Climate Dynamics (2023) 61:1257–1276
https://doi.org/10.1007/s00382-022-06623-4

How to choose credible ensemble members for the sub-seasonal to seasonal prediction of precipitation?

Weihua Jie[1] · Tongwen Wu[1] · Frederic Vitart[2] · Xiangwen Liu[1] · Yixiong Lu[1] · Junchen Yao[1] · He Zhao[1]

Received: 3 August 2022 / Accepted: 3 December 2022 / Published online: 27 December 2022

Abstract
The sub-seasonal to seasonal (S2S) prediction of precipitation is not only a hot topic but also a challenge. The traditional ensemble mean and ensemble probabilistic forecast methods cannot avoid the uncertainty of the initial value in the S2S prediction. Is there a more suitable ensemble postprocessing method for the S2S prediction? In this study, the hindcast data during the 1999–2010 summers from nine operational models in the international S2S prediction project has been evaluated. Based on the quantitative objective precipitation evaluation methods, such as the Equitable Threat Score and frequency bias methods, the climatological spatio-temporal distribution of the optimal probabilistic threshold on the S2S scale is proven to exist, and it can be used as the standard to judge how many ensemble members are credible. Then, different ensemble forecast strategies are adopted in different regions to construct a Deterministic Ensemble Forecast using an Optimal Probabilistic Threshold (DEFOPT) method for precipitation prediction. The hindcast data of eight S2S models outside the period 1999–2010 are used to verify the applicability of the DEFOPT method by using the historical optimal probabilistic threshold during 1999–2010. The results show that the DEFOPT outperforms the deterministic forecast from one initial value, the ensemble mean, and the deterministic ensemble forecast using a probabilistic threshold for the occurrence days of rainfall at the 1 mm and 5 mm thresholds (≥ 1 mm and ≥ 5 mm) over China during each pentad in most S2S models.

Keywords Sub-seasonal to seasonal · Ensemble prediction · Precipitation · Deterministic forecast

1 Introduction

Meteorological disasters caused by extreme weather and extreme climate events have become one of the major problems faced by human beings since the twentieth century. For example, the flood disaster in the Yangtze-Huaihe River Basin of China in the summer of 1998 and the freezing rain and snow disaster in southern China in the winter of 2008 (National Climate Center 1998; Li and Gu 2010) led to great losses to the national economic and social development. Currently, numerical models have made great progress in the medium-range and short-term weather forecasts and the forecasts longer than the seasonal scale. However, there is still a gap between the 2-week forecasts and seasonal forecasts. Therefore, the 15–60 day sub-seasonal to seasonal (S2S) prediction is getting increasing attention worldwide (Vitart et al. 2017; Zhou et al. 2019).

In 2015, the World Weather Research Program and the World Climate Research Program jointly launched an international project of S2S prediction (Vitart et al. 2017). This project aims to enhance the 15–60 day prediction skills, improve the overall understanding of high impact weather events, such as the tropical low-frequency Madden–Julian Oscillation, monsoon, and extreme precipitation, and promote the relevant research carried out by international operational forecast centers and institutions. At present, the operational forecast centers in 11 countries have participated to the S2S project and released a large number of historical hindcast and real-time forecast data of the models (http://www.s2sprediction.net/). The China Meteorological Administration (CMA) also participated in the S2S project, submitted the experiment data based on the Beijing Climate Center Climate System Model (BCC_CSM1.2) of the National Climate Center, and undertook the task of Asian database

✉ Weihua Jie
jiewh@cma.gov.cn

[1] CMA Earth System Modeling and Prediction Center (CEMC), China Meteorological Administration, 46 Zhongguancun Nandajie, Beijing 100081, People's Republic of China

[2] ECMWF, Reading, UK

center for the S2S project. As one of the important issues in this S2S project, the sub-seasonal prediction (15–60 day) of precipitation has attracted wide attention (Ebert 2001). However, the understanding of the predictability and prediction method of precipitation is limited.

Nowadays the weather numerical model performs better in the daily forecast of the geopotential height, air temperature, and precipitation in the leading time of about a week (Saha and Van den Dool 1988; Qin and Van den Dool 1996; Buizza 2008; Schmeits and Kok 2010). Considering the interactions of the atmosphere with the ocean, land surface, and sea ice, the climate model can predict the average and variability of meteorological factors longer than the seasonal scale (Collins and Coauthors 2006; Wu et al. 2013). However, due to the chaos in the atmosphere (Lorenz 1963, 1982; Chou 1989; Hoffman 2002), the inevitable initial value errors and model errors lead to forecast biases in the weather and climate models. Thus, the ensemble mean and probabilistic forecast based on multiple initial values or multiple models are usually carried out to represent the forecast uncertainty caused by one initial value and model errors (Gneiting and Raftery 2005). In the last 20 years, the ensemble forecasting methods such as the Monte Carlo forecast method (Leith 1974), Time-Lagged Average Forecast method (LAF, Hoffman and Kalnay 1983), breeding growing mode method (Toth and Kalany 1993, 1997), singular vectors method (Molteni et al. 1996), ensemble Kalman filter method (Houtekamer and Mitchell 1998), stochastically perturbed parameterization tendencies method (Buizza et al. 1999a, b), multi-model ensemble prediction method (Fritsch et al. 2000), and machine learning approach (Hwang et al. 2019) have gradually become important tools to improve the skill of the weather forecasts, S2S prediction, and even long-term climate change simulation in the national operational centers, which include the National Centers for Environmental Prediction (NCEP), the European Center for Medium-Range Weather Forecasts (ECMWF), the United Kingdom Met Office (UKMO), the Japan Meteorological Agency (JMA), and the Chinese Meteorological Administration (CMA) (Sivillo et al. 1997; Moore and Kleeman 1998; Krishnamurti et al. 2000; Yang 2001; Buizza 2019; Zhang et al. 2021).

For the prediction from day 7 to 60, previous studies have discussed the influence of ensemble forecasting methods, such as the ensemble probabilistic prediction (Pan and Van den Dool 1998; Chessa and Lalaurette 2001), the conditional nonlinear optimal perturbation ensemble (defined as a kind of initial perturbation which makes the cost function acquire their maximum under an initial constraint condition; Jiang et al. 2009), the weather type ensemble forecast (designed for the purpose of post-processing forecast output from ensemble prediction systems and understanding how forecast models perform under different circulation types; Neal et al. 2016), predictability-based extended-range ensemble prediction (proposed for the predictable components and random components obtained with different ensemble prediction strategies; Zheng et al. 2012), on the prediction skill of geopotential height, wind, and air temperature.

Other studies have analyzed and evaluated the impact of ensemble forecast methods on the sub-seasonal precipitation prediction skill (Hamill et al. 2004; Whitaker et al. 2006; Vitart and Molteni 2009; Jie et al. 2013; Bombardi et al. 2017; Liang and Lin 2018; Li et al. 2019). However, the improvement of weekly to sub-seasonal precipitation prediction is limited compared to other time-scale forecasts (Tan and Chen 2013; Jie et al. 2013), and the statistical post-processing ensemble of precipitation is far more challenging than that of weather variables like surface temperature or wind speed (Scheuerer 2014). At present, both the ensemble probabilistic forecast method and ensemble mean method are not good enough in the day 7 to sub-seasonal precipitation prediction due to the excessive increase of the ensemble spread after 1 week (Jie et al. 2014). For the ensemble probabilistic forecast, a few of the ensemble probabilistic thresholds become less skillful with lead times (Buizza et al. 1999a, b; Hamill et al. 2008). In the late period of forecasting (high lead times), the precipitation is significantly underestimated (overestimated) by the forecast with a high (low) probabilistic threshold (Jie et al. 2014). For the ensemble mean forecasting, the precipitation bias is more likely to be caused by the false extreme precipitation predicted by a certain ensemble member if the ensemble size is not large enough (Jie et al. 2014). Considering this, Jie et al. (2014) proposed a method of Deterministic Ensemble Forecast using a Probabilistic Threshold (DEFPT), which selects ensemble members through a certain ensemble probabilistic threshold. It can greatly improve the 6–15 day forecast skill in summer precipitation of different intensities in China, although the spatio-temporal variation of the probabilistic threshold is not considered and only the applicability of this method in a time-lagged ensemble system is verified. Meanwhile, it can avoid the influence of the false extreme value of the precipitation forecasted by a certain ensemble member on the ensemble forecasting. However, to some extent, there are still deviations in the precipitation prediction by using the DEFPT method based on a same probabilistic threshold from ensemble members in different regions, which may be related to the different regional systematic forecast errors of the model.

In this work, the quantitative objective statistical methods are used to explore the spatio-temporal variation of the available probabilistic forecast information in the sub-seasonal forecast. The credible ensemble members (i.e. smaller biases and more skillful members) are selected, based on the spatio-temporal variation, and the optimal ensemble strategy is provided for different regions to be used in the

S2S precipitation ensemble forecast. The applicability of the method in different S2S operational models is verified. This article is organized as follows: Model data and ensemble methods are introduced in Sect. 2; the verification and evaluation of the ensemble methods are provided in Sect. 3; the results are explored in Sect. 4; and Sect. 5 provides a summary and discussion.

2 Model data and ensemble methods

2.1 Model data

In this study, the precipitation data from hindcast experiments from eleven operational prediction models in the S2S project are used (Table 1). All the data cover the period of 1999 to 2010, and are downloaded from http://www.s2sprediction.net/. Although the S2S models have different horizontal resolutions, each operational center uploaded model output to the S2S database archiving centers with a unified horizontal resolution 1.5° × 1.5° except the BoM model with a lower resolution 2.5° × 2.5°. As the Institute of Atmospheric Sciences and Climate of the Italian National Research Council only provided a single re-forecast sample and there are some errors in the ensemble forecast data submitted by the Hydrometeorological Center of Russia (Jie et al. 2017), the ensemble forecasting methods are evaluated and analyzed based on only nine operational models. In this study, the observed rainfall-gauge data over China are interpolated to the corresponding horizontal resolution of each S2S model, and daily accumulated precipitation from each S2S model is analyzed. In addition, eight models with longer re-forecast length than the NCEP (1999–2010) are further examined (Table 3).

2.2 Ensemble methods

In this paper, the Deterministic Ensemble Forecast using an Optimal Probabilistic Threshold (DEFOPT) method is proposed for the S2S (15–60 days) precipitation prediction. The DEFOPT is different from the traditional probabilistic forecast method as it does not predict the probability of precipitation event in each grid, but it uses the available probabilistic forecasting information in the S2S real-time scale to decide how many ensemble members predicting the occurrence of rainfall event should be trusted, and then determines the optimal ensemble forecast in different regions. The details are as follows.

First, in order to avoid the excessive overestimation or underestimation in the ensemble probabilistic forecast of the rainfall event at a certain intensity (the precipitation with the threshold of 1 mm) at each grid point, the rainfall forecasting frequency bias for the probabilistic threshold P_c is limited by a quantitative objective evaluation method—BIA score (see Appendix 1 for details) based on the multi-year hindcast results. This limitation is $\alpha \le BIA(P_c) \le \beta$, where α and β are empirical coefficients artificially selected according to the BIAs of the forecasts using different probabilistic thresholds from each model (e.g. Fig. 1 for ≥ 1 mm; Fig. S1 for ≥ 5 mm in the supplementary material). In this study, α and β are first tuned to have good performances of the DEFOPT in the climatology, and then are used to examine the hindcasts outside this period.

Second, within the reasonable range of the forecasting frequency bias, the optimal probabilistic forecasting threshold $P_{threshold}$ is defined when the skill of the daily precipitation prediction is highest during 12 years at each grid point. Here, the skill is examined by using Equitable Threat Score (ETS; Schaefer 1990; see Appendix 1 for details). The calculation formula is as follows.

