Atmospheric Research 156 (2015) 142–159

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmos

Using a deterministic time-lagged ensemble forecast with a probabilistic threshold for improving 6–15 day summer precipitation prediction in China

Weihua Jie [a], Tongwen Wu [a,*], Jun Wang [b], Weijing Li [a], Thomas Polivka [b]

[a] Beijing Climate Center, China Meteorological Administration, Beijing, China
[b] Department of Earth and Atmospheric Sciences, University of NE-Lincoln, USA

ARTICLE INFO

Article history:
Received 24 January 2014
Received in revised form 15 December 2014
Accepted 11 January 2015
Available online 22 January 2015

Keywords:
Precipitation
6–15 day forecasts
Ensemble forecasting

ABSTRACT

A Deterministic Time-lagged Ensemble Forecast using a Probabilistic Threshold (DEFPT) method is suggested for improving summer 6–15 day categorical precipitation prediction in China from the Beijing Climate Center Atmospheric General Circulation Model version 2.1 (BCC_AGCM2.1). It is based on a time-lagged ensemble system that consists of 13 ensemble members separated sequentially at 6 hour intervals lagging the last three days. The DEFPT is not intended to predict the probability of rainfall, but rather to forecast rainfall (yes/no) occurrence for different categories of precipitation at any model grid box. A given categorical precipitation is forecasted to occur at one gridbox only when the ensemble probability for that categorical precipitation exceeds a certain threshold. This method is useful for providing an estimate of whether precipitation events will occur to decision-makers based on probabilistic forecasts during days 6–15. A large number of hindcast experiments for 1996–2005 summers reveal that this threshold can be best (and empirically) set as 5/13 and 4/13 respectively for the 6–15 day prediction of 1+ mm (i.e., above 1 mm per day) and 5+ mm rainfall events, using the Relative Operating Characteristic (ROC) curve, the Equitable Threat Score (ETS), the Hanssen and Kuipers (HK) score, and frequency bias (BIA) to achieve best prediction performance. With this set of thresholds, the DEFPT shows skill improvement over the corresponding single deterministic forecast using one initial value and the Time-Lagged Average Forecast (LAF) ensemble method. Similar improvements by the DEFPT are also found for the prediction of several other categories of precipitation between 1+ mm and 10+ mm per day. Application of DEFPT to larger ensemble size and BCC_AGCM version 2.2 with a higher horizontal resolution also demonstrates the effectiveness of the DEFPT for 6–15 day categorical precipitation forecasts.

1. Introduction

Precipitation is one of the most difficult variables to accurately predict when compared with the other weather fields routinely forecasted by operational numerical prediction models. Improving the accuracy of precipitation forecasts is one of the primary goals for weather prediction centers and is a major challenge facing the numerical weather prediction and climate research community (Ebert, 2001; Mullen and Buizza, 2001; Romatschke and Houze, 2011).

Over the last two decades, ensemble forecasting has been a popular technique for weather forecasting, seasonal prediction,

[*] Corresponding author at: National Climate Center, China Meteorological Administration, 46 Zhongguancun Nandajie, Beijing 100081, China. Tel.: +86 10 68406403; fax: +86 10 68406403.
E-mail address: twwu@cma.gov.cn (T. Wu).

http://dx.doi.org/10.1016/j.atmosres.2015.01.004

and even climate change studies in various operational centers (Sivillo et al., 1997; Krishnamurti et al., 2000; Martin et al., 2010; Vich et al., 2011). Based upon various ensemble systems using various modes, including breeding growing modes, singular-vector modes, and time-lagged modes, many previous studies (e.g., Du et al., 1997; Eckel and Walters, 1998; Buizza et al., 1999; Ebert, 2001; Mullen and Buizza, 2001; Hamill et al., 2004; Walser et al., 2004; Whitaker et al., 2006; Hohenegger and Schär, 2007; Lu et al., 2007; Tippett et al., 2007; McLay, 2008; Vitart and Molteni, 2009; Yuan et al., 2009) demonstrated the effectiveness of those ensemble methods for improving the ensemble mean and probability of precipitation prediction. However, these works are mainly focused on the usage of ensemble predictions for daily precipitation within a week and pentad or weekly averages of precipitation within one month. Improving the 1–2 week forecasts of daily precipitation is still needed.

It is well-known that forecasting whether a rainfall event of concern will occur in a deterministic or probabilistic way is quite useful for the enterprise, government, and agricultural decision-makers (Katz and Murphy, 1997; Lee and Lee, 2007). As for the traditional Probabilistic Quantitative Precipitation Forecast (PQPF), many works (e.g., Eckel and Walters, 1998; Buizza et al., 1999; Hamill et al., 2008) have shown its limitation for 6–15 day precipitation events since the corresponding skill generally decreases significantly after 6 days; this is because it lacks the sharpness to discriminate which events occurred and which events did not in this period. In this paper, another forecast type is used to provide a deterministic (yes/no) forecast for different precipitation categories from the ensemble probability forecasts by using optimal probabilistic thresholds.

Indeed, our recent study (Jie et al., 2013) has also attempted to improve the 6–15 day precipitation forecasts of the Beijing Climate Center Atmospheric General Circulation Model (BCC_AGCM) using a time-lagged ensemble mean technique (hereafter Time-Lagged Average Forecast, LAF). It provided a deterministic forecast for separate precipitation categories depending on the rainfall intensity of the ensemble mean. Although the LAF is more effective than the single deterministic forecast using only one initial value (hereafter, single forecast), it may lead to forecast error for the low threshold precipitation (see details in Section 3 and Jie et al. (2013)). In this work, we suggest another Deterministic Ensemble Forecast method using a Probabilistic Threshold (hereafter DEFPT, see details in Section 3) replacing the ensemble mean.

When compared to the LAF and the single forecast, the effectiveness of the DEFPT for 6–15 day daily precipitation forecasts during China's summer season is evaluated using the Equitable Threat Score (ETS, e.g., Schaefer, 1990), the Hanssen and Kuipers score (HK, e.g., Hanssen and Kuipers, 1965), the frequency bias (BIA, e.g., Wilks, 1995), and the Root-Mean-Square Error (RMSE).

The model and data used in this study are described in Section 2. The design of the experiment, the forecast, and the evaluation methods are presented in Section 3. Section 4 mainly shows the validation results of the DEFPT as compared to the LAF and the single forecast. Section 5 presents the conclusion and discussion.

2. Model and data

The Beijing Climate Center Atmospheric General Circulation Model (BCC_AGCM) is a global spectral model based on the National Center for Atmospheric Research (NCAR) Community Atmosphere Model and subsequently developed by the National Climate Center at the China Meteorological Administration (Wu et al., 2008, 2010; Wu, 2012). The dynamical core of the model is described in Wu et al. (2008). A precedent version, BCC_AGCM2.0, is detailed in Wu et al. (2010). Most of the physical processes in BCC_AGCM2.0 are from CAM3, developed by NCAR, and a few new schemes are implemented including parameterizations for the deep cumulus convection, dry adiabatic adjustment, snow-cover fraction, and latent/sensible heat fluxes over the ocean surface (Wu et al., 2010). Previous studies have shown that BCC_AGCM2.0 at a T42 resolution reproduces the present-day climate well (Wu et al., 2010), the heavy precipitation events in the summer of 1998 over east China (Jie and Wu, 2010), the intra-seasonal oscillation of 850-mb wind in the tropics (Dong et al., 2009), the Asian–Australian Monsoon inter-annual variability (Wang et al., 2009), and the inter-decadal changes of rainfall, temperature, and circulation in east Asia (Chen et al., 2011).

BCC_AGCM2.1 and BCC_AGCM2.2 used in this work are the atmospheric component in the Beijing Climate Center Climate System Model (BCC_CSM1.1, Wu et al., 2013), and they include the updated versions of BCC_AGCM2.0 with a new "deep penetrative convection" scheme suggested by Wu (2012). BCC_AGCM2.1 has a horizontal resolution of T42 (approximately 2.8125° × 2.8125° transformed grid) and 26 levels in a hybrid sigma/pressure vertical coordinate system. BCC_AGCM2.2 has a higher horizontal resolution of T106 (approximately 1.125° × 1.125° transformed grid) and has been used for the Atmospheric Model Intercomparison Project (AMIP) and short-term climate operational prediction.

Due to the lack of data assimilation systems for BCC_AGCM, the model's initial values are generated by using the Initial Coordinated Integration Method (ICIM) considered as a simple data assimilation procedure for the hindcast simulations (Jie and Wu, 2010). In ICIM, we spin up the model for a period of 10 days by using the atmospheric temperature and the wind fields from NCEP-II reanalysis data (horizontal resolution of 2.5° × 2.5°and 17 levels), which are interpolated to the model grid. During the spin-up, the model boundary conditions of sea surface

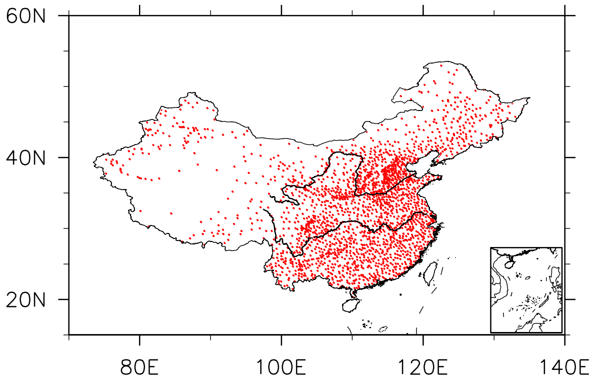

Fig. 1. The locations of 2466 rain gauge stations over China.

temperature and sea ice are specified according to the NCEP-OI2 data (1981–2006) (http://www-pcmdi.llnl.gov/projects/amip/AMIP2EXPDSN/BCS/amipbc_dwnld_files/T42/nc/).

The observed precipitation data are based on reports from 2466 rain gauge stations (Fig. 1) and have undergone quality control processes such as climate and gauge outlier processing and homogenization conducted by the Chinese National Meteorological Information Center. In this work, these precipitation data are interpolated to model grids by the Cressman interpolation method (Cressman, 1959) in which the 1°, 2° and 5° latitudes are used successively as scanning radii, and the initial guess field is set to zero. This interpolation process may generate some missing values in the regions of low-density rain gauges over western China. To avoid the effect of missing data, we only verified the model forecasts at the corresponding grids that have sufficient valid observation data.

3. Methodology and experiment design

In this work, the time-lagged ensemble forecast system is constructed from all forecasts valid for the same time, but initialized at different lagging times for every 6 h. Thus, for example, a 3-day ensemble forecast system can render a total of 13 ensemble members, including the one from the single forecast. For this study, forecasts within a lead time of 30 days from each ensemble member are archived.

The Deterministic Ensemble Forecast using a Probabilistic Threshold (DEFPT) method is established based upon this time-lagged ensemble system. It is not a probability forecast, but rather a deterministic categorical forecast of precipitation intensity at any model grid box. A given precipitation category is forecasted only when the ensemble probability for the category to occur at one grid box exceeds a certain threshold. The DEFPT is defined as:

$$A_{DEFPT} = \left\{ \begin{array}{l} 1,\ \ if \sum_{i=1}^{n} \alpha_i \geq N_{threshold} \\ 0,\ \ if \sum_{i=1}^{n} \alpha_i < N_{threshold} \end{array} \right\}, \qquad (1)$$

where A_{DEFPT} shows yes/no occurrence of rainfall for a certain given category such as 1 mm and above per day (hereafter 1+ mm) at a model grid box, α_i denotes the occurrence (equal to 1) or nonoccurrence (equal to 0) of the predicted rainfall event for the ith of total n forecast members, and $N_{threshold}$ is the threshold number of all members to forecast occurrences of precipitation for a given category. The value of $N_{threshold}$ depends on the category of precipitation and the performance of BCC_AGCM model. The optimal probabilistic threshold ($N_{threshold}/n$) of the DEFPT is chosen based on the condition that the corresponding ETS and HK scores are relatively higher than that for the other probabilistic thresholds and that the corresponding frequency biases (BIA) are not far from 1.0.

