Stochastic Environmental Research and Risk Assessment (2021) 35:515–529
https://doi.org/10.1007/s00477-020-01923-9

ORIGINAL PAPER

The adaptability of typical precipitation ensemble prediction systems in the Huaihe River basin, China

Han Wang[1] · Ping-an Zhong[1,2] · Fei-lin Zhu[1] · Qing-wen Lu[1] · Yu-fei Ma[1] · Sun-yu Xu[1]

Accepted: 23 October 2020 / Published online: 5 November 2020

Abstract

Evaluating the adaptability of precipitation forecasting is of great importance for regional flood control and drought warnings. This study conducted evaluations using the 1–9 days cumulative precipitation forecast data of five typical operational global ensemble prediction systems (EPSs) from TIGGE (i.e., The Observing System Research and Predictability Experiment Interactive Grand Global Ensemble) and the observed daily precipitation data of 40 meteorological stations over the Huaihe River basin (HB). A series of verification metrics is used to evaluate the performances of quantitative precipitation forecasts (QPFs) and probabilistic quantitative precipitation forecasts (PQPFs) from the five EPSs from April to December 2015 in terms of overall performance, different precipitation thresholds, different lead times and the spatial distribution over the HB. The adaptability of the multimodel superensemble integrated from the five EPSs by the Bayesian model average is also examined during the main flood season. The results show that (1) the forecast quality of the China Meteorological Administration EPS is the worst for all lead times, which may relate to its having the fewest ensemble members. The European Centre for Medium-Range Weather Forecasts (ECMWF) EPS performs the best in terms of QPF and PQPF qualities for longer lead times because ECMWF has the largest ensemble members. (2) All EPSs have better discrimination at low thresholds, indicating the reference value for drought warnings. ECMWF is expected to obtain the best PQPF skill for a large threshold through postprocessing; (3) due to the differences in climates in the North and South of the basin, QPF and PQPF qualities are better in the northern HB than in the southern HB; (4) except for climate, the PQPF skill is also influenced by precipitation type, while the QPF accuracy is affected by terrain. The PQPF is good at forecasting the precipitation caused by ocean effects but not by mountain topography. The QPF accuracy decreases in mountainous areas; and (5) the multimodel superensemble has little effect on PQPF skill improvement but can improve QPF accuracy when raw EPSs have significantly different QPF accuracies.

Keywords Huaihe river basin · Ensemble prediction system · Precipitation forecast · Bayesian model average

✉ Ping-an Zhong
zpa_hhu@163.com; pazhong@hhu.edu.cn

Han Wang
wanghan_hhu@163.com

Fei-lin Zhu
zhufeilin@hhu.edu.cn

Qing-wen Lu
luqw@hhu.edu.cn

Yu-fei Ma
mayufei@hhu.edu.cn

Sun-yu Xu
3022529385@qq.com

[1] College of Hydrology and Water Resources, Hohai University, No.1 Xikang Road, Nanjing 210098, China

[2] National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Hohai University, No.1 Xikang Road, Nanjing 210098, China

1 Introduction

Floods and droughts are major natural disasters that occur in China and have enormous destructive strength, restricting economic and social development (Wang 2017; Zhang Xu et al. 2019). As an essential part of the hydrological forecasting system, skillful precipitation forecasts are of vital importance in mitigating risk associated with extreme events, which supports decision-making for water resource utilization. In addition, precipitation forecasts provide decision-makers with uncertainty information in precipitation and flood forecasts (Ying et al. 2019).