Table 1 Operational models of the S2S project

	Time range	Resolution	Re-forecast frequency	Ensemble member	Re-forecast length	Ocean coupled	Sea-ice coupled
ECMWF	d 0–46	Tco639/319L91	3/4 days	11	1995–2015	Yes	No
NCEP	d 0–44	T126L64	daily	4	1999–2010	Yes	Yes
CMA	d 0–60	T106L40	daily	4	1994–2014	Yes	Yes
JMA	d 0–33	T319L60	10 days	5	1981–2010	No	No
CNRM	d 0–61	T255L91	15 days	15	1993–2014	Yes	Yes
HMCR	d 0–61	1.1 × 1.4L28	7 days	10	1985–2010	No	No
BoM	d 0–62	T47L17	5 days	33	1981–2013	Yes	No
CNR-ISAC	d 0–31	0.75 × 0.56L54	5 days	1	1981–2010	No	No
UKMO	d 0–60	N216L85	8 days	3	1993–2015	Yes	Yes
ECCC	d 0–32	0.45 × 0.45L40	5 days	4	1995–2014	No	No
KMA	d 0–60	N216L85	daily	3	1991–2010	Yes	Yes

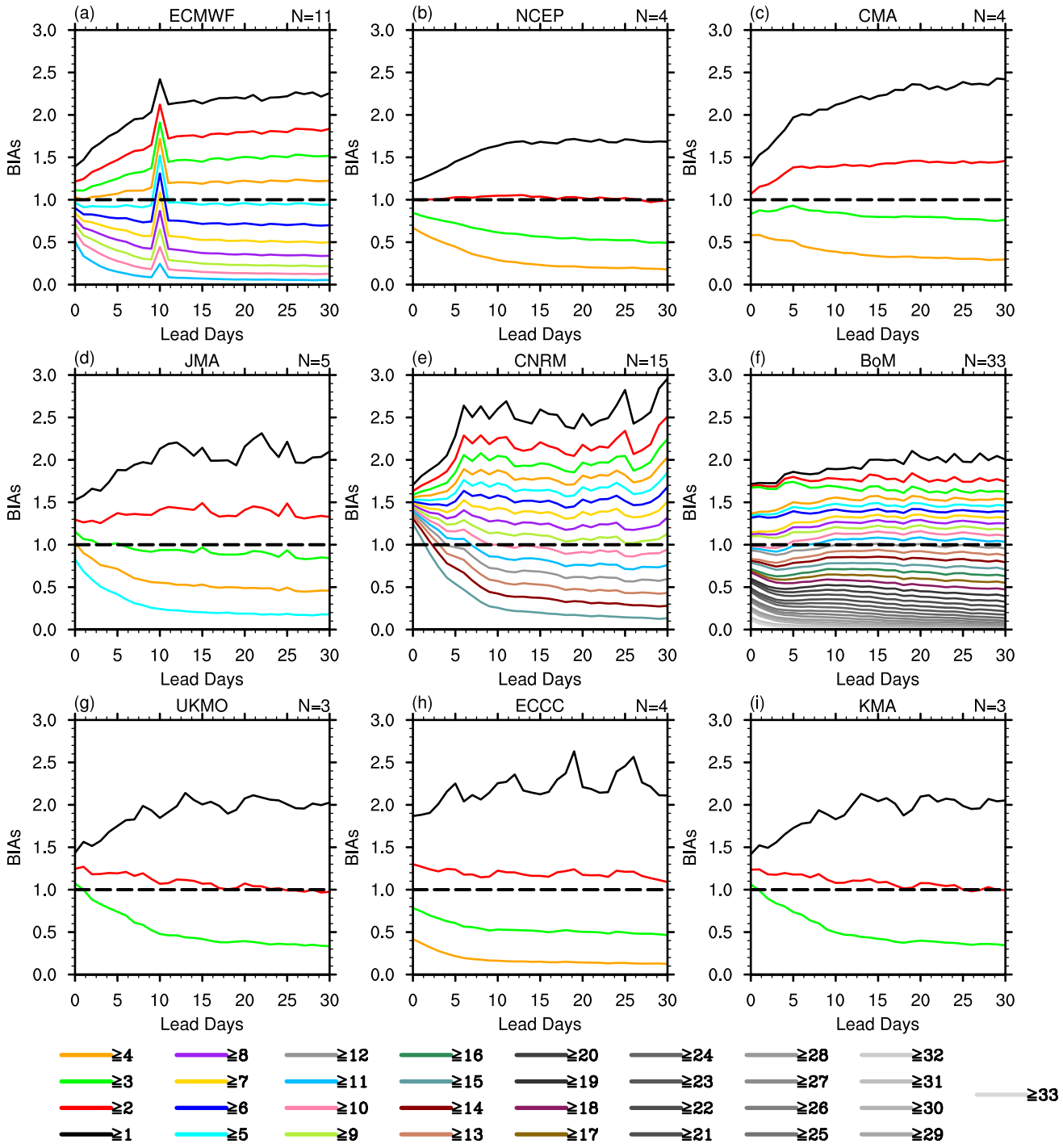

Fig. 1 Temporal variation of the BIA averaged over each grid for the ≥ 1 mm precipitation in summers from 1999 to 2010 based on the different probabilistic threshold forecasts of the S2S multiple models. *N* is the total number of ensemble members from each model, and the numbers in the legend indicate the numbers of ensemble members predicting the occurrence of rainfall event. The BIA > 1.0 (BIA < 1.0) means overestimation (underestimation) of precipitation frequency

$$P_{threshold} = \begin{cases} \mathrm{P_{min}}, \ \ BIA\left(\mathrm{P_{min}}\right) \le \beta \\ P_{\mathrm{c}}, \ \ ETS\left(\mathrm{P_c}\right) = Max\left(\mathrm{ETS}\begin{pmatrix} \mathrm{P_{min}} \\ \cdot \\ \cdot \\ P_{\mathrm{c}} \\ \cdot \\ \cdot \\ P_{\mathrm{max}} \end{pmatrix}\right) \\ P_{max}, \ \ BIA\left(P_{\mathrm{max}}\right) \ge \alpha \end{cases} \quad (1)$$

In the equation, $\mathrm{P_{min}}$ and $\mathrm{P_{max}}$ are the minimum and maximum values of $\mathrm{P_{threshold}}$, respectively (e.g. $\mathrm{P_{min}} \le \mathrm{P}_c \le \mathrm{P_{max}}$), when the $BIA(\mathrm{P}_c)$ scores are within the reasonable range. After calculating the $\mathrm{P_{threshold}}$ at each grid, the spatio-temporal distribution of $\mathrm{P_{threshold}}$ in different regions can be achieved.

Third, according to the spatio-temporal distribution characteristics of the optimal probabilistic threshold, the threshold of the credible ensemble number ($\mathrm{N_{threshold}}$) is selected. $\mathrm{N_{threshold}} = \mathrm{P_{threshold}} \times \mathrm{n}$, where n is the total number of ensemble members. Then, whether the forecasted rainfall event occurs or not is redefined, that is, the forecasted rainfall event occurs when the number of ensemble members that predict the rainfall event (N) is greater than or equal to the $\mathrm{N_{threshold}}$. Otherwise, the forecasted rainfall event does not occur. The formulas are as follows.

$$\mathrm{A_{DEFOPT}} = \mathrm{A_{threshold}} \times \phi \quad (2)$$

$$\phi = \begin{cases} 1, \ N \ge \mathrm{N_{threshold}} \\ 0, \ N < \mathrm{N_{threshold}} \end{cases} \quad (3)$$

In the equation, $\mathrm{A_{threshold}}$ is the amount of rainfall at a certain threshold (for example, ≥ 1 mm). $\mathrm{A_{DEFOPT}}$ is the final result of ensemble forecasting at this threshold. ϕ indicates whether the precipitation event occurs (1) or not (0). If N is greater than or equal to $\mathrm{N_{threshold}}$, ϕ is 1. Otherwise, ϕ is 0.

To evaluate the benefits of the DEFOPT method, the temporal variation of the BIA (averaged over the rainfall events in the re-forecast length at each grid point) over China for the ≥ 1 mm and ≥ 5 mm precipitation in summers from 1999 to 2010 in China are shown based on the $\mathrm{P_c}$ of the nine S2S models (Fig. 1). For the ≥ 1 mm rainfall, the deviation between the high-threshold (e.g. 11 ensemble members for e.g. ECMWF) and low-threshold (e.g. 1 ensemble member) probabilistic forecasts of the S2S models increases rapidly within 10 days and tends to be steady after 10 days, and the forecast with the low (high) probabilistic threshold significantly overestimates (i.e. BIA > 1.0) [underestimates (i.e. BIA < 1.0)] the observed rainfall. The range of the corresponding BIA is significantly larger than that in the early stage of the forecast. The ranges of the BIAs of the ECMWF, NCEP, CMA, JMA, and ECCC models increase from 0.5–1.5 to 0.2–2.0. The ranges change from 1.0–1.5 to 0.3–2.0 for the UKMO and KMA models, increase from 1.3–1.7 to 0.1–3.0 for the CNRM model, and change from 0.1–1.7 to 0.0–2.0 for the BoM model. It is similar for the 5 mm precipitation, where the corresponding BIA begins to deviate from one standard deviation after 5 days (Fig. S1). Based on the temporal variation characteristics of BIA, the limitation of the BIA for the $\mathrm{P_c}$ of each S2S model is given to avoid excessive overestimation or underestimation: $\alpha \le \mathrm{BIA}\left(\mathrm{P_c}\right) \le \beta$. The values of the empirical coefficients α and β are shown in Table 2, but the variability of α and β with lead time is not considered in this study.

Within the proper range of the BIAs, the spatio-temporal distribution of the $\mathrm{P_{threshold}}$ with the highest ETS for the ≥ 1 mm precipitation prediction is calculated according to the S2S re-forecast data of daily precipitation in summers during 1999–2010 in China (Fig. 2). Figure 2a shows the average spatial distribution of the $\mathrm{P_{threshold}}$ from the 11th to the 15th day for the ≥ 1 mm rainfall in the summers of China predicted by the ECMWF model with eleven ensemble members. The $\mathrm{P_{threshold}}$ is within 30%–40% in most areas of northern China, central China, and eastern China, and within 50%–70% in some areas of southern China, southwestern China, and the southern part of the Qinghai-Tibet Plateau. However, the $\mathrm{P_{threshold}}$ in the arid and semi-arid

Table 2 Empirical coefficients α and β for precipitation events with different thresholds forecasted by the S2S models

BIAs ranges	ECMWF	NCEP	CMA	JMA	CNRM
1 mm	1.2–1.5	1.1–1.6	1.1–1.6	1.2–1.6	1.1–1.7
5 mm	1.2–3.5	1.1–3.0	1.1–4.0	1.3–3.5	1.1–3.5
BIAs Ranges	BoM	UKMO	ECCC	KMA	
1 mm	0.9–1.6	0.9–1.6	1.2–1.6	0.9–1.6	
5 mm	1.5–3.0	1.6–3.5	1.3–3.5	1.7–3.5	