The DEFPT may overcome the LAF's disadvantages. The LAF rainfall at a certain gridbox depends on the rainfall intensity of each LAF member. It can be expressed as

$$\overline{A_{LAF}} = \frac{A_1 + A_2 + \ldots + A_n}{n}, \qquad (2)$$

where A_i is the amount of predicted rainfall for the ith ensemble member, and n is the number of ensemble members. Following our previous work (Jie et al., 2013), the $\overline{A_{LAF}}$ is also classified into different precipitation categories in a verification procedure. Although a six hour time-lagged LAF using the five most recent members is more effective than the single forecast and other LAF schemes, it is still easy to forecast inaccurately the occurrence of low-intensity rainfall. This primarily occurs if

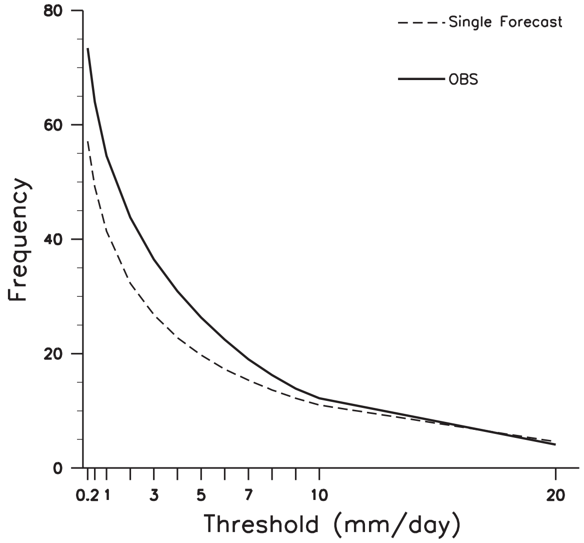

Fig. 2. The averaged frequency of daily precipitation amounts over China as a function of threshold from 0.2+ mm to 20+ mm per day at lead times of 6–15 days during the summer of 1998. The solid line is from the observed precipitation, and the dashed black line is from the single forecast. In units of percentages.

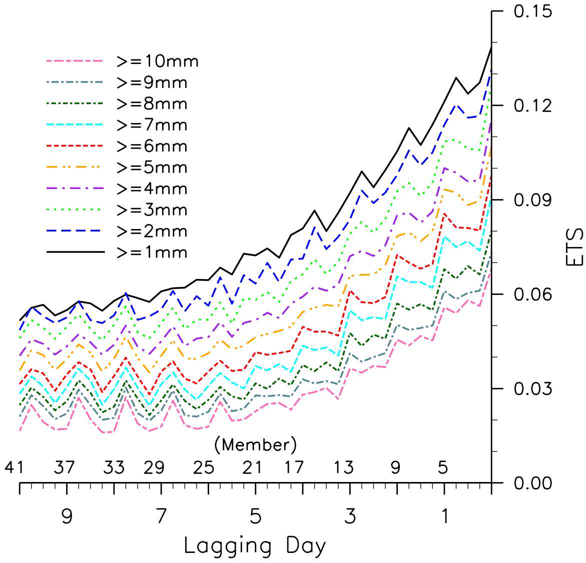

Fig. 3. The averaged ETS scores for the 6–15 day forecasts for the different precipitation categories as a function of ensemble member lagged from the 1st to 10th days (i.e., from the 1st to 41st members) during the 1998 summer over China.

some members with excessive or minor rainfall amounts appear as the lead time increases (Jie et al., 2013). For example, if one member ($i = 1$) from total five members ($n = 5$) predicts a 15 mm rainfall occurrence ($A_1 = 15$ mm), but the other members do not predict any rainfall ($A_{2–5} = 0$ mm), the final result ($\overline{A_{LAF}} = 3$ mm) will lead to the occurrence of 3+ mm categorical rainfall at this gridbox. This can lead to errors, especially when there was no observed rainfall event. Similarly, if two members ($i = 1, 2$) forecast no rainfall event ($A_{1–2} = 0$ mm) but the other members predict rainfall higher than 7 mm ($A_{3–5} = 7.1$, 7.8 and 8.1 mm), the ensemble mean ($\overline{A_{LAF}} = 4.6$ mm) will belong to the 4+ mm categorical rainfall; this ensemble mean is not representative of the majority of members. The DEFPT, however, forecasts the occurrence of a low-intensity of rainfall event only when the number of ensemble members predicted this occurrence at a given location exceeds the threshold number $N_{threshold}$, regardless of the excessive or miniscule rainfall amounts predicted from one ensemble member.

Due to the importance of summer precipitation prediction over China (e.g., Ding and Hu, 2003; Fan et al., 2008; Liu and Fan, 2014), we have conducted a large number of numerical experiments using BCC_AGCM2.1 at the T42 resolution for this season. 92 hindcast cases (using the DEFPT method), initiated respectively at 00 UTC in each day during the 1st June to the 31st August, 1998 are analyzed. The selection of the number of ensemble members is discussed in Section 4.2. 13 ensemble time-lagged runs (separated 6 h) in each case are mainly selected for this work. In the summer of 1998, there are several typical persistent precipitation events in central China (Ding and Hu, 2003). In order to test the effectiveness of the DEFPT method in other years, 27 additional hindcast cases are investigated, with initial times respectively of 00 UTC for 1 June, 1 July and 1 August in 9 summers of 1996–1997 and 1999–2005. In this work, we also tested the influence of different horizontal resolutions on the effectiveness of the DEFPT, and 30 cases resembling to those described above (i.e., initiated at 00 UTC on the 1st June, the 1st July and the 1st August in the 10 summers of 1996–2005) are conducted by applying the DEFPT to a higher resolution version of BCC_AGCM2.2.

In the following section, the discriminating skill and reliability of each probability bin generated from ensemble members are evaluated using an attribute diagram (Murphy, 1973) and Relative Operating Characteristic curve (ROC, e.g., Mason, 1982; Harvey et al., 1992; Jolliffe and Stephenson, 2003). The Equitable Threat Score (ETS, e.g., Schaefer, 1990) is also utilized to evaluate the skill of rainfall event prediction minus the random forecast skill, the Hanssen and Kuipers score (HK, e.g., Hanssen and Kuipers, 1965) is used to verify the accuracy of both events and nonevents, while the frequency bias (BIA, e.g., Wilks, 1995) is used to evaluate the bias of rainfall frequency; the rank histograms (RHs) (Hamill and Colucci, 1998; Hamill, 2001) are used to assess the spread of ensemble members. The Root-Mean-Square Error (RMSE) is applied to measure forecast error.

4 . Results

4.1 . The selection of precipitation category

Fig. 2 shows the observed frequencies of daily precipitation in the summer of 1998 for different categories including 0.2+ mm, 0.5+ mm, 1+ to 10+ mm (separated by 1 mm intervals), and 20+ mm per day. They are calculated using daily precipitation data from rain gauge observations and averaged for the time corresponding to 92 cases at lead times of 6–15 days. The 1+ mm, 5+ mm and 10+ mm per day rainfall frequency categories (solid line in Fig. 2) from the averaged grid points across China account for more than 55%, 25% and 10%, respectively. The following analyses will mainly focus on these categories of 1+ mm to 10+ mm per day.

In Fig. 2, we also see that single forecasts from the BCC_AGCM2.1 at lead times of 6-15 days underestimate the frequency of the observed rainfall during the summer of 1998. It denotes the existence of a dry bias in the BCC_AGCM2.1.

4.2 . The selection of ensemble members

To explore what the longest useful lag time would be for 6–15 day forecasts, we calculated the ETS scores for each member forecast (lagging from 1 day up to 10 days) at lead times of 6–15 days for the various precipitation categories of the summer of 1998. Fig. 3 indicates that the ETS scores for all these categories decrease rapidly and then gradually stabilize after 6 days (including 25 members at 6-hour intervals). The decrease in the ETS score might be caused by the increase of the real forecast error for earlier members (Dalcher et al., 1988). The score within the last three day lags (including 13 members) for each precipitation category is generally higher

Table 1
The averaged precipitation forecast errors correlation matrix of 13 members at lead times of 6–15 days.

Number	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	0.489	0.469	0.477	0.460	0.445	0.449	0.452	0.440	0.444	0.442	0.448	0.429
2		1	0.497	0.483	0.467	0.459	0.463	0.451	0.438	0.456	0.444	0.445	0.446
3			1	0.511	0.489	0.479	0.472	0.456	0.450	0.458	0.451	0.452	0.447
4				1	0.492	0.486	0.481	0.467	0.461	0.462	0.469	0.456	0.452
5					1	0.511	0.484	0.490	0.470	0.456	0.458	0.457	0.445
6						1	0.515	0.495	0.474	0.469	0.476	0.456	0.443
7							1	0.531	0.501	0.494	0.483	0.463	0.456
8								1	0.509	0.500	0.492	0.476	0.467
9									1	0.533	0.503	0.503	0.489
10										1	0.535	0.513	0.482
11											1	0.558	0.519
12												1	0.530
13													1

than the corresponding averaged ETS value during 10-lagged days. In addition, our previous work (Jie et al., 2013) has demonstrated that the time-lagged members within the last three days can contain useful information for the 6–15 day ensemble mean. Thus, 13 ensemble members with 6-hour intervals within lagging times of three days are mainly used in this study and 25 members are further tested in Section 4.8.

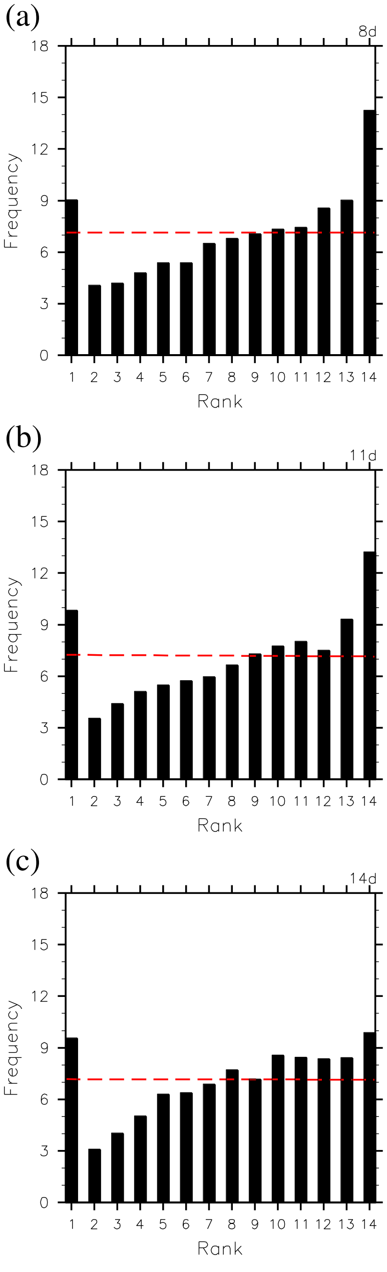

Fig. 4. The rank histograms (RHs) of time-lagged ensemble members lagged three days (with 6-hour time-lagged intervals) for the 8th (a), 11th (b) and 14th (c) day precipitation forecasts using a T42 resolution. The histograms are based upon the results of 30 cases during 1996 to 2005. The abscissa indicates the rank of the observation among all ensemble members. The ordinate indicates the frequency of the total sample for each rank. The black line denotes an averaged rank. In units of percentages.