Numerical weather prediction (NWP) models have been developed and improved since the 1940s (Trenberth 1992), and the forecast accuracy has also steadily improved (Buizza et al. 1999; Lan et al. 2011). In the past 25 years, NWP has evolved from deterministic forecasting to a new stage of ensemble prediction systems (EPSs) (Molteni et al. 1996). In contrast to deterministic forecasting, EPS generates a forecast set through initial perturbation and model uncertainty, thereby providing the most likely forecast value as well as the uncertainty of the forecast. The improved EPS performance is attributed to advances in the initial perturbation strategy, (Meng 2011; Whitaker and Hamill 2002), model uncertainty simulation strategy, resolution, number of members and forecast length (Roberto 2019; Roebber et al. 2004). Currently, EPS is not only used operationally to generate forecasts valid for different time scales, such as short-term forecasts (up to 2–3 days), medium-term forecasts (up to 2 weeks), seasonal forecasts and subseasonal (10–90 days) forecasts but also for different hydrometeorological variables, such as temperature, precipitation, wind speed, and tropical cyclone paths (François et al. 2018; Hemri et al. 2014). Currently, EPS is widely used in many fields. For example, in hydrology, coupled with hydrological models, precipitation ensembles can generate runoff forecasts (Cloke and Pappenberger 2009; Lan et al. 2011; Pappenberger et al. 2005). In the energy field, different weather scenarios are created by ensembles to estimate the uncertainty of electricity demand forecasts (Taylor and Buizza 2003). In aviation, ensembles are used to provide the probability of convectional calamity and flying conditions, guiding air traffic control (Robert 2018; Verlinden 2017). Furthermore, the inherent forecast limitations of a single model are difficult to measure. It is common to combine ensembles from multiple independent models in a scheme called a multimodel ensemble. This practice considerably reduces systematic errors in forecasts and improves reliability (Kirtman et al. 2014; Krishnamurti et al. 1999, 2016).

The Observing System Research and Predictability Experiment Interactive Grand Global Ensemble (TIGGE) provides a solid technical and data support for studies on the operational global ensemble forecasts (Park et al. 2008). In recent years, regional cases on ensemble forecast systems have been extensively carried out in quantitative precipitation forecasts (QPFs) and probabilistic quantitative precipitation forecasts (PQPFs): Hamill (2012) examined the PQPFs from four TIGGE EPSs over the contiguous United States during July–October 2010 and discussed the TIGGE multimodel and the European Centre for Medium-Range Weather Forecast (ECMWF) reforecast-calibrated PQPFs. The author concluded that PQPFs from the Canadian Meteorological Centre (CMC) EPS are the most reliable, while those from the U.S. National Centre for Environmental Prediction (NCEP) and the United Kingdom Meteorological Office (UKMO) EPSs are the least reliable. In addition, the TIGGE multimodel shows better forecast skills, while the accuracy of ECMWF reforecast-calibrated PQPFs is reduced. Xiang et al. (2014) evaluated the QPFs and PQPFs from six TIGGE EPSs during June–August 2008–2010 in the Northern Hemisphere (NH) midlatitude and tropics, as well as the change in performance after being upgraded. Their study indicated that the overall forecast skill is better in the NH midlatitude than in the NH tropics, and generally, the ECMWF EPS performs best. After the upgrade, the overall QPF and PQPF errors from CMC EPS increase due to its excessively enlarged ensemble spread. Louvet et al. (2016) compared PQPFs from seven TIGGE EPSs with satellite rainfall estimates over West Africa during 2008–2012 and examined the performance of the ensemble mean of all models. They found that the skills of UKMO and ECMWF EPSs are better than others. For a lead time from 1 to 15 days, the skill of TIGGE forecasts decreases, and the performance of the multimodel overcomes that of any individual models. Karuna et al. (2017) assessed the skills of three TIGGE EPSs in predicting 15 rainstorm events over India during 2007–2015. Their results showed that NCEP EPS has the least spread, but its QPFs are not well predicted. The displacement and pattern errors contribute more to the total root mean square error (RMSE). Using deterministic, dichotomous (yes/no) and probabilistic techniques, Aminyavari et al. (2018) verified the precipitation forecast performance of three TIGGE EPSs over Iran for the period of 2008–2016. This study concluded that all EPSs underestimate precipitation in high precipitation regions and overestimate precipitation in other regions. ECMWF EPS has better scores than others, while UKMO EPS yields higher scores in mountainous regions. The multimodel superensemble is recommended to improve the forecast quality.