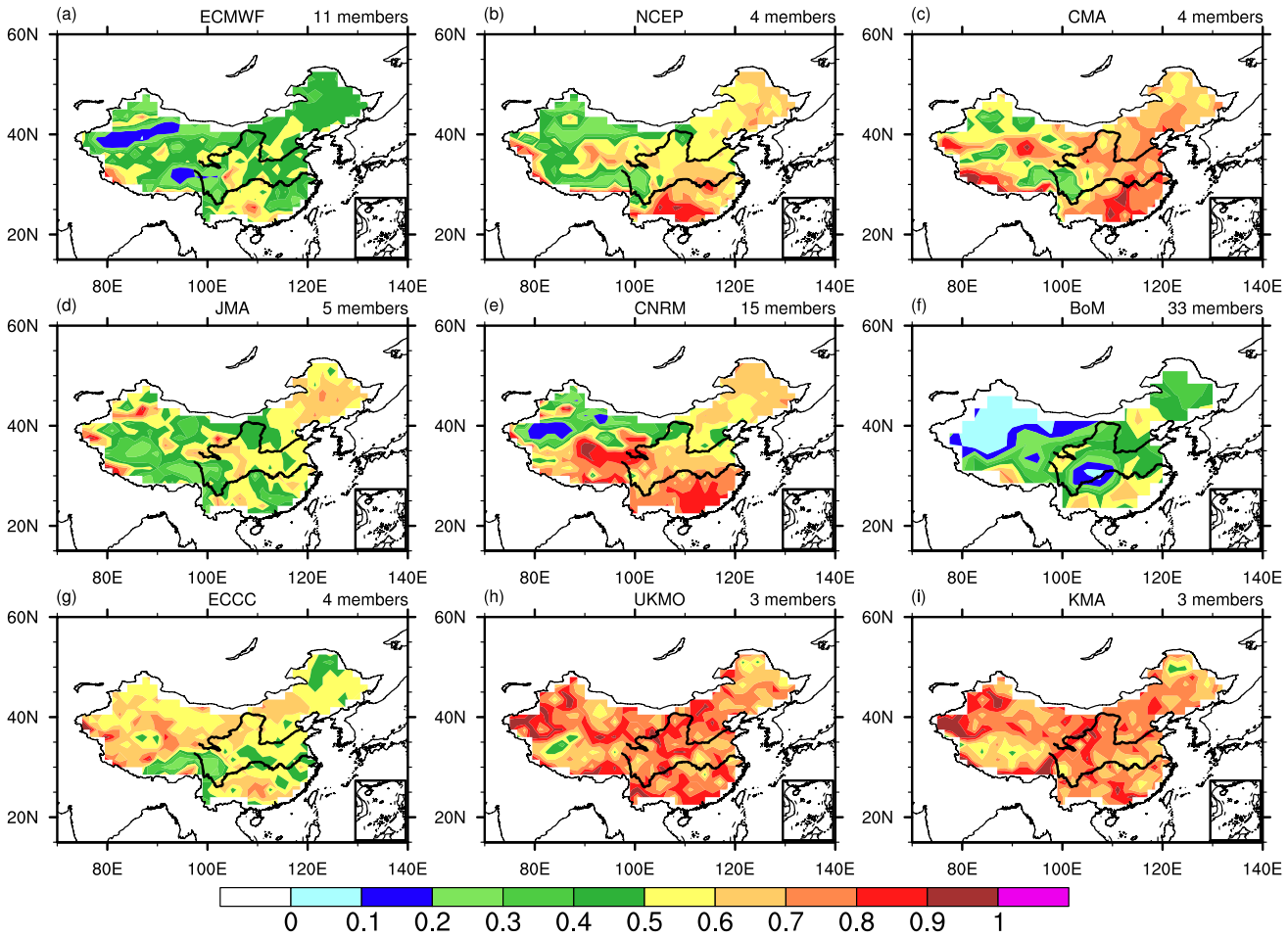

Fig. 2 Average spatio-temporal distribution of the $P_{threshold}$ with the highest ETS of the 1 mm precipitation from the 11th to the 15th day (the third pentad) calculated by using the historical hindcast data of daily precipitation in summers from 1999 to 2010 in China based on the S2S models. **a**–**f** Results of the ECMWF, NCEP, CMA, JMA, CNRM, BoM, UKMO, ECCC, and KMA models, respectively

areas in northwestern China is about 20%. The results of the JMA model with five ensemble members are similar to those of the ECMWF model, but the $P_{threshold}$ is slightly higher in northeastern China and the lower reaches of the Yangtze River, which is about 50%–60% (Fig. 2d). For the CMA, NCEP, and CNRM models, the $P_{threshold}$ is above 50% in the areas to the east of 110°E and exceeds 70% in the southern and some parts of northeastern China. It is within 20–30% in the arid and semi-arid areas in northwestern China and the upper reaches of the Yangtze River (Fig. 2b, c, and e). For the UKMO and KMA models containing three ensemble members, the $P_{threshold}$ stays about 70% in general (Fig. 2h, i). The $P_{threshold}$ of the ECCC model is around 60% in most areas of China, but relatively low in the reaches of the Yangtze River (about 40%, Fig. 2g). The $P_{threshold}$ of the BoM model with 33 ensemble members is within 10%–20% in the arid and semi-arid areas and central China, which is significantly lower than other models (Fig. 2f). Figure 3 further shows the average spatial distribution of the $P_{threshold}$ from the 25th to the 30th day for the S2S models. It shows that the spatial variation of the $P_{threshold}$ for each model is similar in the different pentads. In addition, the average spatial distributions of the $P_{threshold}$ in other pentads within 30 days based on the S2S models are compared and analyzed (Fig. S2–S5). The results also show that the spatial distributions of $P_{threshold}$ in each model are similar on the sub-seasonal scale (6–30 days) with only a 10% difference between each pentad. Here, in order to display the main tendency of $P_{threshold}$ and filter the high-frequency information, we do not show the optimal probabilistic threshold day by day.

The pentad spatial distributions of the $P_{threshold}$ of the $\geq$5 mm rainfall within 30 days are further analyzed. The 3rd pentad-averaged $P_{threshold}$ for the ECMWF model is within 20%–40% in most areas of China, but within 50–70% in the part of the Qinghai-Tibet Plateau (Fig. 4a). For the NCEP, UKMO and KMA models, the $P_{threshold}$ is above 50% in the southern and some parts of northeastern China, and close to 80%–90% over the northern Qinghai-Tibet Plateau (Fig. 4b,

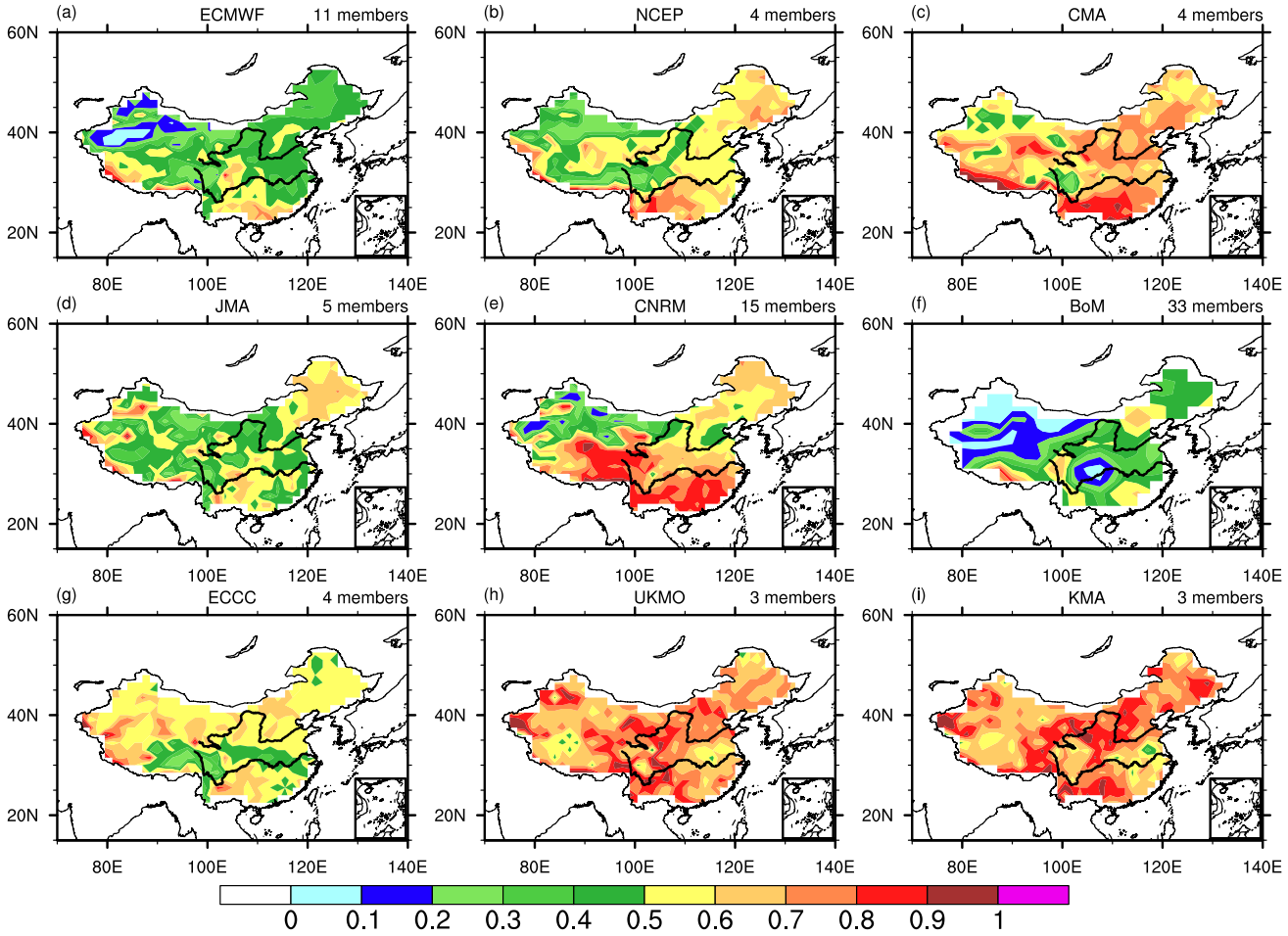

Fig. 3 Same as Fig. 2, but for the 26th to the 30th day

h and i). The results of the CMA and JMA are similar to those of the above three models, but the $P_{threshold}$ is lower than 50% in southern China (Fig. 4c, d). For the CNRM and ECCC models, the $P_{threshold}$ is generally 30%–40% in the areas to the east of 110°E and exceeds 50% in the other areas (Fig. 4e, g). The $P_{threshold}$ of the BoM model is generally lower than other models over China, especially 10–20% in central China and the arid and semi-arid areas (Fig. 4f). For each S2S model, the results of the spatial distributions of $P_{threshold}$ of the ≥ 5 mm rainfall in the 3rd pentad are also similar to other pentads during 6–30 days (Fig. S6–S10).

Therefore, the spatio-temporal distribution characteristics of the $P_{threshold}$ enable us to select the credible ensemble members by using the $P_{threshold}$ calculated from numerous hindcast results. Then, the DEFOPT ensemble forecast can be constructed to carry out the deterministic forecasts for the precipitation events with different intensities, for example, 1–5 mm. As compared to the DEFOPT, the DEFPT method proposed in our previous work (Jie et al. 2014) predicts "yes" or "no" occurrence of rainfall event with a given intensity only by judging whether or not the forecast probability exceeds a constant threshold without spatio-temporal variability. To demonstrate the added value of DEFOPT, we therefore compare the DEFOPT (spatio-temporally variable threshold) with the DEFPT (constant threshold), as well as with the deterministic forecast from control run (CTL) and the classical ensemble mean (ENS mean).