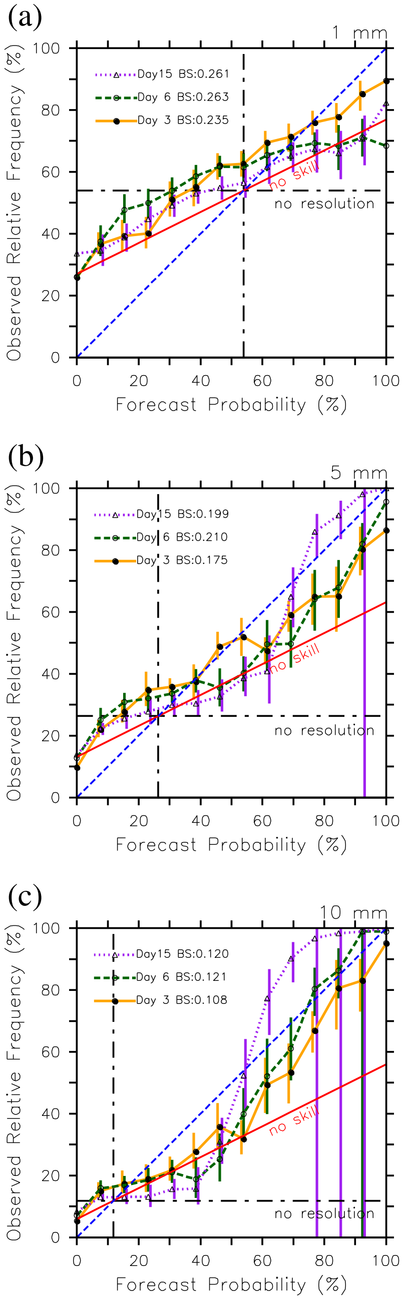

Fig. 5. The attribute diagrams of the PQPF of (a) 1+ mm, (b) 5+ mm and (c) 10+ mm per day rainfall for 14 probability bins (e.g., 0–1/13, 1/13–2/13, ... or 12/13–1, etc.) from 92 cases at lead times of 3 days, 6 days and 15 days during the 1998 summer in China. The colored curves with different marks respectively denote the results at the different lead times. The short lines plotted on each curve are the 95% bootstrap CIs. The blue dashed line represents perfect reliability (i.e., forecasted probability = observed relative frequency). The "no resolution" line is the climatic frequency of observed rainfall. The red line is a criterion for yes/no skill of PQPF.

Table 1 lists the correlation matrix of forecast errors among all the 13 members for the 6–15 day forecasts of precipitation over China. All of these correlations are significant at the 1% level. It indicates that the 13 members are, to some extent, independent to each other and their correlations of forecast errors are between 0.429 and 0.558. To further examine the spread of 13 ensemble members within 3-lagged days, the rank histograms of the ensemble members are calculated from 30 hindcast cases that are separately initiated at 00 UTC on the 1st June, the 1st July and the 1st August in the 10 summers (1996–2005). Fig. 4a shows the rank histogram for the 8th day forecast and the black dashed line denotes the value of 1 divided by the number of the total ranks. The 13 ensemble members present a reverse L-shaped RH with sloped distributions of the rank histograms toward one side. Based on the work of Yuan et al. (2009), this shape of RH reflects that these ensemble members have insufficient variability and dry biases in the model. The frequency in the highest rank almost doubles the perfect rank histogram (black dashed line), indicating that the intensity of rainfall is under-forecasted by almost all the ensemble members. But, as the lead time increases, the ensemble forecasts for both the 11th day and 14th day become more disperse and the corresponding RHs are relatively uniform (Fig. 4b–c), although the slightly higher frequencies in rank 1 and rank 14 suggest the ensemble spread is still not large enough. There are four possible reasons accounting for this slightly insufficient variability of the time-lagged ensemble system: (i) the time-lagged ensemble forecast system is only a single-model ensemble system: it does not account well for the uncertainty of the model processes; (ii) there is no breeding cycle (as in the NCEP Short-Range Ensemble Forecast system) and no maximization procedure cycle (like the singular-vector method used in the ECMWF); (iii) the description of the initial uncertainties are limited to the number of ensemble members over the past three days (Lu et al.,

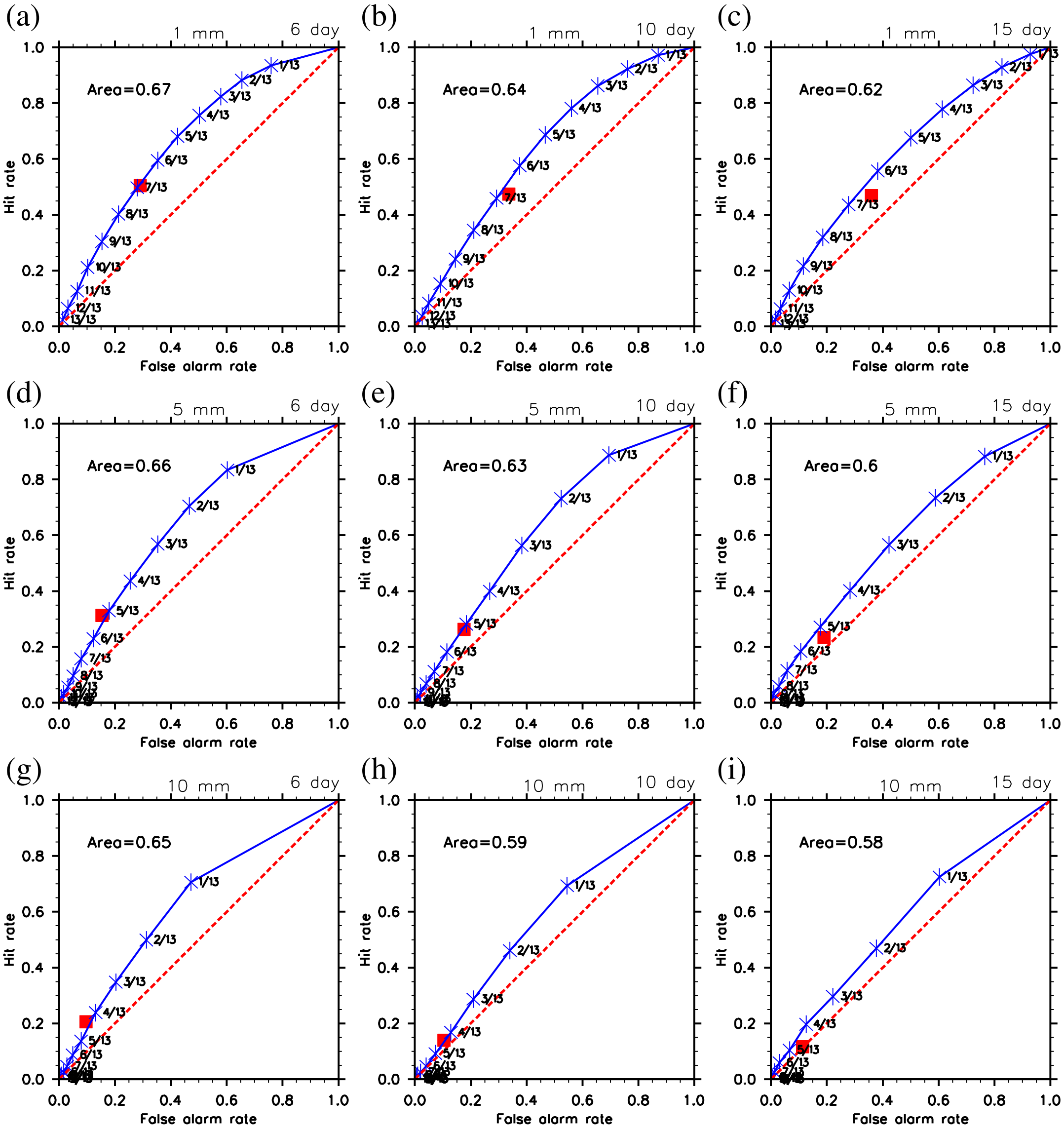

Fig. 6. The averaged ROC curves of the DEFPT using different probabilistic thresholds for (a–c) 1 + mm, (d–f) 5 + mm and (g–i) 10 + mm per day rainfall at lead times of (left panel) 6 days, (middle panel) 10 days and (right penal) 15 days during the 1998 summer in China. The ROC area is shown on the top-left corner and the black square indicates the single forecast.

2007); (iv) there are dry biases in the BCC_AGCM model simulation of high intensity rainfall events, but wet biases in that of low-intensity rainfall events and nonevents.

4.3 . The evaluation for all probability forecasts

First, the traditional PQPF based upon the 13 ensemble members (with ΔT = 6 h) within 3-lagged days is evaluated by using an attribute diagram (see details in Appendix 5). We divided all the forecast probabilities generated from 13 ensemble members into 14 probability bins (mutually exclusive) as 0/13–1/13, 1/13–2/13, …, 12/13–13/13, and 13/13 (hereafter [0]/13; [1]/13; [2]/13 etc.) for each precipitation category, and then we compare the number of rainfall events contained in each probability bin with the observed counterparts (Murphy and Winkler, 1987).

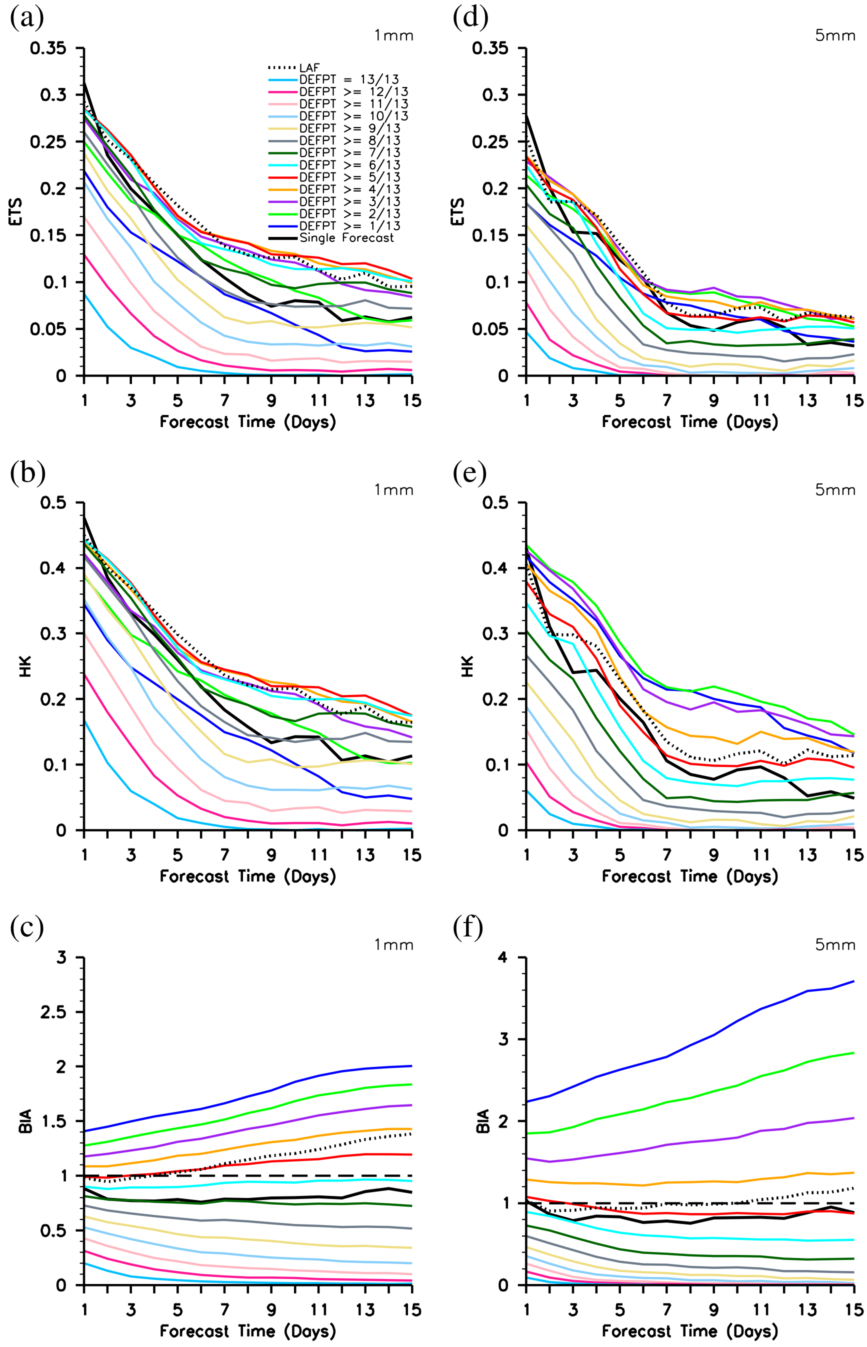

Fig. 7. Evaluation of LAF-based and DEFPT-based rainfall prediction with a time-lagged interval of 6 h. The averaged ETS scores for prediction of 1 + mm and 5 + mm per day rainfall as a function of forecast length for up to 15 days during June, July, and August 1998 in China are respectively shown as panels (a) and (d) in the first row. The second and third rows are respectively the same as the first row except for showing the HK (b, e) and BIA scores (c, f).