However, systematic studies on regional TIGGE precipitation forecasts are scarce. Thus, a more comprehensive study is needed to reveal the detailed properties of regional precipitation EPSs. In addition, statistical postprocessing can construct a multimodel superensemble from EPSs to remove systematic biases and improve the accuracy and robustness of EPSs (Qingyun et al. 2019). It is of interest to analyze the forecast skill of a multimodel superensemble in a particular area.

This study focuses on the QPFs and PQPFs generated from individual TIGGE centers from April to December 2015 over the Huaihe River basin (HB). The forecast quality is assessed in many aspects to obtain a comprehensive understanding and summary of the precipitation forecast properties of five selected operational global EPSs in the HB. The overall forecast quality is verified, and the forecast quality at different precipitation thresholds is further discussed. Forecast quality changes for different lead times are also examined in this study. We evaluate the spatial distribution of forecast performance to reveal the adaptabilities of EPSs to the terrain and climate background. In addition, the multimodel superensemble is integrated from five EPSs using the Bayesian model average (BMA), and its performance is evaluated with reference to individual EPSs.

The rest of the paper is organized as follows: Sect. 2 describes the study area, datasets and methods. Section 3 provides the results and discussions. A summary is presented in Sect. 4.

2 Study area, datasets and methods

2.1 Study area

The HB is located at 111°55′ E–121°25′ E and 30°55′ N–36°36′ N (Fig. 1). The left bank of the Huaihe River is almost all plain rivers with large concentration areas, while the right bank is all hilly rivers with small, concentrated areas. In addition, HB is a transitional zone between the northern and southern climates of China (Robert 2018). In contrast to the warm zone with a semihumid monsoon climate in the northern region, the southern region is a subtropical zone with a humid monsoon climate.

The average annual precipitation over the HB is approximately 910 mm, and the precipitation decreases from South to North. June–September is the flood season in the HB, and precipitation is usually 500–600 mm during this period, accounting for 50–80% of the annual precipitation. During the unique plum rain season (June and July), rainfall lasts for 1 or 2 months, covering the whole basin. The atmospheric system is complex and changeable over the HB. The spatial and temporal distribution of precipitation is uneven and prone to floods, droughts and other disasters. The complex terrain and unique climate background make it difficult to forecast precipitation in this region.

2.2 Datasets

2.2.1 Observed data

The observed data set comes from the National Meteorological Information Centre of China. The data set is the collection of surface meteorological records submitted monthly by the data-processing departments of provinces, municipalities and autonomous regions. The data set comprises the daily data of 752 meteorological stations in China from 1951 to 2015. Daily precipitation data from 40 stations over the HB are used in this study. Some dates are missing data or contain outliers, and these dates are culled.