3 Verification and evaluation of the DEFOPT method

Based on the spatio-temporal variation characteristics of the $P_{threshold}$ of the ≥ 1 mm precipitation during 1999–2010 in the S2S models, the DEFOPT method is applied to each S2S model to predict the ≥ 1 mm daily rainfall in the summer of 1999–2010 over China. The quantitative objective precipitation evaluation results from the ETS and Hanssen-Kuipers scores (abbreviated as HK, Hanssen and Kuipers 1965, see Appendix 1 for details) indicate that the DEFOPT is outperforming the CTL and the ENS mean at the lead time 0–30 days in

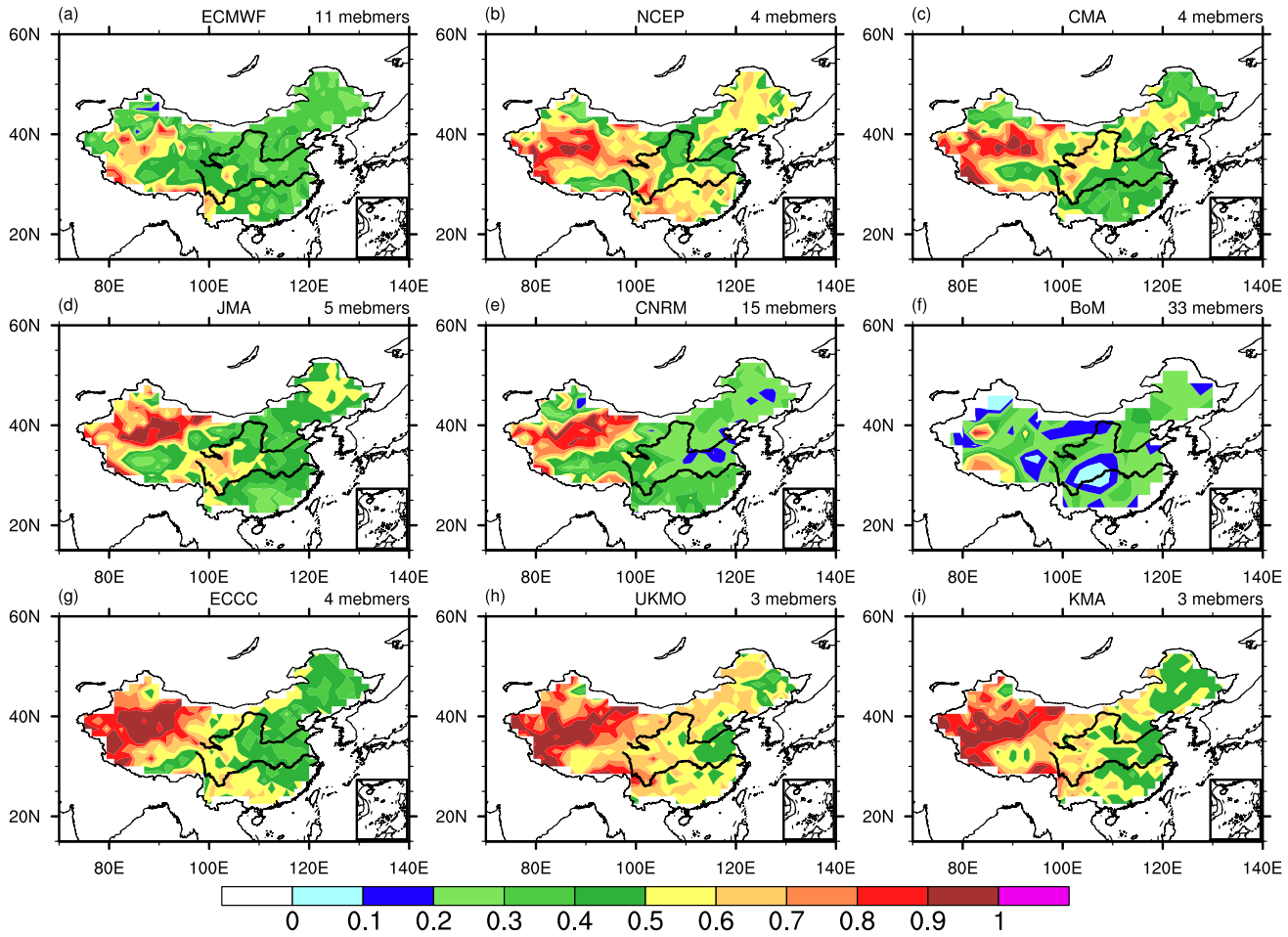

Fig. 4 Same as Fig. 2, but for the ≥ 5 mm rainfall

each S2S model, and also better than majority of forecasts using the different probabilistic thresholds in forecasting the ≥ 1 mm rainfall in the NCEP, CMA, JMA, BoM, UKMO, ECCC and KMA models, although the performance of the DEFOPT is not better than the forecasts produced by using 5/11 and 6/11 probabilistic thresholds in the ECMWF model and 10/15 and 11/15 in the CNRM model after 10 days (Figs. 5 and 6). Generally, the corresponding ETS and HK scores can increase by about 20% by using the DEFOPT compared to the CTL and ENS mean methods. Meanwhile, the BIAs reveals that the frequency bias of DEFOPT is smaller than the ENS mean and most of the forecasts using the probabilistic thresholds during the sub-seasonal range in each S2S model, although the CTL is better than the DEFOPT for many models (Fig. 7). The DEFOPT's BIA scores are not far away from 1.0, and the values are approximately equal to 1.3.

For the ≥ 5 mm rainfall, the skill of the DEFOPT for all S2S models is substantially higher than that of the CTL, ENS mean and the forecasts by using different probabilistic thresholds in general, as the corresponding ETS (Fig. 8) and HK (Fig. S11) is highest, and the BIA is close to or slightly larger than that of ENS mean (Fig. S12).

The DEFOPT method is further evaluated for the prediction of the frequencies of the daily ≥ 1 mm and ≥ 5 mm daily rainfall events within each pentad and 10-day periods in summer (Fig. 9). The Pearson correlation between the number of observed and forecasted ≥ 1 mm rainfall days in each pentad from each S2S model in summers during 1999–2010 shows that the ensemble forecasting skill of the DEFOPT (red solid line) is higher than that of the CTL (black solid line), the ENS mean (black dotted line), and the DEFPT using the same probabilistic threshold for the entire region (blue solid line). The corresponding correlation coefficients increase by about 0.1–0.2, 0.05–0.1, and 0.05, respectively. The predictions of the pentad frequency for the ≥ 5 mm rainfall events show that the DEFOPT method (marked red solid line) can improve the forecast skills within 30 days for each S2S model compared to other ensemble forecasting methods. The improvement is particularly large relative to CTL (marked black dotted line) and ENS mean (marked

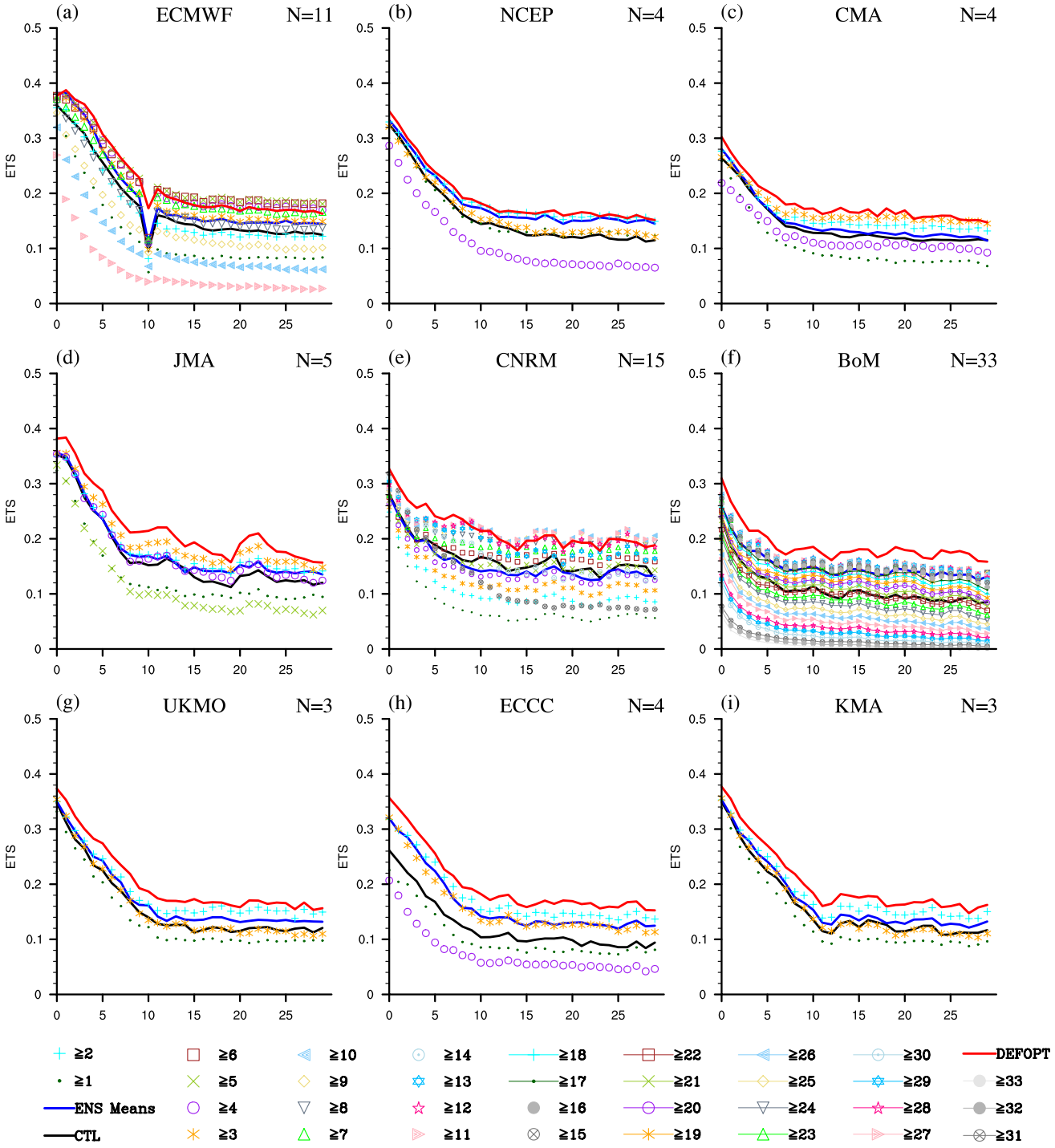

Fig. 5 The ETS of the ≥ 1 mm daily precipitation in the summer of 1999–2010 in China forecasted by the S2S models at lead time 0–30 days. The black solid line, colored markers, blue solid line, and red solid line represent the results of the control run, the forecast by using different probabilistic thresholds, the ensemble mean and the DEFOPT, respectively. *N* is the total number of ensemble members from each model, and the numbers in the legend indicate the numbers of ensemble members predicting the occurrence of rainfall event. The higher ETS score, the better prediction

blue solid line). The corresponding correlation coefficient increases by about 0.1–0.2 (0.05–0.1) compared to that of CTL (ENS mean).

Based on the maximum lead time provided by each S2S model (Table 1), the forecast skills for the frequency of the daily ≥ 1 mm and ≥ 5 mm precipitation in each period of ten