Fig. 5 shows the attribute diagram of the 14 forecasted probability bins (x-axis) and their corresponding observed relative frequencies (y-axis) for 1 + mm, 5 + mm and 10 + mm per day rainfall from 92 cases at lead times of 3 days, 6 days, and 15 days during the 1998 summer in China. The 95% bootstrap Confidence Intervals (CIs) suggested by Efron (1979, 1981) are used to reveal the sampling variability in each probability bin (short lines plotted on the reliability curves in Fig. 5). In this bootstrap procedure, the sub-sample in each bin based on 92 cases was randomly resampled 1000 times. On the 3rd day, as shown in Fig. 5, PQPF is skillful in some probability bins within [7–13]/13 for the 1 + mm per day rainfall, [4–13]/13 for the 5 + mm per day rainfall, and [2–13]/13 for the 10 + mm per day rainfall. This is evidenced by the corresponding orange lines with filled circles (i.e., reliability curve) belonging to the positive skill area, although part of bins are relatively unreliable (far away from the perfect reliability 1:1 line) such as [8]/13 and [11]/13 for 5 + mm per day and [7]/13 for 10 + mm per day. The positive skill area is the region where the reliability curve is above (below) the no skill line when the forecasted probability is larger (less) than the climatic frequency of observed rainfall. But for other probability bins, the forecasted probability of precipitation is obviously less than the observed relative frequency and out of the positive skill area. Overall, the BS (including the reliability, resolution, uncertainty terms) of the traditional PQPF for 1 + mm, 5 + mm and 10 + mm per day are 0.235, 0.175 and 0.108 on the 3rd day, respectively.

Beyond 5 days, these BS values generally increase and stabilize around 0.26, 0.2, and 0.12 (the higher the BS, the

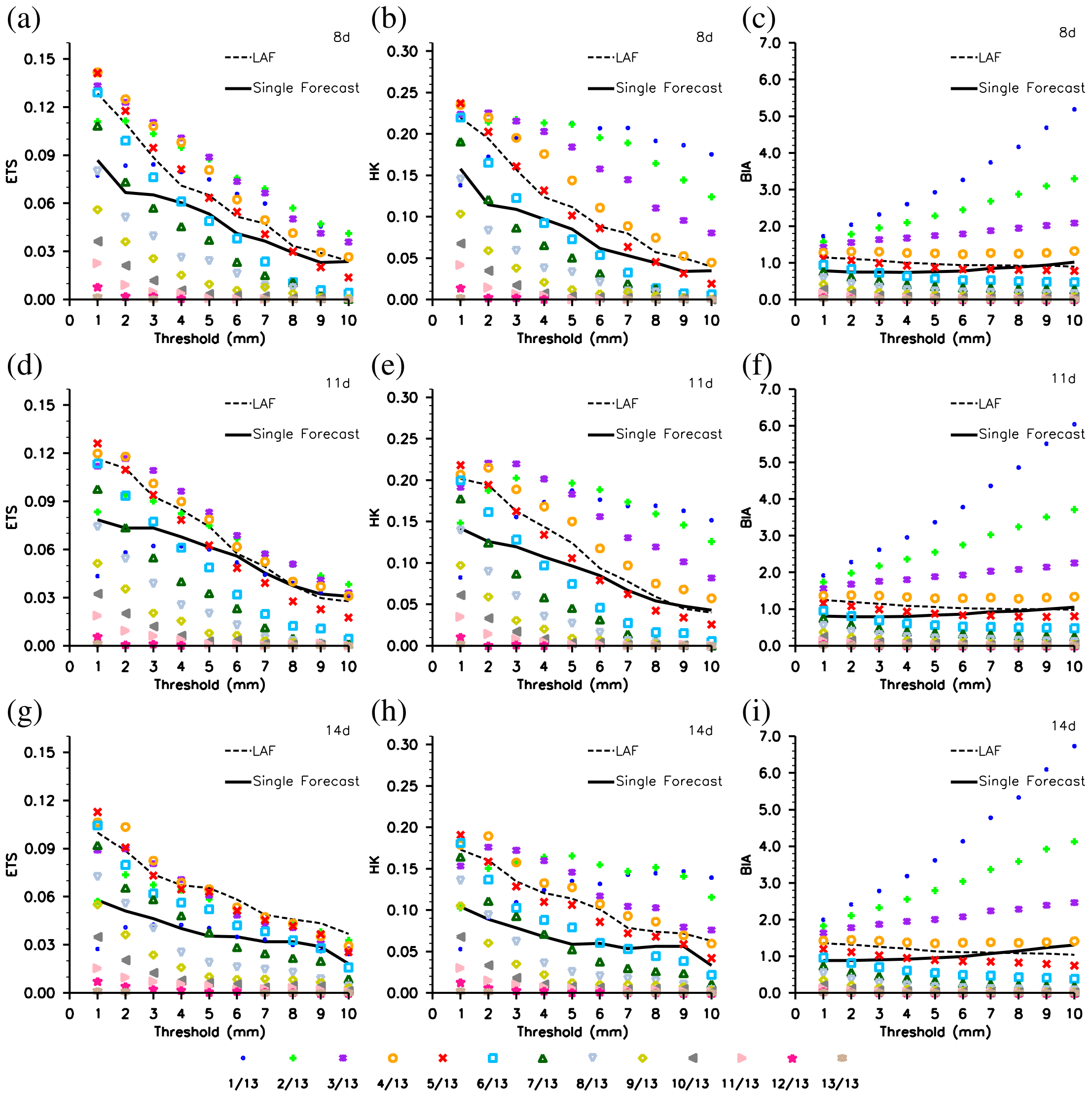

Fig. 8. The (a) ETS, (b) HK and (c) BIA scores (y-axis) for the 8th forecast of precipitation as a function of the categories (from 1 + mm up to 10 + mm per day) of daily rainfall amounts (x-axis in mm). The black solid line shows the evaluation score for the single forecast, while the black dashed line is the LAF using five ensemble members at 6-hour time-lagged intervals. The colored dots respectively show the evaluation scores for the forecasts using the DEFPT method with different probabilistic thresholds (e.g., 1/13–13/13 color coded according the legend at the bottom of each panel). The second and third rows are respectively the same as the first row except for showing the 11th day (d–f) and 14th day (h–i).

worse the forecast). But as the lead time increases, some probability bins (green dashed lines with hollow circles, representing the 6th day, and purple dashed lines with triangles, representing the 15th day in Fig. 5) are still located in the positive skill area when the forecasted probability bins exceed the “no resolution” line, except the [5–8]/13 bins for the 5 + mm per day rainfall, and the [3–6]/13 bins for the 10 + mm per day rainfall. It implies that there is still information available for precipitation forecasts longer than 5 days using probabilities.

The 95% bootstrap CIs plotted over the corresponding reliability curves are much larger for the high probability bins and also generally increase with lead time. In particular, the CIs for high probabilities with the 5 + mm and 10 + mm per day precipitation forecasts are quite large after 5 days. It indicates that sampling variability is involved in the high probability bins, although the corresponding reliability curves belong to the positive skill area in the attribute diagram.

4.4 . Case Study of DEFPT method for the summer of 1998

In the following sections, we mainly discuss the usage of probabilistic thresholds from 1/13 to 13/13 for a certain categorical precipitation forecast. Here, the probabilistic thresholds are different from the probability bins. All the probability bins are mutually exclusive to each other, but each probabilistic threshold includes all the probability bins larger than a certain value (i.e., $\geq k/13$; where k is the number of forecasted occurrences members at a certain grid box).

Fig. 6a–i shows the ROC curves, which reveal the differences in discriminating skill between rainfall events and nonevents for different probabilistic thresholds. We found that the 5/13 to 6/13 thresholds for 1 + mm per day and the 2/13 to 4/13 thresholds for 5 + mm and 10 + mm per day categorical precipitation have higher hit rates in contrast to the single forecasts (black square) as well as larger *hit-rate/false-alarm-rate* ratios than other thresholds with higher hit rates (e.g., 1/13). In the following sections, the DEFPT method defined by the different probabilistic thresholds is examined.

Fig. 7a shows the ETS scores of DEFPT using different ensemble probabilistic threshold values varying from 1/13 to 13/13 to predict the 6–15 day precipitation of 1 + mm per day by category, as well as their comparisons with those from the LAF and the single forecast. Quantitatively, the ETS score for the single forecast (solid black line in Fig. 7a) decreases from about 0.3 to 0.15 during the first 5 days, and thereafter gradually declines to about 0.07 for a lead time beyond 5 days (Fig. 7a). In comparison, the ETS scores for the DEFPT (solid color lines in Fig. 7a) using thresholds in 4/13–5/13 range are larger than 0.1 and are enhanced significantly by 0.05 for the forecasts beyond 5 days, and these values are slightly higher than the values obtained with LAF (dashed black line in Fig. 7a) after 6 days.

This enhancement from using the DEFPT is also supported by the analysis of HK score that accounts for the forecast accuracy of rainfall events and nonevents. Fig. 7b shows that the HK score for single forecast (solid black line) beyond 5 days is less than 0.2 and continues to decrease with longer lead time. In contrast, its counterparts for DEFPT (solid color lines) with 4/13 to 5/13 thresholds are persistently larger than 0.2 beyond 5 days, and they are also slightly higher than that for the LAF (dashed black line) as a whole (Fig. 7b).

In Fig. 7a and b, we also note that the DEFPT with thresholds less than 20% (e.g., 1/13–2/13) or larger than 50% (e.g., 7/13–13/13) shows no improvement when compared to the DEFPT using thresholds in 4/13–5/13 range. It may be caused by the lack of the sharpness in ensemble system to discriminate among rainfall events and nonevents as the lead time increases (as shown in Fig. 6a–c). This is also supported by the evaluation of BIA scores (Fig. 7c), which presents an overestimation from the forecasts with low probabilistic thresholds (BIA scores of 1.5–2.0) and an underestimation from those with high probabilistic thresholds (BIA scores of 0.0–0.5). In addition, the BIA scores for DEFPT with thresholds of 5/13 or 6/13 are generally closer to 1.0 than that for the single forecast (a persistent underestimation with BIA scores of 0.8–0.9) and the LAF (an overestimation with the scores of 1.1–1.4 beyond 6 days) (Fig. 7c). This further suggests that the disadvantage of the LAF for the low threshold precipitation can be reduced. Overall, the 5/13 threshold can be regarded as an optimal threshold value for the DEFPT to predict 6–15 day 1 + mm per day categorical precipitation in the region of this study.

The prediction of heavier precipitation using the DEFPT method is also evaluated. Fig. 7d–f shows the ETS, HK and BIA scores of 5 + mm per day rainfall prediction. Similar to the 1 + mm per day categorical rainfall prediction, the 6–15 day forecasts' skill for 5 + mm per day categorical rainfall is enhanced by the DEFPT, especially with 2/13–4/13 ensemble probabilistic thresholds. The resultant ETS scores are close to 0.1 and are higher than the single forecast and the LAF (Fig. 7d). The corresponding HK scores with 2/13–4/13 thresholds are also higher than the single forecast by about 0.1 beyond 5 days, which indicates that the forecast accuracy for precipitation events and nonevents can be improved significantly with

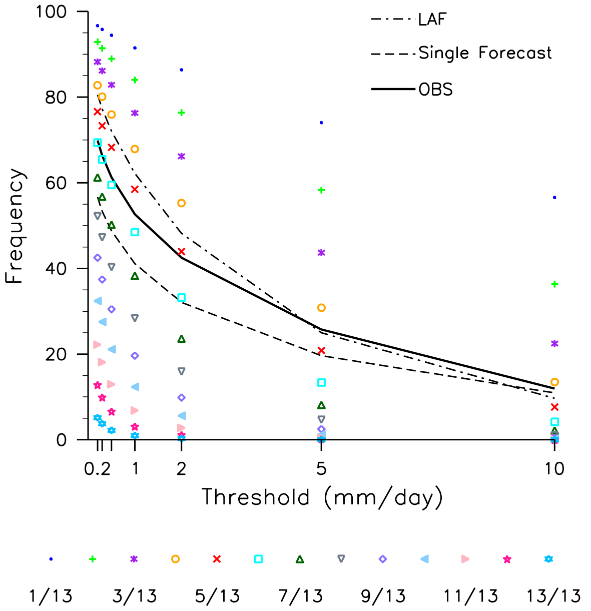

Fig. 9. The averaged frequency of daily precipitation amounts over China as a function of threshold from 0.2 mm to 10 mm at lead times of 6–15 days during the 1998 summer. The solid line is from the observed precipitation, the dashed black lines are from the single forecast and the ensemble mean, and dots are from the DEFPT (e.g., 1/13–13/13 coded according the legend at the bottom of each panel). In units of percentages.