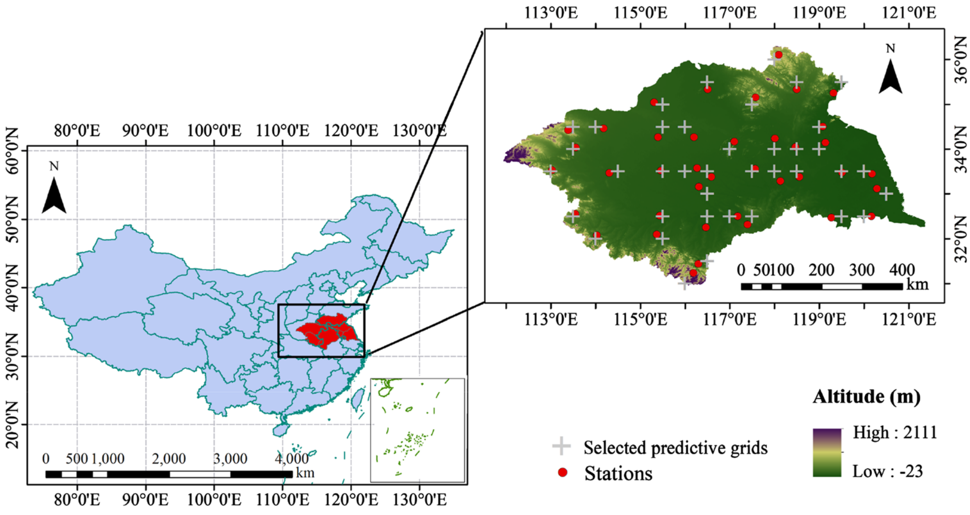

Fig. 1 The location of HB, predicted grid points and stations

2.2.2 Precipitation forecast data

The cumulative precipitation forecast data of 1–9 days provided by the Japan Meteorological Agency (JMA), China Meteorological Administration (CMA), UKMO, U.S. NCEP and the ECMWF EPSs in the TIGGE data set are adopted for evaluation. The regional range is 112°–121° E and 30.5°–36.5° N. The original precipitation data are converted into the same 0.5° × 0.5° grid before downloading using the bilinear interpolation software provided by the ECMWF TIGGE data portal. The configurations of the selected operational global EPSs are shown in Table 1.

The JMA EPS starts to provide data in February 2014; in addition, the CMA EPS is missing data from October 2014 to March 2015 due to the system upgrade, and thus, the verification period covers April–December 2015 in this study. The negative values of the forecast are set as 0, and linear interpolation of time is used to estimate the missing values. The nearest-neighbor approach is used to obtain the forecast of a specific station from the gridded forecast data (Vogel et al. 2017). Figure 1 illustrates the distribution of the selected predicted grid points and the meteorological stations. The background grid in gray is the original grid.

2.3 Verification metrics and postprocessing method

The QPFs and PQPFs of five typical EPSs are verified in terms of different verification metrics. A multimodel superensemble is constructed by the BMA. The principles of each part are described as follows.

2.3.1 Verification methods

2.3.1.1 Verification metrics of PQPFs The direct output of EPS is a set of possible values (i.e., PQPF); thus, PQPF is a probabilistic prediction. Sharpness, skill, reliability and resolution are the most common aspects of probabilistic prediction quality. The sharpness describes the concentration of the probabilistic prediction distributions. The skill represents the forecast accuracy compared with a reference forecast. The reliability relates to the average consistency between the forecast and observation when a specific forecast is issued, measuring how well forecast probabilities match observed frequencies. The resolution shows differences in outcomes for the different forecasts issued, which means that the distribution of outcomes when “A” was forecast is different from the distribution of outcomes when “B” is forecast (Qingyun et al. 2019). In this study, the continuous ranked probability skill score (CRPSS) is applied to assess the forecast skill of PQPFs (Hersbach 2000). The reliability diagram and Brier score resolution are used to intuitively and quantitatively evaluate the reliability of PQPFs, respectively, for dichotomous events. Dichotomous events refer to events whose results can be divided into occurrence and nonoccurrence through thresholds. Brier score skill and Brier score resolution represent the prediction skill and resolution of PQPFs for dichotomous events, respectively.

The CRPSS is calculated by normalizing the continuous ranked probability score (CRPS) with the reference forecast, which is defined as follows:

$$CRPSS_j^T = \frac{CRPS_{ref,j} - CRPS_j^T}{CRPS_{ref,j}} \tag{1}$$

where $CRPSS_j^T$ represents the CRPSS of EPS T at station j and $CRPS_j^T$ is the CRPS of EPS T at station j (Tilmann and Raftery 2007):

$$CRPS_j^T = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\left[G_{ij}^T(x) - H\left(x - o_{ij}\right)\right]^2 dx \tag{2}$$

$$H\left(x - o_{ij}\right) = \begin{cases} 1 & x \geq o_{ij} \\ 0 & x < o_{ij} \end{cases} \tag{3}$$

where x represents accumulated precipitation; o_{ij} is the observation on i day at j station; N is the number of days in the verification period; and G_{ij}^T represents the predictive cumulative distribution function of T EPS on i day at j station.