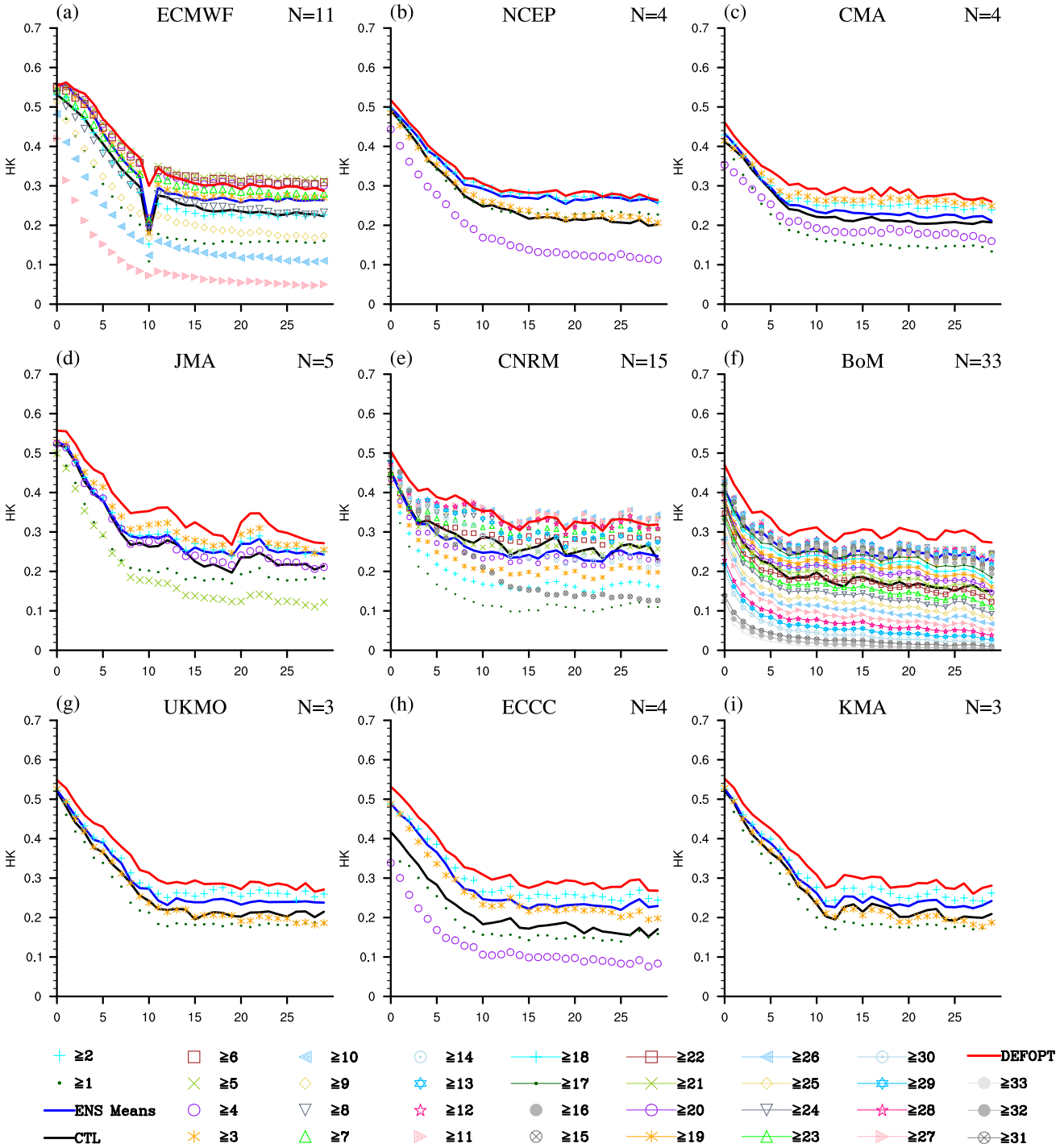

Fig. 6 Same as Fig. 5, but for the HK

days at a longer time range are evaluated (Figs. 10). The results show that the DEFOPT method (red solid line) can significantly improve the sub-seasonal to seasonal forecast skills of each S2S model compared to the CTL (black solid line) by about 0.1–0.2, and the ENS mean (black dotted line) and DEFPT (blue solid line) methods. In addition, it was noticed that the ensemble mean forecast skill for the CNRM model with fifteen ensemble members is lower than that of the CTL in the ≥ 1 mm rainfall prediction (Fig. 10e). This could be caused by the overestimation of the rainfall intensity by most ensemble members (the BIA score is high, Fig. 7e) which leads to a substantial increase of the false forecast for the ≥ 1 mm rainfall events when using the ENS mean. For the ≥ 5 mm rainfall, the performance of the

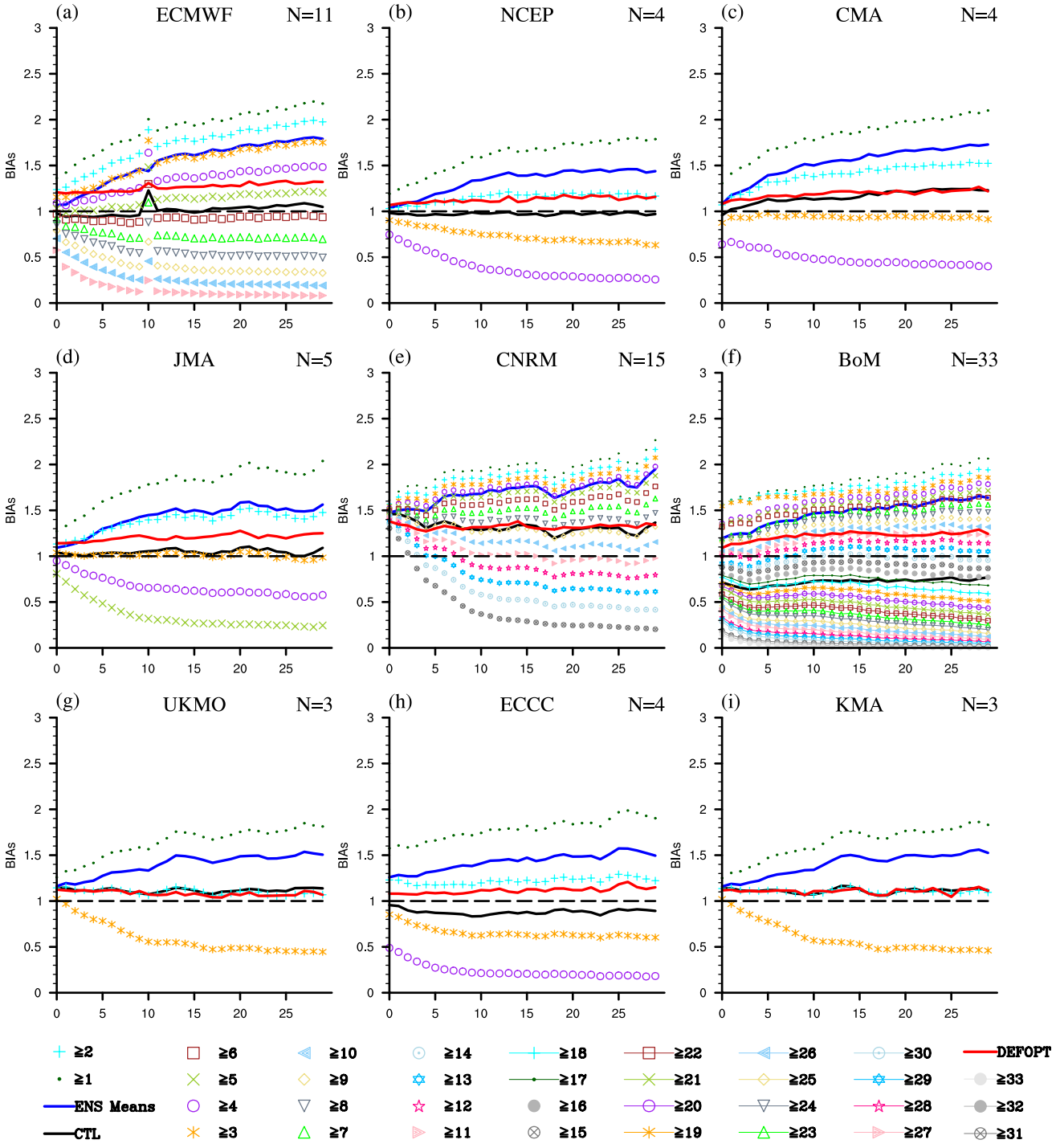

Fig. 7 Same as Fig. 5, but for the BIAs. The long black dotted line is the standard of the BIA that equals 1.0. The BIA > 1.0 (BIA < 1.0) means overestimation (underestimation) of precipitation frequency

DEFOPT in each model is much better than that of ≥ 1 mm rainfall with the correlations increasing by about 0.1 – 0.2 from all the other methods (marked lines).

In order to further verify the applicability of the DEFOPT method, the frequencies of the daily ≥ 1 mm and ≥ 5 mm precipitation in each period of ten days are evaluated during other re-forecast periods excluding 1999–2010 (Table 3). Except for the NCEP model, the other eight models have at least 8 years samples for evaluation. Whether for ≥ 1 mm or ≥ 5 mm rainfall, the DEFOPT is still better than other methods in most S2S models (Fig. 11).

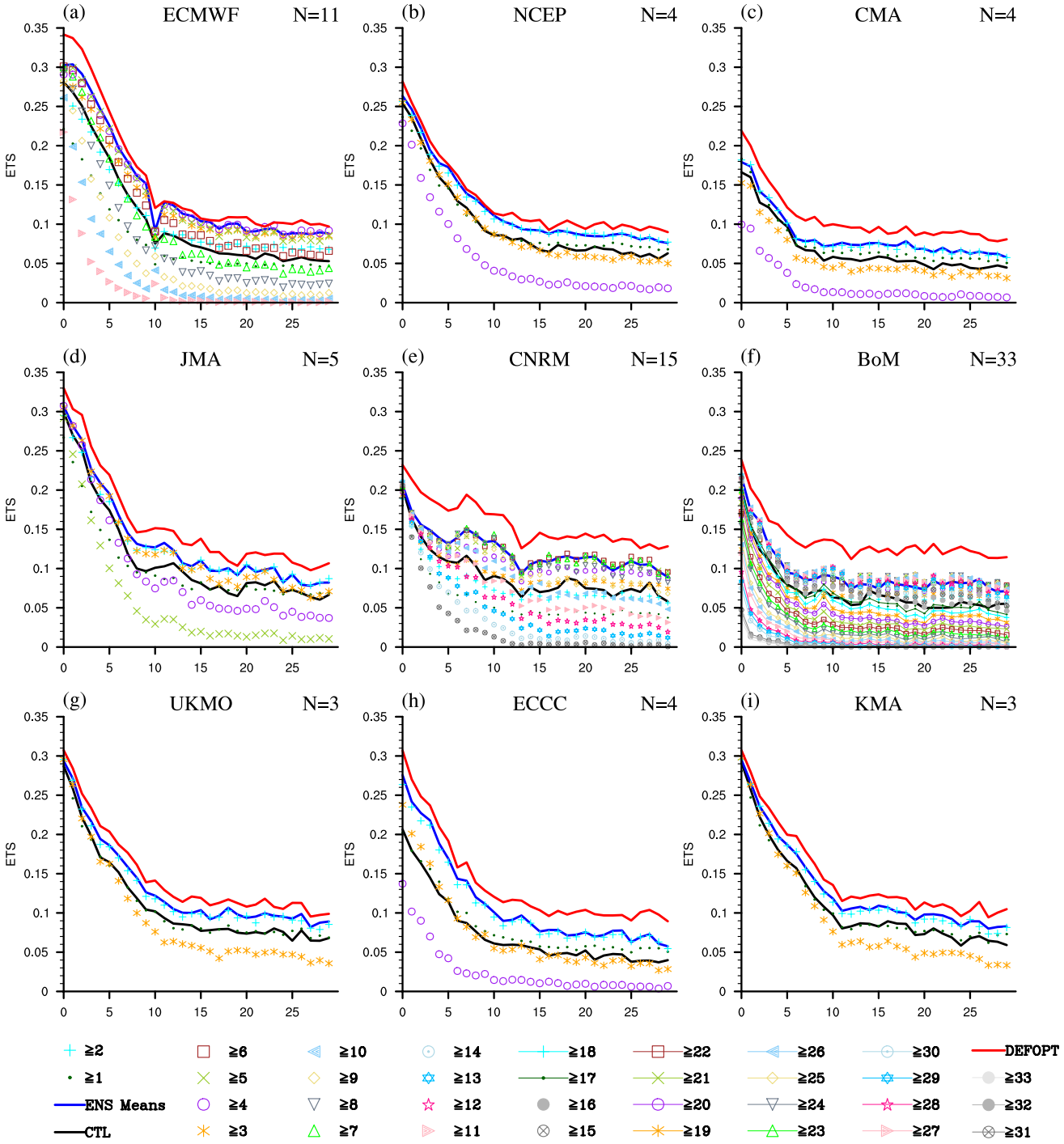

Fig. 8 Same as Fig. 5, but for the ≥ 5 mm

We further analyzed the performance of individual ensemble member chosen by the DEFOPT at the different lead times. Figure 12 shows the proportion of each ensemble member predicting the occurrence of ≥ 1 mm rainfall event in all ensemble members chosen by the DEFOPT in the S2S models during the summer of 1999–2010 over China. It is clear that the proportion of individual ensemble member in the ECMWF, NCEP, CMA, JMA, CNRM, KMA, UKMO and BoM is similar or shows a slight fluctuation in the different lead times at 5-days intervals. It indicates that the ensemble members selected by the DEFOPT are random. However, the proportions of the