DEFPT using 2/13–4/13 thresholds (Fig. 7e). As for the high probabilistic thresholds (6/13–13/13), the ETS and HK scores are not enhanced. In terms of BIA evaluation (Fig. 7f), the result with 4/13 threshold has relatively smaller bias than that with 2/13 or 3/13 thresholds. As a whole, the 4/13 threshold is a good selection for the DEFPT to forecast the 5+ mm per day categorical precipitation over China.

Furthermore, other category precipitation forecasts from 1+ mm to 10+ mm per day are verified. These evaluations are plotted in Fig. 8 in terms of ETS, HK, and BIA scores for the 8th, the 11th and the 14th days as a function of rainfall category. Fig. 8a–b indicates that the DEFPT, for a certain probabilistic threshold, generally improves the 8th day forecast skill for all the precipitation categories from 1+ to 10+ mm (in comparison to the single forecast and the LAF), because the corresponding ETS and HK scores are often higher than those of the single forecast and the LAF. It also shows that the optimal thresholds for the DEFPT method to achieve the best results slightly decreases as the precipitation category increases (e.g., 5/13 for 1+ mm; 4/13 for 2+ mm; 3/13 for 5+ mm; 2/13 for 10+ mm per day). The BIA scores for the single forecast, the LAF, and the DEFPT using the 4/13–5/13 probabilistic

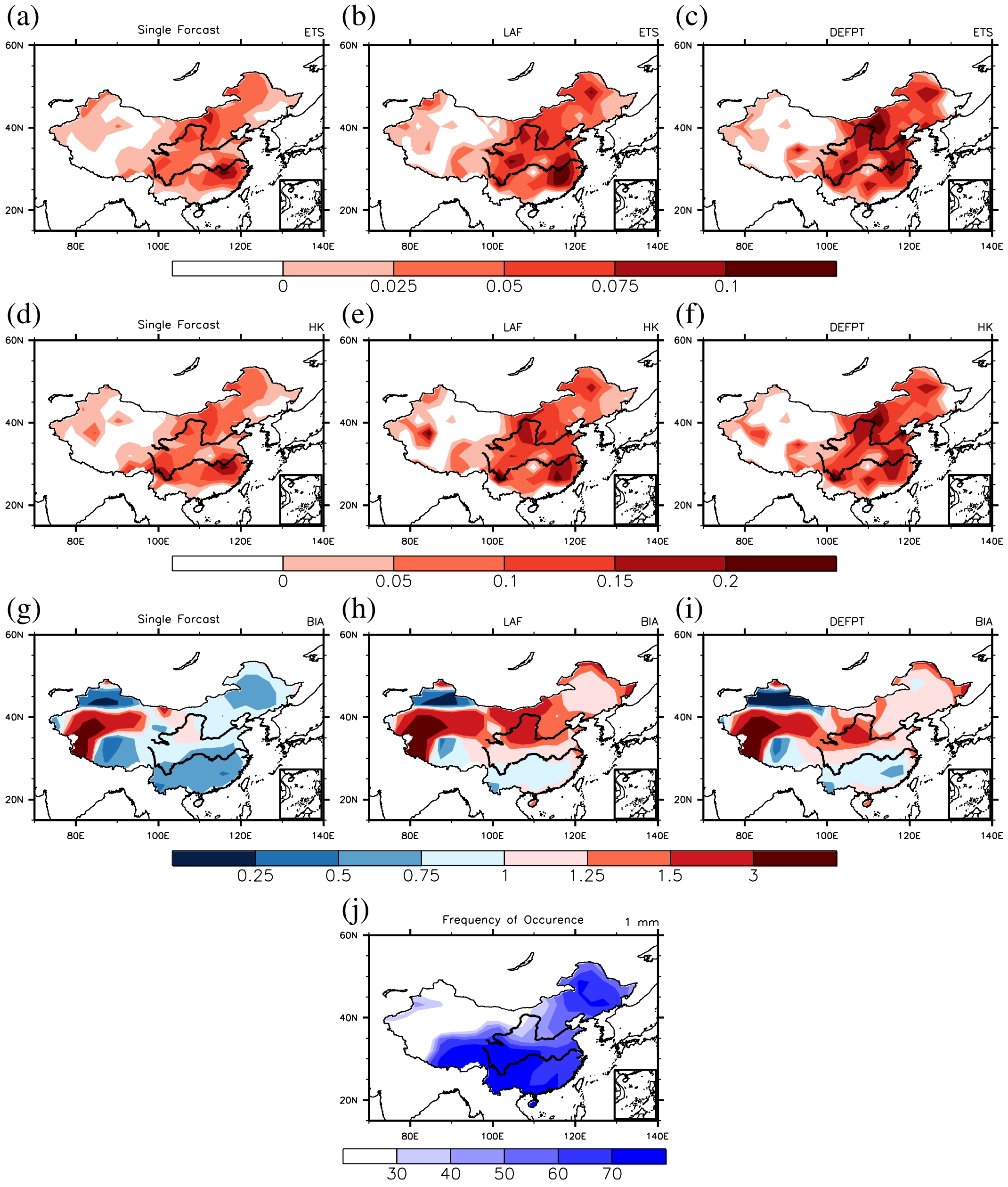

Fig. 10. Geographic distribution of the averaged ETS, HK and BIA scores for 1+ mm per day precipitation forecasts at lead times of 6–15 days from June to August 1998 for (a, d, g) the single forecast, (b, e, h) the LAF using five ensemble members at 6-hour time-lagged intervals, (c, f, i) the DEFPT using a 5/13 probabilistic threshold at the same intervals, and (j) the frequencies of occurrence for 1+ mm per day categorical rainfall. In units of percentages.

thresholds do not visibly depart from the BIA = 1.0. Overall, the probabilistic thresholds with 5/13 for 1 + mm and 4/13 for 2 + mm to 10 + mm per day are relatively good choices for the DEFPT. The evaluations for the predictions on the 11th (Fig. 8d–f) and 14th day (Fig. 8g–i) are similar to those on the 8th day.

In addition, as lead time increases, the number of samples at some probabilistic thresholds may be few and unrepresentative, although their evaluation scores may be high. To some extent, the forecasted rainfall frequency of the DEFPT in summer for different categories can reflect the number of samples within the 1/13–13/13 probabilistic thresholds. Fig. 9 shows the averaged frequency of daily precipitation amounts as a function of threshold from 0.2 mm to 10 mm per day at lead times of 6–15 days over China. The observed rainfall frequency (solid line) decreases from 70% to 10% as the threshold increases and the single forecast frequency is underestimated. Using the optimal probabilistic thresholds (e.g., 5/13 for 1 + mm, 4/13 for 5 + mm–10 + mm), the DEFPT

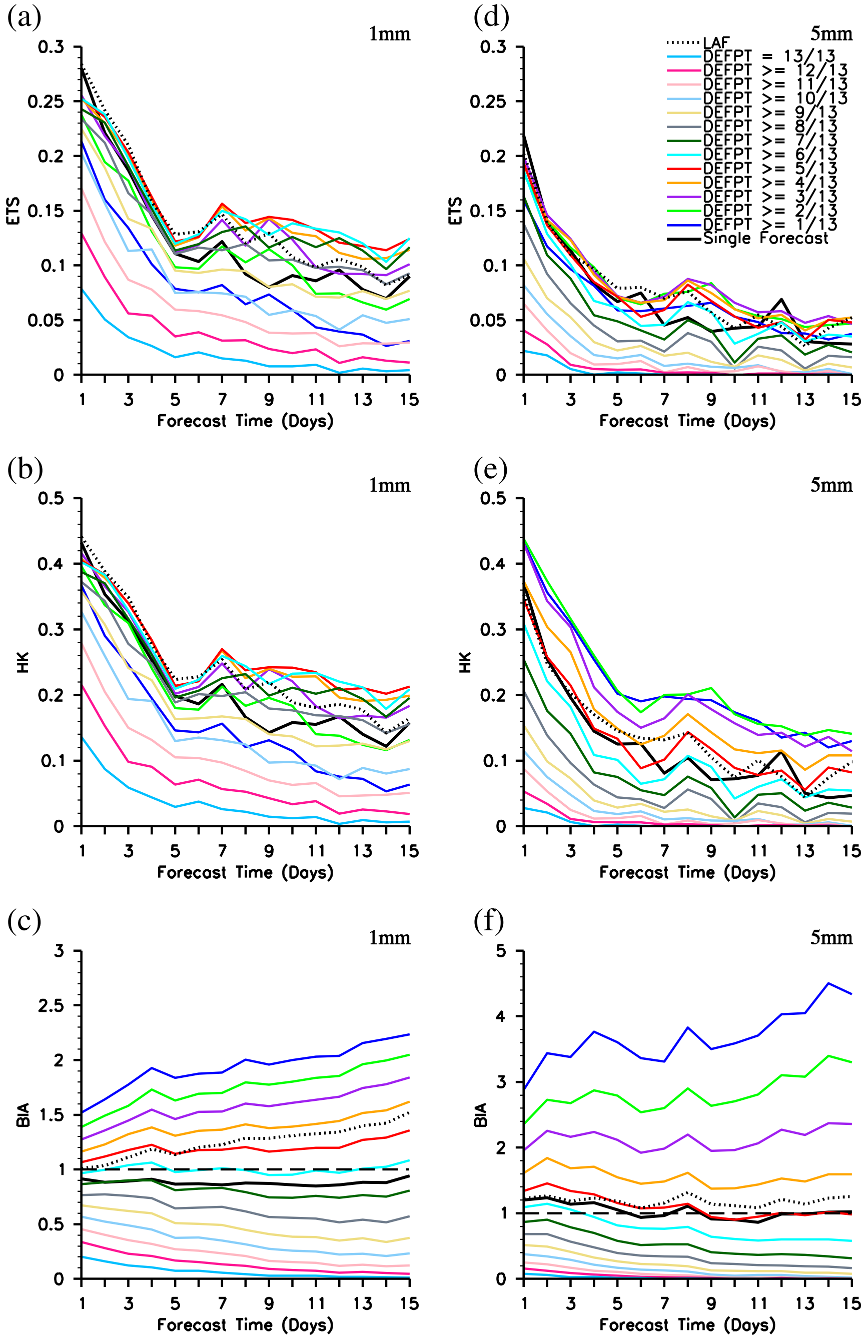

Fig. 11. Same as Fig. 7, but for the ETS (a, d), HK (b, e), and BIA (c, f) for (left panel) the 1 + mm and (right panel) the 5 + mm per day categorical precipitation forecasts from BCC_AGCM2.1 (T42 resolution) for 30 cases during 1996 to 2005.

and the LAF are generally closer to the observed frequency than the single forecast. However, the occurrence frequencies of the DEFPT with higher probabilistic thresholds are always lower than 10% and even close to 0 when the category is larger than 5 mm. It indicates that there may be sampling variability involved in these high probabilistic threshold samples.