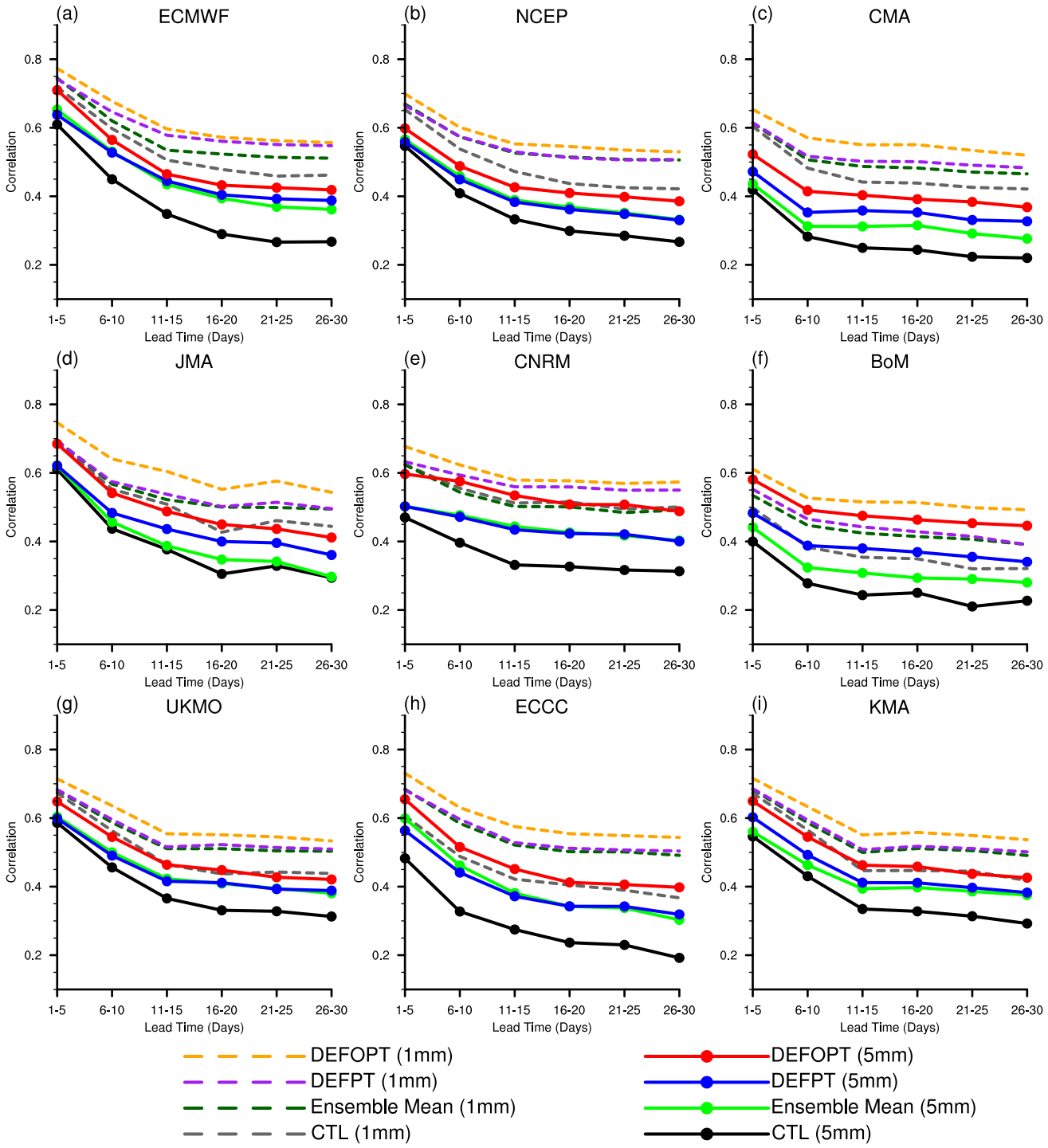

Fig. 9 Temporal variation of the correlation coefficient between the observed and forecasted frequencies of days with ≥ 1 mm (dash colored lines) and ≥ 5 mm (the marked lines) precipitation by using the S2S multiple models in each pentad in summers during 1999–2010. The different color lines represent the results of the CTL, the ENS, the DEFPT (the most skillful forecast by using a probabilistic threshold), and the DEFOPT, respectively

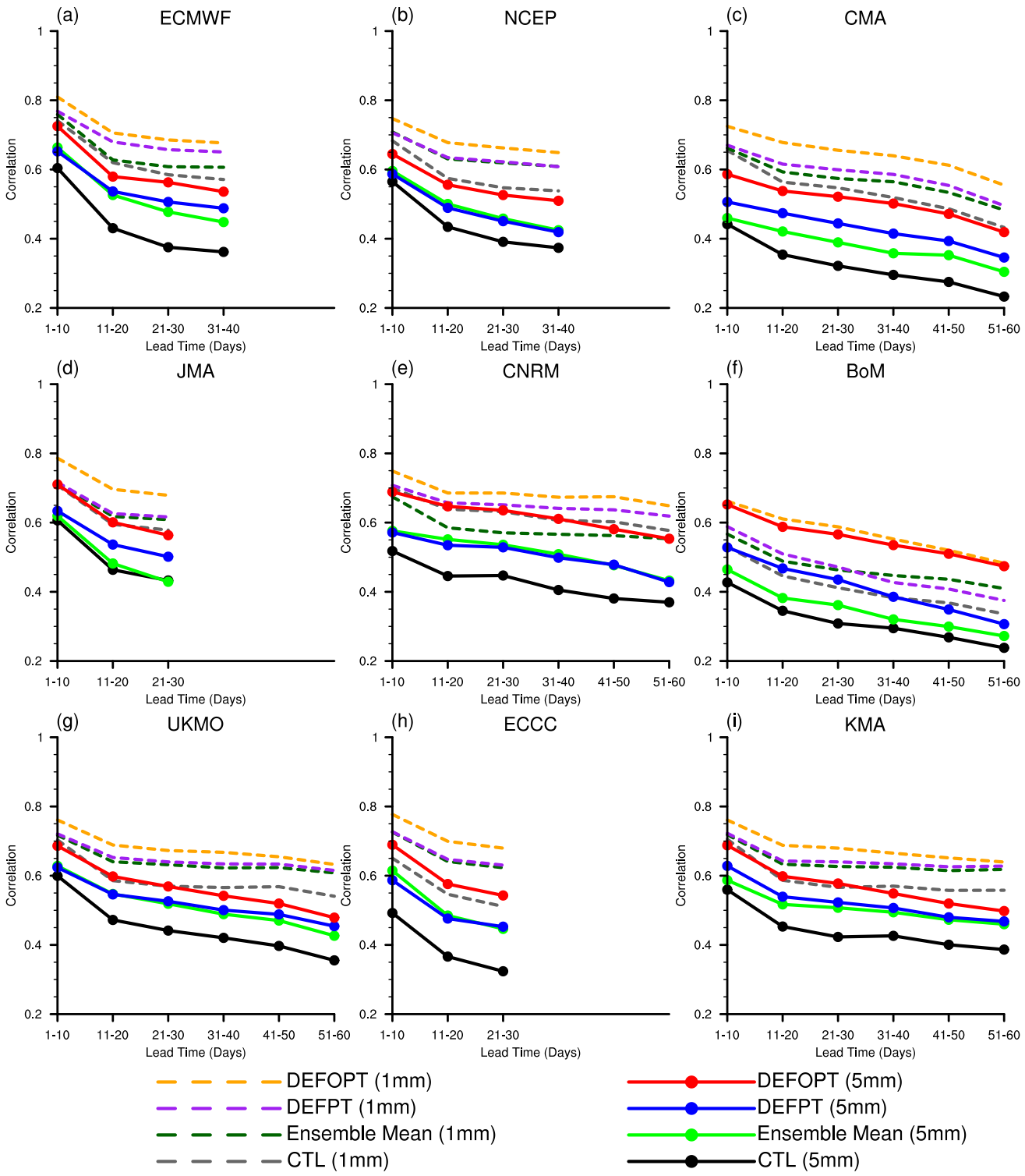

Fig. 10 Same as Fig. 9, but for the correlation coefficient between the observed and forecasted frequency of precipitation events every ten days

member 2 and 4 in the ECCC are higher than the member 1 and 3. It is possible that the member 2 and 4 have systematic biases. The similar result can also be found in the ≥ 5 mm rainfall event (Fig. S13).

Table 3 The analyzed S2S models outside the period 1999–2010

	Re-forecast length	outside the period 1999–2010	Remaining years
ECMWF	1995–2015	1995–1998 2011–2015	9
NCEP	1999–2010	–	–
CMA	1994–2014	1994–1998 2011–2014	9
JMA	1981–2010	1981–1998	18
CNRM	1993–2014	1993–1998 2011–2014	10
BoM	1981–2013	1981–1998 2011–2013	21
UKMO	1993–2015	1993–1998 2011–2015	11
ECCC	1995–2014	1995–1998 2011–2014	8
KMA	1991–2010	1991–1998	8

4 Discussion and conclusion

The DEFOPT method is proposed to choose credible ensemble members for the sub-seasonal to seasonal prediction of precipitation in this paper. It uses the spatio-temporal distribution characteristics of the optimal probabilistic threshold which were proven to exist in the climatology as the standard to decide how many ensemble members should be trusted on the S2S scale. The optimal ensemble strategy is made for S2S precipitation prediction by following 3 steps:

1. Based upon hindcasts with long period, exclude the probabilistic thresholds with large frequency biases by using the limitation of BIA score at each grid point.
2. Find out a most skillful probabilistic threshold (with the highest ETS) from the leftover probabilistic thresholds (after step 1) via the ETS score at each grid to generate a climatological spatio-temporal distribution of the optimal probabilistic threshold.
3. Determine the number of skillful ensemble members in the real-time prediction by judging whether the number is greater than or equal to the optimal probabilistic threshold or not, based upon the spatio-temporal distribution characteristics of the optimal probabilistic threshold from the climatology.

Here, all these steps are just part of the post-processing based on 12 years of hindcasts and are not part of the numerical modeling integration. By using Fortran codes on a regular UNIX workstation, the process of selecting the credible ensemble members (including step 1 and 2) spends about 2 min (clocktime) on a model with horizontal resolution 1.5° × 1.5° and ~ 10 ensemble members over China during 12 hindcast years, and about 30 s on generating an adjusted real-time forecast (step 3). Thus, the DEFOPT will not be computationally expensive in the operational application.

In this work, the quantitative objective evaluation scores including ETS, HK and BIA widely used to evaluate model precipitation forecasts (Accadia et al. 2010; Weusthoff et al. 2010) are selected. All these scores are constructed by hits, false alarms and misses of rain event forecast, and no-rain event accurate forecast (correct reject) as shown in Table 4, respectively. The ETS and HK scores focus on the prediction skill of rainfall and no rainfall events, meanwhile the BIA score shows the overestimation or underestimation of the frequency of rainfall events. The evaluation results of the application of the DEFOPT method on the nine S2S operational models show that this methodology can substantially improve the forecast skill of precipitation events with 1 mm and 5 mm thresholds in the S2S summer over China, and its skill is better than that of the CTL, the ENS mean, and the DEFPT as shown in ETS and HK evaluation methods. Meanwhile, the frequency bias of the DEFOPT for the ≥ 1 mm precipitation is smaller than the ENS mean, although it is close to or slightly larger than the CTL. For the ≥ 5 mm precipitation, the DEFOPT frequency bias is generally greater than the ENS mean and the CTL, but is not far away from them. The main reason for the improvements is that the DEFOPT can substantially increase the hit rates of ≥ 5 mm rainfall, although slightly increase the false alarm rates in part of S2S models (such as ECMWF, JMA, CNRM and BoM) as compared to the traditional ensemble mean method, the control run; and it also can increase the hits or decrease the false alarms in comparison to the DEFPT using a uniform probabilistic threshold in most S2S models (Fig. 13). For the low intensity rainfall (e.g. ≥ 1 mm), the DEFOPT can substantially decrease the false alarms compared to the ENS which shows not only high hits but also too high false alarms in most S2S models, and it is also more close to the left top corner of subplot compared with the DEFPT in each S2S model except the ECMWF (Fig. S14).

As compared to some ensemble reduction techniques (reduction by ‘‘uncorrelation’’ method, by principal component analysis, etc.) in recent 10 years (Knutti 2010; Knutti et al. 2017; Riccio et al. 2012; Sanderson et al. 2015; Stein et al 2015; Mendlik and Gobiet 2016; Dalelane et al. 2018), which are proposed for weather forecasting, seasonal prediction or climate projection to make optimal use of the information inherent in the full ensemble, the DEFOPT does not essentially reduce ensemble size or discard any ensemble member, but only makes use of the information of the optimal ensemble members to determine whether a forecasting event occurs or not in a given region.

It is notable that the calculation of the optimal probabilistic threshold in DEFOPT may be slightly affected by the length of the hindcast period (for example, only 12 years

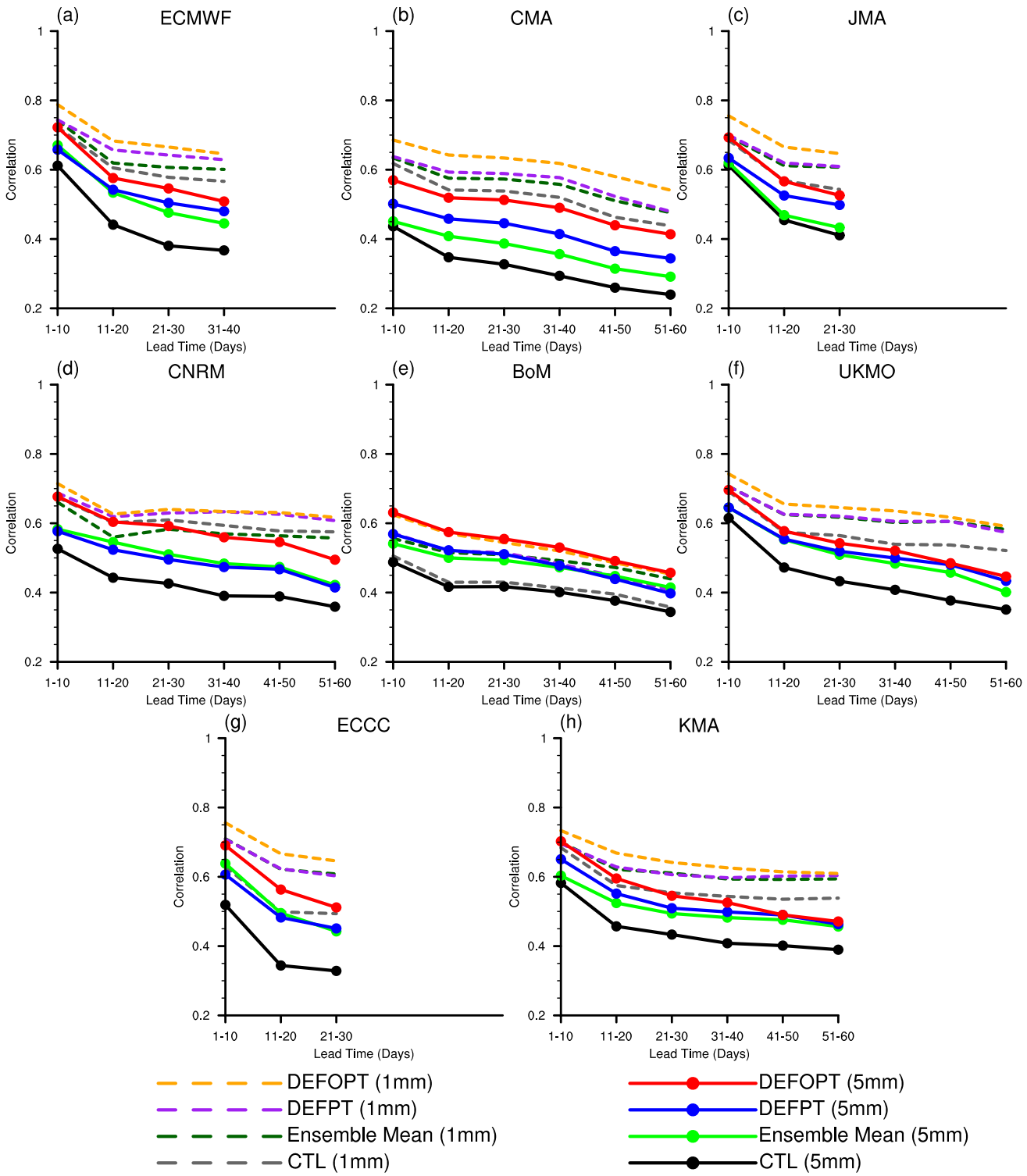

Fig. 11 Same as Fig. 10, but for the other re-forecast periods excluding 1999–2010

in this study). In addition, it is found that the α and β in Formula 3 are dependent on the model in the selection of the $P_{threshold}$ due to the different systematic forecasting biases in different models, and these coefficients can be considered to change with lead time, which will be investigated in our future work. There may be some potential application values of the DEFOPT for the multi-model ensemble. When the ensemble mean from each model is

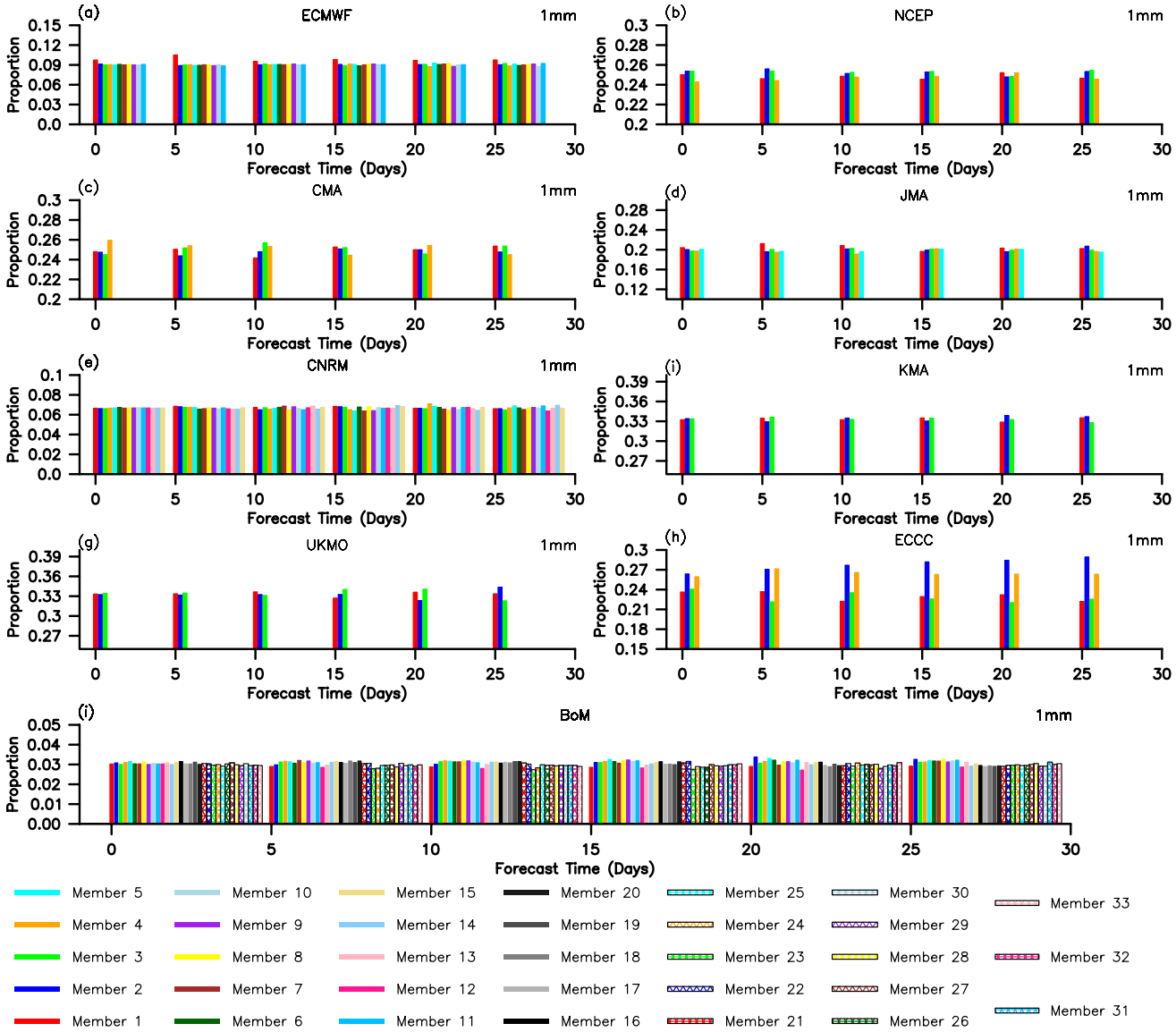

Fig. 12 The proportion of each ensemble member predicting the occurrence of ≥ 1 mm rainfall event in the ensemble members chosen by the DEFOPT in the S2S models in the different lead times at 5-days intervals during the summer of 1999–2010 over China

Table 4 2 × 2 matrix for a precipitation event with a certain threshold

	Rain observed	
	Yes	No
Rain forecast		
Yes	*a*	*b*
No	*c*	*d*

considered as an individual ensemble member, this method could determine how many models can be trusted in different regions and lead times to achieve a best performance of S2S multi-models precipitation prediction. Moreover, only the DEFOPT predictions for the summer precipitation over China has been investigated in this study, and its application to other areas, other seasons or other variables (air temperature, anomaly of height, etc.) can be evaluated in the future.

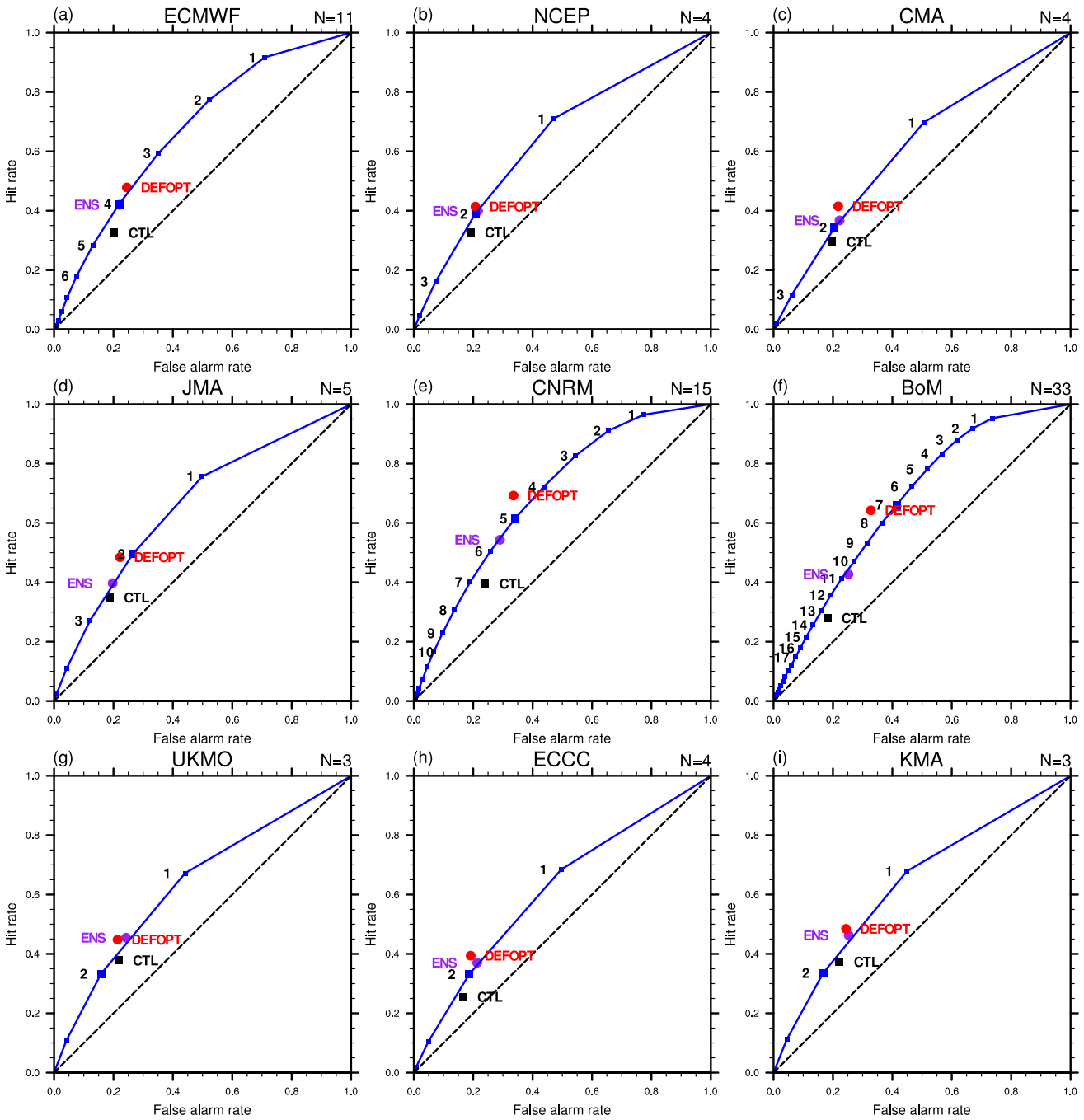

Fig. 13 The averaged Relative Operating Characteristic (ROC; Jolliffe and Stephenson 2003) curves of the predictions for the fourth pentad-averaged ≥ 5 mm precipitation during the summer of 1999–2010 over China. The black square indicates the CTL, the blue square larger than other blue squares means the DEFPT using a given probabilistic threshold, and the colored circles are the ENS and DEFOPT, respectively. The *N* is the total number of ensemble members and the numbers in subplot are the thresholds of ensemble member number