We further assessed the geographical performance in ETS, HK and BIA scores for 1+ mm per day precipitation forecast using the DEFPT method with a 5/13 optimal threshold. The hit, miss, false alarm, and correct no-rain forecasts which comprise these evaluation scores for any 6–15 day lead time (i^{th} day) are calculated for each model grid box over China, which is accomplished by comparing the 92 hindcasts at the *i*th day lead time with observations from the summer of 1998. In comparison to the single forecast, Fig. 10a–c shows that the areas that have improved the ETS scores from the DEFPT and LAF for the 1+ mm per day categorical precipitation are located in most parts of southern and northeastern China as well as the eastern part of the Tibetan Plateau. The DEFPT method is substantially better than the LAF in the semi-arid regions over northern China as the corresponding ETS score increases by 0.025–0.05. The above results are also supported by the spatial distribution of HK scores (Fig. 10d–f). For the frequency biases (Fig. 10g–i), the single forecast under-forecasts the 1+ mm per day rainfall in the most rainy regions of China (|1-BIA|<0.5). The LAF can reduce these biases in the most regions where the frequencies of observed rainfall days are larger than 40%–50% except central to northern China, as the values of |1-BIA| are in the range of 0–0.25 (Fig. 10h and j). However, the DEFPT method can further partly reduce the LAF overestimation in central to northern China, as we expected. Overall, the regions of significant improvement from the DEFPT are mainly located in the areas where the frequencies of observed 1+ mm per day rainfall days are above 40%–50% when compared to both the single forecast and the LAF.

4.5 . Verification of DEFPT method for 10 summers

To evaluate the performance of the DEFPT for different precipitation processes that are likely to occur in China, we applied the DEFPT with different probabilistic thresholds to the 1+ mm (5+ mm) per day precipitation in the summers of 1996–2005. Fig. 11a–c shows the averaged ETS, HK and BIA scores for 30 cases from the first day of every month from June to August in 10 years. Similar to the results of the 1998 summer, the ETS and HK scores for the 6–15 day 1+ mm per day rainfall are greatly increased by the DEFPT (solid color lines) using 4/13–6/13 probabilistic thresholds as compared to the single forecast (solid black line) and the LAF (dashed line). This indicates that the prediction skills of rainfall and no rainfall events can be enhanced by the DEFPT. The BIA values (indicating frequency biases) for the DEFPT with the 5/13–6/13 probabilistic thresholds are also closer to 1.0 than that for the LAF. The above result is similar to what is shown in the wet summer of 1998 (Fig. 7a–b). In addition, as compared to the LAF, we also note that the improvement from the DEFPT for the 1996–2005 summers seems to be more significant than that for the extreme year of the 1998 summer. It is mainly caused by the LAF's relatively poor skill for the normal year (Jie et al., 2013). For the 5+ mm per day rainfall prediction, the significant improvements are still obtained when 4/13 threshold is applied to the DEFPT (Fig. 11d–f).

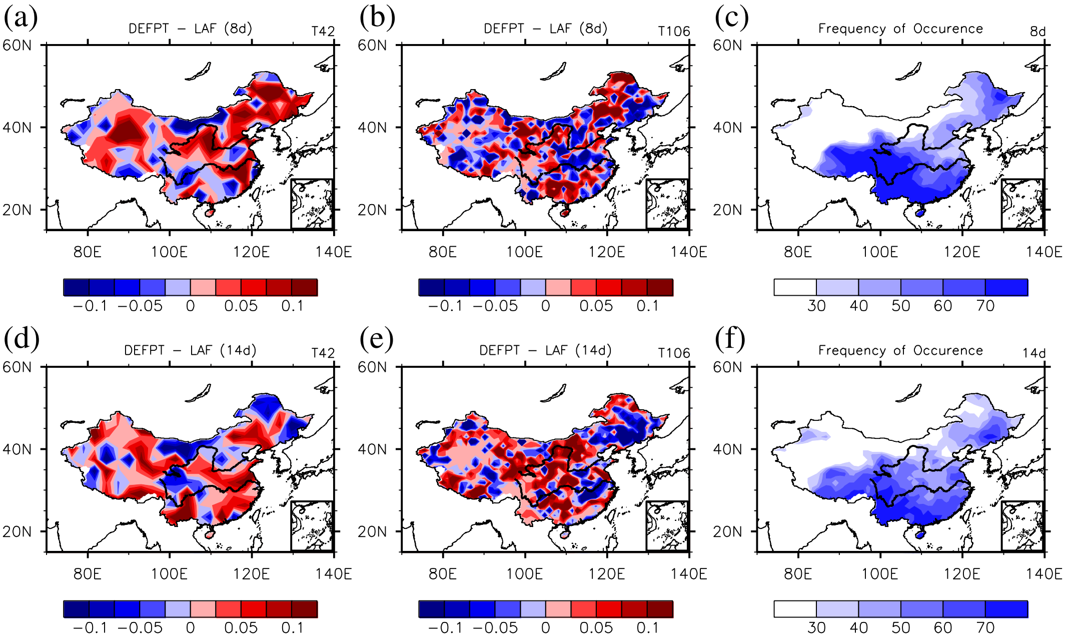

Fig. 12. Geographic distribution of the differences of ETS scores at lead times of (a–b) 8 and (d–e) 14 day for 1+ mm per day categorical precipitation between the DEFPT (5/13 probabilistic threshold) and the LAF. 30 cases in 10 summers (1996–2005) from BCC_AGCM2.1 (T42 resolution) and BCC_AGCM2.2 (T106 resolution) were used; and (c, f) the corresponding frequencies of occurrence for 1+ mm per day categorical rainfall. In units of percentages.

Fig. 12a shows the geographical distribution of differences in ETS scores for 1 + mm per day categorical rainfall between the DEFPT with a 5/13 threshold and the LAF for the 30 cases during 10 years. It shows that improvements using the DEFPT method on the 8th day predicted rainfall are located in the rainiest places and part of the arid and semi-arid regions over China (Fig. 12c). In these locations, the corresponding ETS scores are generally about 0.05–0.1 higher than the counterparts of the LAF. On the 14th day, the DEFPT also improves generally the forecast of 1 + mm per day categorical rainfall as compared to the LAF, although there are some negative skill regions in part of northern and western China. The corresponding ETS scores increase by about 0.025–0.075 (Fig. 12d and f). The distribution of false alarm difference (not shown) between the DEFPT and the LAF can further explain why the DEFPT is better than the LAF in these areas. It shows that, geographically, the ETS values are always positive if the corresponding false alarm rates of DEFPT are lower than the LAF. This also supports our claims that the DEFPT can decrease the over-forecast errors from the LAF for the low categorical precipitation.

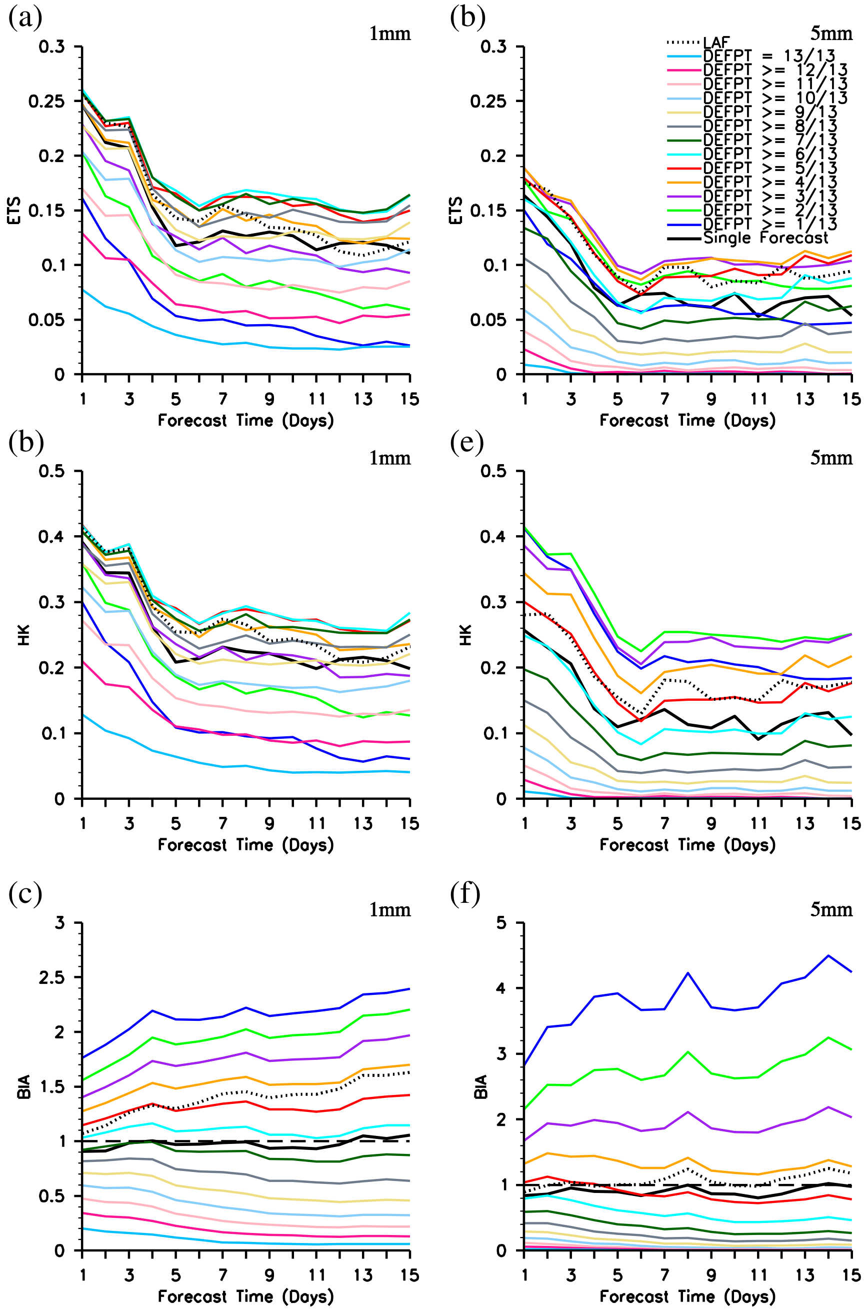

Fig. 13. Same as Fig. 11, but for the forecasts from BCC_AGCM2.2 (T106 resolution).

4.6 . Tests for a higher horizontal resolution

The above evaluations of the DEFPT are all conducted from a large number of hindcasts using the BCC_AGCM2.1 with T42 horizontal resolution. To further evaluate the DEFPT, we applied this method to the higher resolution version (T106) of BCC_AGCM2.2 for 10 summers to study the influence of horizontal model resolutions on the DEFPT's ensemble results of. As shown in Fig. 13a–c, when compared to the single forecast, the DEFPT at the T106 resolution with 5/13–6/13 thresholds can dramatically enhance the skill of 6–15 day rainfall event and no event forecasts, as the corresponding ETS and HK scores separately reach above 0.15 and 0.25 and are obviously higher than the scores from the single forecast (Fig. 13a–b). The frequency biases of the DEFPT with 5/13–6/13 thresholds (Fig. 13c) are much closer to 1.0 than with other thresholds. For a heavier rainfall prediction in the 5 + mm per day category (Fig. 13d–f), the results of the DEFPT at the T106 resolution are basically consistent with those at the T42 resolution, and the optimal probabilistic threshold for DEFPT is still 4/13.

We also show the spatial distribution of ETS score differences for the T106 resolution in Fig. 12. Similar to model results at the T42 resolution, the geographical distribution of ETS score differences for the DEFPT's 1 + mm per day rainfall also shows significant improvements (in contrast to the LAF) for both the 8th day and 14th day forecasts in most parts of China, except for the middle and lower reaches of Yangtze River or parts of northern China (Fig. 12b and e). Compared with the LAF, the DEFPT shows no improvement regions where the DEFPT has higher false alarms rate (e.g., the northeastern China; not shown).

Overall, the optimal thresholds for DEFPT to predict the 1 + mm and 5 + mm per day rainfall do not appear sensitive to the model horizontal resolutions (T42 and T106).

4.7 . The RMSE of DEFPT

Finally, we used the RMSE to measure the forecast error from the DEFPT method with optimal probabilistic thresholds. Fig. 13 shows the RMSE of the 1 mm–10 mm per day categorical precipitation predicted from the single forecast, the DEFPT, and the LAF as a function of forecast length for up to 15 days. This test is based on the 30 cases in 10 summers (1996–2005) at the T106 resolution. In the DEFPT, this forecast is a combination of the 1 + mm per day precipitation predicted using 5/13 probabilistic threshold and the 2 + mm–10 + mm per day using 4/13 probabilistic threshold. During 6–15 day forecasts, the forecast error of the DEFPT is generally smaller than the single forecast (LAF) error as the RMSE value of single forecast decreases by about 0.25 (0.15) mm per day in Fig. 14.

4.8 . The influence of size of ensemble on optimal probabilistic thresholds

We explored the impact of ensemble size on the DEFPT's skill. Fig. 15 shows the ETS, HK and BIA scores of the DEFPT using the 25 members with a 6 hour interval for 6–15 day forecasts during the 1998 summer as a function of rainfall categories (from 1 + mm to 10 + mm per day). Here, 25 members are obtained by extending the lagging time from 3 to 6 days because the ETS scores of all precipitation categories for each ensemble member stabilized after about 6 days (Fig. 3). As shown in Fig. 15, the optimal thresholds for the DEFPT method to get the best forecast results for the 8, 11, and 14 day forecasts also decrease as precipitation category increases, such as 9/25 for 1 + mm, 8/25 for 2 + mm, 7/25 for 5 + mm, and 6/25 for 10 + mm. Overall, for the 1 + mm to 5 + mm per day rainfall, optimal probabilistic thresholds show little change as the maximum lagging time is changed from 3 to 6 days (i.e.,4/13–5/13 or 31%–38% vs. 7/25–9/25 or 28%–36%), and the DEFPT with these optimal probabilistic thresholds is generally better than the LAF.

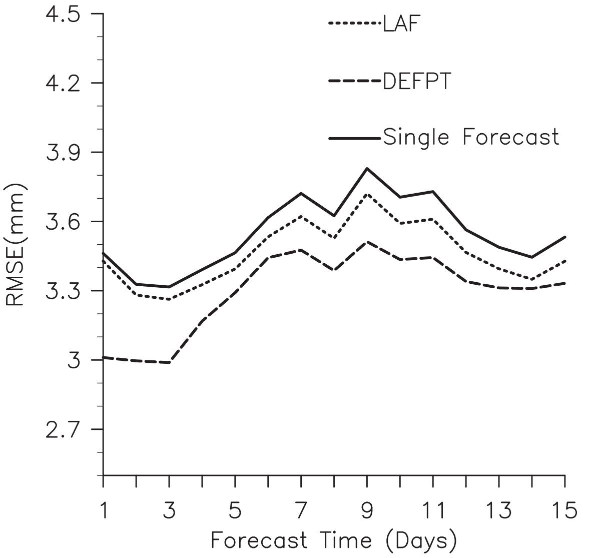

Fig. 14. The RMSE of the 1 mm–10 mm categorical precipitation (i.e., 1 + mm, 2 + mm, ……, 10 + mm) forecasts from the single forecast, the DEFPT and the LAF as a function of forecast length for up to 15 days for 30 cases in 10 summers (1996–2005) from BCC_AGCM2.2 (T106 resolution).

5 . Summary and discussion

This paper studies ensemble forecast methods for 6–15 day daily summer precipitation over China using the BCC_AGCM model. On the basis of the observations of rainfall occurrence, the discussion for categories of 1 + mm to 10 + mm per day rainfall forecasts in this work is meaningful. Although the traditional ensemble mean and probability ensemble forecast methods have their limitations for the 6–15 day precipitation prediction, the DEFPT method based upon 13 6-hour time-lagged ensemble members by using optimal probabilistic thresholds shows significant improvement for predicting precipitation and provides a deterministic (yes/no) forecast from ensemble probability forecasts.

Our analysis via the evaluations of the ETS, HK, BIA scores and ROC curves for a large number of hindcast experiments of 1996–2005 summers shows that the DEFPT method, when compared to a single forecast and the LAF method, can enhance rainfall forecast skill for the 1 + mm and 5 + mm per day categories if the probabilistic threshold for 13 ensemble members is set in the range of 5/13 and 4/13, respectively. These evaluations also support the improvements by the DEFPT

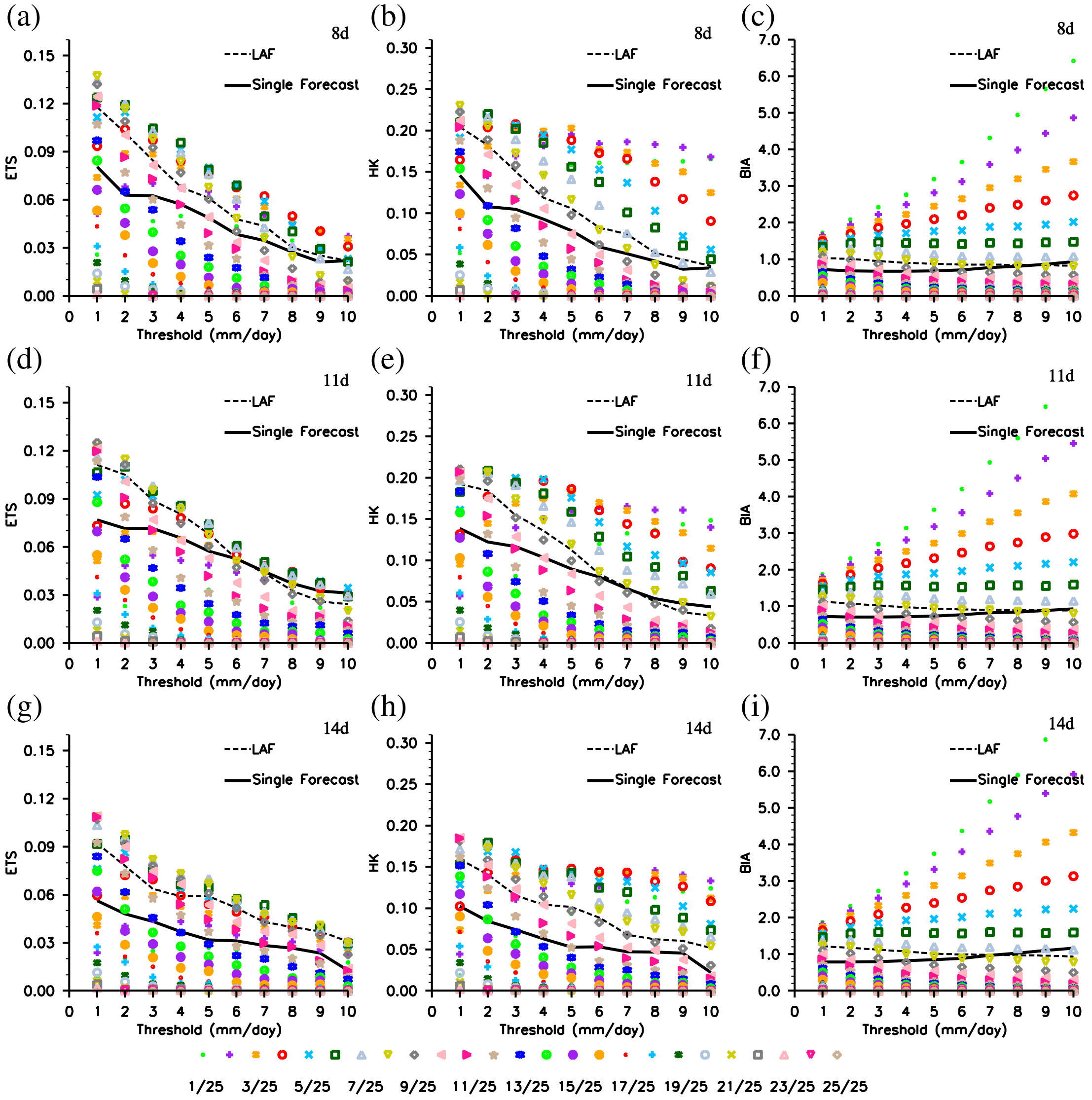

Fig. 15. Same as Fig. 8, but for the DEFPT using 25 time-lagged ensemble members.

for other categories of precipitation between 1+ mm and 10+ mm per day. The forecasts with optimal probabilistic thresholds not only dramatically enhance the single forecast skill of rainy areas over China where the frequencies of rainfall days are higher, but also improve the LAF ensemble in the semi-arid drought regions over China. The RMSE further demonstrates that the forecast errors can be smaller than the single forecast and the LAF. The influence of larger ensemble sizes on the selection of optimal DEFPT thresholds appears to be small.

The comparison of the DEFPT for precipitation hindcasts of the 1996–2005 summers using two different horizontal resolutions (T42 and T106) with BCC_AGCM further indicates the effectiveness of the DEFPT for 6–15 day categorical precipitation forecasts. The optimal probabilistic thresholds of the DEFPT are not sensitive to the model horizontal resolution.

Presently, the physical interpretation of the effectiveness of the DEFPT ensemble method is not completely clear. Despite this, it can reduce the initial uncertainty in long-to-medium range forecasts, because (i) the DEFPT method constitutes a simple form of post-processing in that the flexible choice of the threshold in Eq. (1) permits different adjustments for different precipitation categories; (ii) the DEFPT uses the proper forecasted probabilities to produce a precipitation forecast whose occurrence frequency can be closer to the observation; (iii) this ensemble method does not depend on the intensity of excessive precipitation predicted from only few members, so it is more reliable than the ensemble mean method; (iv) the precipitation events cannot be fully captured by a single forecast, but they might still be predicted by some DEFPT ensemble members.

This study has shown that the selection of the probabilistic threshold for different categories of precipitation is empirical in nature and is expected to depend on uncertainties, error characteristics of current climate models, and ensemble processes. The results of this study are limited to the ensemble techniques for medium range forecasts of the summer rainfall in eastern Asia. Additional analyses are needed to verify the usefulness of the DEFPT method in other parts of the world and other seasons in future. Moreover, the optimal probabilistic thresholds of the DEFPT method selected for the different regions over China could possibly be different, so it is worth studying further. In addition, if under the constraint of practical limitation, models can only run once or twice per day (instead of four times per day), so the DEFPT using the selection of the time-lagged intervals (24 h or 12 h) also needs to be verified. We can possibly develop this method for other ensemble systems in future work.

Finally, better surface observation data are needed to characterize the effects of observation uncertainty (Yuan et al., 2005) in our evaluation of the DEFPT. For example, the improvement from the DEFPT over western China (longitude $\leq$ 90°E), where rain gauges are relatively very rare, is not more remarkable than in other regions of China. The corresponding ETS and HK values are lower and frequency biases are larger (not shown). It possibly relates to the low density of rain gauges. Thus, the results in these regions should be verified using other high-resolution observation data.

Acknowledgments

This work was supported by Special Program for China Meteorology Trade (Grant No. GYHY201306020) and Special Fund for Public Welfare Industry (Meteorology) (Grant No. GYHY201406019). We would like to thank Dr. Alberto CASADO-LOPEZ for his suggestions for this paper.

Appendix A

The BIA, ETS, and HK scores are based on a categorical dichotomous statement (e.g., a yes–no statement). It is then possible, with a given set of matched rain forecasts and observations, to build a 2 × 2 contingency table (Table 2). Each event in this table is identified when a forecast or the observed precipitation is below or above a precipitation category. For a certain precipitation category, the combination of four possibilities of hits, false alarms, misses, and correct no-rain forecasts (*a*, *b*, *c* and *d* as shown in Table 2) between observations and forecasts define the contingency table.

Table 2
Contingency table of possible events for a selected threshold.