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00382-022-06623-4.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 41505094). We would like to thank Nanjing Hurricane Translation for reviewing the English language quality of this paper.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by WJ, TW, FV, XL, YL, JY and HZ. The first draft of the manuscript was written by WJ and TW, and FV commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by National Key Natural Science Foundation of China (Grant Number. 42230608) and National Natural Science Foundation of China (Grant Number. 41505094).

Data availability The datasets generated during the current study are available in the S2S project database center (http://www.s2sprediction.net/).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Accadia C, Mariani S, Casaioli M, Lavagnini A, Speranza A (2010) Verification of precipitation forecasts from two limited-area models over Italy and comparison with ECMWF forecasts using a resampling technique. Weather Forecasting 20:276–300

Bombardi RJ, Pegion KV, Kinter JL, Cash BA, Adams JM (2017) Subseasonal predictability of the onset and demise of the rainy season over monsoonal regions. Front Earth Sci 5(14):1–17

Buizza R (2008) The value of probabilistic prediction. Atmos Sci Lett 9:36–42

Buizza R (2019) Chapter 13—ensemble generation: the TIGGE and S2S ensembles. In: Robertson AW, Vitart F (eds) Sub-seasonal to seasonal prediction. Elsevier, pp 261–303

Buizza R, Hollingsworth A, Lalaurette F, Ghelli A (1999a) Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. Weather Forecasting 14:168–189

Buizza R, Miller M, Palmer TN (1999b) Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Q J R Meteorol Soc 125:2887–2908

Chessa PA, Lalaurette F (2001) Verification of the ECMWF ensemble prediction system forecasts: a study of large-scale patterns. Weather Forecasting 16:611–619

Chou JF (1989) Predictability of the Atmosphere. Adv Atmos Sci 6:335–346

Collins WD et al (2006) The Community Climate System Model version 3 (CCSM3). J Clim 19:2122–2143

Dalelane C, Frueh B, Steger C, Walter A (2018) A pragmatic approach to build a reduced regional climate projection ensemble for Germany using the euro-cordex 8.5 ensemble. J Appl Meteorol Climatol 57(3):477–491

Ebert EE (2001) Ability of a poor man's ensemble to predict the probability and distribution of precipitation. Mon Weather Rev 129:2461–2480

Fritsch JM, Hilliker J, Ross J, Vislocky RL (2000) Model consensus. Wea Forecasting 15:571–582

Gneiting T, Raftery AE (2005) Weather forecasting with ensemble methods. Science 310(5746):248–249

Hamill TM, Whitaker JS, Wei X (2004) Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. Mon Weather Rev 132:1434–1447

Hamill TM, Hagedorn R, Whitaker JS (2008) Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. Mon Weather Rev 136:2620–2632

Hanssen AW, Kuipers WJA (1965) On the relationship between the frequency of rain and various meteorological parameters. Meded Verh 81:2–15

Hoffman RN (2002) Controlling the global weather. Bull Am Meteorol Soc 83:241–248

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

Houtekamer PL, Mitchell HL (1998) Data assimilation using an ensemble Kalman filter technique. Mon Wea Rev 126:196–811

Hwang J, Orenstein P, Cohen J, Pfeiffer K, Mackey L (2019) Improving subseasonal forecasting in the Western U.S. with machine learning. In: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA, p 14. https://doi.org/10.1145/3292500.3330674

Jiang Z, Mu M, Wang D (2009) Ensemble prediction experiments using conditional nonlinear optimal perturbation. Sci China 52:511–518

Jie W, Wu T, Wang J, Li W, Liu X (2013) The improvement of 6–15 day precipitation forecast using a time-lagged ensemble method. Adv Atmos Sci 31:293–304

Jie W, Wu T, Wang J, Li W (2014) Using a deterministic time-lagged ensemble forecast with a probabilistic threshold for improving 6–15 day summer precipitation prediction in China. Atmos Res 156:142–159

Jie W, Vitart F, Wu T, Liu X (2017) Simulations of the Asian summer monsoon in the sub-seasonal to seasonal prediction project (S2S) database. Q J R Meteorol Soc 143(706):2282–2295

Jolliffe IT, Stephenson DB (2003) Forecast verification: a practitioner's guide in atmospheric science. Wiley, New York, p 66

Knutti R (2010) The end of model democracy? Clim Change 102:395–404

Knutti R, Sedláček J, Sanderson B, Lorenz R, Fischer E, Eyring V (2017) A climate model projection weighting scheme accounting for performance and interdependence. Geophys Res Lett 44:1909–1918

Krishnamurti TN, Kishtawal CM, Zhang Z, LaRow T, Bachiochi D, Williford E (2000) Multimodel ensemble forecasts for weather and seasonal climate. J Clim 13:4196–4216

Leith CE (1974) Theoretical skill of Monte Carlo forecasts. Mon Wea Rev 102:409–418

Li C, Gu W (2010) An analyzing study of the anomalous activity of blocking high over the Ural mountains in January 2008. Chin J Atmos Sci 34(5):865–874

Li W, Chen J, Li L, Chen H, Li X (2019) Evaluation and bias correction of s2s precipitation for hydrological extremes. J Hydrometeorol 20(9):1887–1906

Liang P, Lin H (2018) Sub-seasonal prediction over East Asia during boreal summer using the ECCC monthly forecasting system. Clim Dyn 50:1007–1022

Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141

Lorenz EN (1982) Atmospheric predictability experiments with a large numerical model. Tellus 34:505–513

Mendlik T, Gobiet A (2016) Selecting climate simulations for impact studies based on multivariate patterns of climate change. Clim Change 135:381–393

Molteni F, Buizza R, Palmeret TN (1996) The ECMWF ensemble prediction system: methodology and validation. Q J R Meteorol Soc 122:73–119

Moore AM, Kleeman R (1998) Skill assessment for ENSO using ensemble prediction. Q J R Meteorol Soc 124:557–584

National Climate Center, China Meteorological Administration (1998) The catastrophic flood and climate anomaly over China in 1998. Beijing Meteorological Press, p 139

Neal R, Fereday D, Crocker R, Comer RE (2016) A flexible approach to defining weather patterns and their application in weather forecasting over europe. Meteorol Appl 23:389–400

Pan J, van den Dool HM (1998) Extended-range probability forecasts based on dynamical model output. Weather Forecasting 13:983–996

Qin J, van den Dool HM (1996) Simple extensions of an NWP model. Mon Weather Rev 124:277–287

Riccio A, Ciaramella A, Giunta G, Galmarini S, Solazzo E, Potempski S (2012) On the systematic reduction of data complexity in multimodel atmospheric dispersion ensemble modeling. J Geophys Res 117:D05314

Saha S, van den Dool HM (1988) A measure of the practical limit of predictability. Mon Weather Rev 116:2522–2526

Sanderson B, Knutti R, Caldwell P (2015) A representative democracy to reduce interdependency in a multimodel ensemble. J Clim 28:5171–5194

Schaefer JT (1990) The critical success index as an indicator of warning skill. Weather Forecasting 5:570–575

Scheuerer M (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. Q J R Meteorol Soc 140(680):1086–1096

Schmeits JM, Kok KJ (2010) A Comparison between raw ensemble output, (Modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. Mon Weather Rev 138:4199–4211

Sivillo JK, Ahlquist JE, Toth Z (1997) An ensemble forecasting primer. Weather Forecasting 12:809–818

Stein AF, Ngan F, Draxler RR, Chai T (2015) Potential use of transport and dispersion model ensembles for forecasting applications. Weather Forecasting 30(3):639–655

Tan N, J Chen (2013) A study of ensemble perturbation method for 1–15 day prediction based on T213 model. Dissertation, Chinese Academy of Meteorological Sciences.

Toth Z, Kalany Y (1993) Ensemble forecasting at NMC: the generation of perturbations. Bull Am Meteorol Soc 74:2317–2330

Toth Z, Kalany Y (1997) Ensemble forecasting at NCEP: the breeding method. Mon Weather Rev 125:3297–3318

Vitart F, Molteni F (2009) Dynamical extended-range prediction of early monsoon rainfall over India. Mon Weather Rev 137:1480–1492

Vitart F, Ardilouze C, Bonet A, Brookshaw A, Chen M, Codorean C, Déqué M, Ferranti L, Fucile E, Fuentes M, Hendon H, Hodgson J, Kang H, Kumar A, Lin H, Liu G, Liu X, Malguzzi P, Mallas I, Manoussakis M, Mastrangelo D, MacLachlan C, McLean P, Minami A, Mladek R, Nakazawa T, Najm S, Nie Y, Rixen M, Robertson A, Ruti P, Sun C, Takaya Y, Tolstykh M, Venuti F, Waliser D, Woolnough S, Wu T, Won D, Xiao H, Zaripov R, Zhang L (2017) The sub-seasonal to seasonal prediction (S2S) project database. Bull Am Meteorol Soc 98:163–176

Weusthoff T, Ament F, Arpagaus M, Rotach MW (2010) Assessing the benefits of convection-permitting models by neighborhood verification: examples from MAP D-phase. Mon Wea Rev 138(9):3418–3433

Whitaker JS, Wei X, Vitart F (2006) Improving week-2 forecasts with multimodel reforecast ensembles. Mon Wea Rev 134:2279–2284

Wilks DS (1995) Statistical methods in the atmospheric sciences. Academic Press 59:1–464

Wu T, Li W, Ji J, Xin X, Li L, Wang Z, Zhang Y, Li J, Zhang F, Wei M, Shi X, Wu F, Zhang L, Chu M, Jie W, Liu Y, Wang F, Liu X, Li Q, Dong M, Liang X, Gao Y, Zhang J (2013) Global carbon budgets simulated by the Beijing climate center climate system model for the last century. J Geophys Res Atmos 118(10):4326–4347

Yang X (2001) The new development and outlook of the operational prediction system. Meteorol Mon 27(6):3–9

Zhang L, Kim T, Yang T, Hong Y, Zhu Q (2021) Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous U.S. J Hydrol 603:Part B. https://doi.org/10.1016/j.jhydrol.2021.127058

Zheng Z, Feng G, Huang J, Chou J (2012) Predictability-based extended-range ensemble prediction method and numerical experiments. Aeta Phys Sin 61(19):1–8

Zhou Y, Yang B, Chen H, Zhang Y, Huang A, La M (2019) Effects of the Madden–Julian Oscillation on 2-m air temperature prediction over China during boreal winter in the S2S database. Clim Dyn 52(11):1–19