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

(1) The BIA score denotes underestimation (overestimation) of rainfall frequency with the value lower (higher) than 1.0, and it is defined as

$$B_{IA} = \frac{a+b}{a+c}. \tag{3}$$

(2) The ETS score is used to verify the skill of predicted rainfall events minus the random forecast skill. An ETS equal to 1 indicates a perfect forecast, while an ETS close to 0 or negative indicates poor rain forecasting skill. The ETS is calculated as

$$E_{TS} = \frac{a-a_r}{a+b+c-a_r}, \tag{4}$$

where a_r is a factor of the model hits expected from a random forecast:

$$a_r = \frac{(a+b)(a+c)}{a+b+c+d}. \tag{5}$$

(3) The HK score is a measure of the accuracy both for events and nonevents. A perfect forecast has an HK score equal to 1.0. This score is computed as:

$$H_k = \frac{ad-bc}{(a+c)(b+d)}. \tag{6}$$

(4) The rank histograms (RHs) (Hamill and Colucci, 1998; Hamill, 2001) are generated by computing the rank of observed precipitation relative to values from an ensemble sorted from lowest to highest for each grid box. A U-shaped rank indicates lack of variability in the ensemble, but a uniform rank shows the ensemble is dispersed as the observation ranks equally among the ensemble members. In addition, wet biases (over forecast) in ensemble forecasts can lead to an L-shaped RH, while dry biases (under forecast) often cause observations to rank highest with a reversed L shape (Yuan et al., 2009).

(5) The attribute diagram is usually used to reveal the properties of PQPF which can be described by three terms (reliability, resolution, and uncertainty of the forecasts) from the Brier score (BS, Murphy, 1973; Murphy and Winkler, 1987). The lower the BS, the better the forecast. Its formulation is:

$$B_S = \underbrace{\frac{1}{N}\sum_{k=1}^{K} N_k\left(f_k - \overline{o_k}\right)^2}_{1} - \underbrace{\frac{1}{N}\sum_{k=1}^{K} N_k\left(\overline{o_k} - \bar{o}\right)^2}_{2} + \underbrace{\bar{o}\left(1-\bar{o}\right)}_{3}. \tag{7}$$

where K is the number of forecasted probability bins (are mutually exclusive) for a certain category of precipitation, the k^{th} bin sub-sample contains N_k events (i.e., $\sum_{k=1}^{K} N_k = N$), f_k $(0 \leq f_k \leq 1)$ denotes the forecasted probability in the k^{th} bin, $\overline{o_k}$ is the relative frequency of observed rainfall occurrence in the k^{th} bin, $\bar{o}$ is the mean of all the frequencies (i.e., $\bar{o} = \frac{1}{N}\sum_{k=1}^{K} N_k\overline{o_k}$). The first term is the averaged squared difference between the f_k and the $\overline{o_k}$. It indicates that the PQPF is reliable, when it is close to 0 (i.e., $f_k \approx \overline{o_k}$). The second term is the mean of squared difference between the $\overline{o_k}$ and $\bar{o}$. It is a measure of the resolution degree of the observed relative frequency in each probability bin and the mean of all these frequencies. The larger second term indicates better resolution. Overall, if the second term is larger than the first one (i.e., $(\overline{o_k}-\overline{o})^2 \geq (f_k-\overline{o_k})^2$), the Brier skill score will be positive. It suggests that the PQPF is skillful as compared to the forecast using climatic frequency of observation (see details in Hsu and Murphy, 1986). Thus, the $\overline{o_k}=(f_k+\overline{o})/2$ can produce a no forecast skill line in attribute diagram and when this line is below (above) the no resolution line, the positive skill is provided when the reliability curve is also below (above) the no skill line. In addition, the work of Hamill and Juras (2006) indicated that the variations of climatological event frequency may partly affect the evaluation of attribute diagram.

(6) The Relative Operating Characteristic (ROC) curve is used to verify the discriminating ability of the ensemble probability forecast (e.g., Mason, 1982; Harvey et al., 1992; Jolliffe and Stephenson, 2003). Its ordinate indicates the hits rate ($a/(a + c)$) and the abscissa is the false alarm rate ($b/(b + d)$), where a, b, c and d are the same as Eq. (3). Therefore, a better discriminating probability forecast can produce a skewed curve in the upper-left corner of the diagram. The area under the ROC curve is the perfect value of 1.0 and no skill value of 0.5. Generally, the discriminatory skill of probability forecast is useful when the ROC area exceeds 0.7 (Buizza et al., 1999; Bright et al., 2004).

(7) The Root-Mean-Square Error (RMSE) is applied to measure the forecast error. The formulation is:

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-Y_i)^{2}} \tag{8}$$

where X_i is forecast value, Y_i is observed value for the i^{th} element, and N denotes total number of elements.

References

Bright, D.R., Weiss, S.J., Wandishin, M.S., Kain, J.S., Stensrud, D.J., 2004. Evaluation of short-range ensemble forecasts during the 2003 SPC/NSSL Spring Program. Preprints, 22nd Conf. on Severe Local Storms, Hyannis, MA, Amer. Meteor. Soc., P15.5.

Buizza, R., Hollingsworth, A., Lalaurette, F., Ghelli, A., 1999. Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. Weather Forecast. 14, 168–189.

Chen, H., Yu, R., Li, J., Xin, X., Wang, Z., Wu, T., 2011. The coherent interdecadal changes of East Asia climate in mid-summer simulated by BCC_AGCM2.0.1. Clim. Dyn. http://dx.doi.org/10.1007/s00382-011-1154-6.

Cressman, G.P., 1959. An operational objective analysis system. Mon. Weather Rev. 87, 367–374.

Dalcher, A., Kalnay, E., Hoffman, R.N., 1988. Medium range lagged forecasts. Mon. Weather Rev. 116, 402–416.

Ding, Y., Hu, G., 2003. A study on water vapor budget over China during the 1998 severe flood periods. Acta Metall. Sin. (in Chin.) 61 (2), 129–145.

Dong, M., Wu, T., Wang, Z., et al., 2009. Simulations of the tropical intra-seasonal oscillations by the AGCM of the Beijing Climate Center. Acta Meteorol. Sin. 67 (6), 912–922.

Du, J., Mullen, S.L., Sanders, F., 1997. Short-range ensemble forecasting of quantitative precipitation. Mon. Weather Rev. 125, 2427–2459.

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

Eckel, F.A., Walters, M.K., 1998. Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. Weather Forecast. 13, 1132–1147.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26.

Efron, B., 1981. Censored data and the bootstrap. J. Am. Stat. Assoc. 76, 312–319.

Fan, K., Wang, H.J., Choi, Y.J., 2008. A physically-based statistical forecast model for the middle–lower reaches of the Yangtze River Valley summer rainfall. Chin. Sci. Bull. 53, 602–609.

Hamill, T.M., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Mon. Weather Rev. 129, 550–560.

Hamill, T.M., Colucci, S.J., 1998. Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. Mon. Weather Rev. 126, 711–724.

Hamill, T.M., Juras, J., 2006. Measuring forecast skill: is it real skill or is it the varying climatology? Q. J. R. Meteorol. Soc. 132, 2905–2923.

Hamill, T.M., Whitaker, J.S., Wei, X., 2004. Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. Mon. Weather Rev. 132, 1434–1447.

Hamill, T.M., Hagedorn, R., Whitaker, J.S., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. Mon. Weather Rev. 136, 2620–2632.

Hanssen, A.W., Kuipers, W.J.A., 1965. On the relationship between the frequency of rain and various meteorological parameters. Meded. Verh. 81, 2–15.

Harvey, L.O., Hammond, K.R.C., Lusk, M., Mross, E.F., 1992. The application of signal detection theory to weather forecasting behavior. Mon. Weather Rev. 120, 863–883.

Hohenegger, C., Schär, C., 2007. Atmospheric predictability at synoptic versus cloud-resolving scales. Bull. Am. Meteorol. Soc. 88, 1783–1793.

Hsu, Wu-ron, Murphy, A.H., 1986. The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. Int. J. Forecast. 2 (3), 285–293.

Jie, W., Wu, T., 2010. Hindcast for the 1998 summer heavy precipitation in the Yangtze and Huaihe River Valley using BCC_AGCM2.0.1 model. Chin. J. Atmos. Sci. 34 (5), 962–978.

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. 13, 1–12. http://dx.doi.org/10.1007/s00376-013-3037-8.

Jolliffe, I.T., Stephenson, D.B., 2003. Forecast Verification: a Practitioner's Guide in Atmospheric Science. Wiley, New York, p. 66.

Katz, R.W., Murphy, A.H., 1997. Economic Value of Weather and Climate Forecasts. Cambridge University Press, Cambridge, New York.

Krishnamurti, T.N., et al., 2000. Multimodel ensemble forecasts for weather and seasonal climate. J. Clim. 13, 4196–4216.

Lee, K.K., Lee, J.W., 2007. The economic value of weather forecasts for decision making problems in the profit/loss situation. Meteorol. Appl. 14, 455–463.

Liu, Y., Fan, K., 2014. An application of hybrid downscaling model to forecast summer precipitation at stations in China. Atmos. Res. 143, 17–30.

Lu, C., Yuan, H., Schwartz, B.E., Benjamin, S.G., 2007. Short-range numerical weather prediction using time-lagged ensembles. Weather Forecast. 22, 580–595.

Martin, M.L., Santos-Muñoz, D., Valero, F., Morata, A., 2010. Evaluation of an ensemble precipitation prediction system over the Western Mediterranean area. Atmos. Res. 98, 163–175.

Mason, I., 1982. A model for assessment of weather forecasts. Aust. Meteorol. Mag. 30, 291–303.

McLay, G.J., 2008. Markov chain modeling of sequences of lagged NWP ensemble probability forecasts: an exploration of model properties and decision support applications. Mon. Weather Rev. 136, 3655–3670.

Mullen, S.L., Buizza, R., 2001. Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. Mon. Weather Rev. 129, 638–663.

Murphy, A.H., 1973. A new vector partition of the probability score. J. Appl. Meteorol. 12, 595–600.

Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Mon. Weather Rev. 115, 1330–1338.

Romatschke, U., Houze, R.A., 2011. Characteristics of precipitating convective systems in the South Asian monsoon. J. Hydrometeorol. 12, 3–26.

Schaefer, J.T., 1990. The critical success index as an indicator of warning skill. Weather Forecast. 5, 570–575.

Sivillo, J.K., Ahlquist, J.E., Toth, Z., 1997. An ensemble forecasting primer. Weather Forecast. 12, 809–818.

Tippett, M.K., Barnston, A.G., Robertson, A.W., 2007. Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles. J. Clim. 20, 2210–2228.

Vich, M., Romero, R., Homar, V., 2011. Ensemble prediction of Mediterranean high-impact events using potential vorticity perturbations. Part II: Adjoint-derived sensitivity zones. Atmos. Res. 102, 311–319.

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

Walser, A., Luthi, D., Schar, C., 2004. Predictability of precipitation in a cloud-resolving model. Mon. Weather Rev. 132, 560–577.

Wang, L., Zhou, T., Wu, T., et al., 2009. Simulation of the leading mode of Asian–Australian monsoon interannual cariability with Beijing Climate Center atmospheric general circulation model. Acta Meteorol. Sin. 67 (6), 973–982.

Whitaker, J.S., Wei, X., Vitart, F., 2006. Improving week-2 forecasts with multimodel reforecast ensembles. Mon. Weather Rev. 134, 2279–2284.

Wilks, D.S., 1995. Statistical Methods in the Atmospheric Sciences. Academic Press (467 pp.).

Wu, T., 2012. A mass-flux cumulus parameterization scheme for large-scale models: description and test with observations. Clim. Dyn. 38, 725–744. http://dx.doi.org/10.1007/s00382-011-0995-3.

Wu, T., Yu, R., Zhang, F., et al., 2008. A modified dynamic framework for atmospheric spectral model and its application. J. Atmos. Sci. 65, 2235–2253.

Wu, T., Yu, R., Zhang, F., et al., 2010. The Beijing Climate Center for Atmospheric General Circulation Model (BCC-AGCM2.0.1): description and its performance for the present-day climate. Clim. Dyn. 34, 123–147.

Wu, T., Li, W., Ji, J., et al., 2013. Global carbon budgets simulated by the Beijing Climate Center Climate System Model for the last century. J. Geophys. Res. 10, 4326–4347.

Yuan, H., Mullen, S.L., Gao, X., Sorooshian, S., Du, J., Juang, H.H., 2005. Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. Mon. Weather Rev. 133, 279–294.

Yuan, H., Lu, C., McGinley, J.A., Schultz, P.J., Jamison, B.D., Wharton, L., Anderson, C.J., 2009. Evaluation of short-range quantitative precipitation forecasts from a time-lagged multimodel ensemble. Weather Forecast. 24, 18–38.