$CRPS_{ref,j}$ represents the referenced CRPS at j station and is generated using the cumulative distribution function

Table 1 Configurations of the five TIGGE EPSs used in this study

EPS	Country	Ensemble members	Lead time	Horizontal resolution archived
UKMO	UK	23	1–9 days	0.45° × 0.3°
ECMWF	Europe	50		O640 (0.14° × 0.14°)*
CMA	China	14		0.28° × 0.28°
JMA	Japan	26		0.37° × 0.37°
NCEP	USA	20		1.0° × 1.0°

*ECMWF uses Gaussian grid O640, which means there are 640 lines between the pole and equator, giving a total of 1280 latitude lines. The resolution is approximately 0.14° × 0.14° over the HB

(CDF) of the observed samples (i.e., sample climatology) (Konstantinos et al. 2019):

$$CRPS_{ref,j} = \frac{1}{N}\sum_{i=1}^{N}\left|o_{ij} - \overline{o}_j\right| \tag{4}$$

where $\overline{o_j}$ is the average observed precipitation at j station during the verification period. CRPSS ranges from $-\infty$ to 1, and a negative value indicates that the forecast skill of EPS T is worse than that of the sample meteorology (Demargne et al. 2010; Ye et al. 2014). In this study, 95% confidence intervals for CRPSS are calculated by the bootstrapping method by randomly selecting the statistics 10,000 times (Xiang et al. 2014).

The reliability diagram represents the frequency of the actual event when the predicted event occurs with a certain probability. The reliability diagram sets the observed relative frequency of an event versus the forecast probability of the event (Fig. 2). Given that m denotes the different M thresholds of forecast probability, the observed relative frequency q_m^T is given by the following equation (Wilks 2009):

$$q_m^T = \frac{1}{n_{mj}^T \times J}\sum_{j=1}^{J}\sum_{i=1}^{n_{mj}^T}\gamma_{mj}^T \tag{5}$$

where n_{mj}^T denotes the number of forecast-observation pairs used in the verification period for EPS T at j station; and J represents the total number of stations. Since the observation of the event is dichotomous for the forecast-observation pair of T EPS at the j station, $\gamma_{mj}^T = 1$ if the event occurs and $\gamma_{mj}^T = 0$ otherwise. According to the forecast probability, the reliability diagram parts the verification dataset into subsamples, which means that the reliability diagram requires a fairly large dataset.

The Brier score resolution and Brier score reliability for EPS T are defined as follows (Wilks 2009):

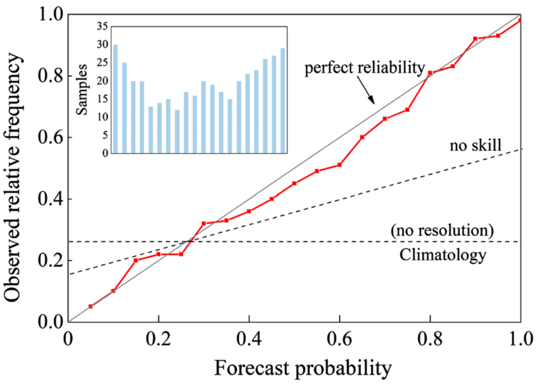

Fig. 2 Schematic of the reliability diagram

$$RES^T = \frac{1}{\sum_{j=1}^{J}\sum_{m=1}^{M} n_{mj}^T}\sum_{i=1}^{n_{mj}^T}\sum_{m=1}^{M}\left(p_m^T - q_m^T\right)^2 \tag{6}$$

$$REL^T = \frac{1}{\sum_{j=1}^{J}\sum_{m=1}^{M} n_{mj}^T}\sum_{i=1}^{n_{mj}^T}\sum_{m=1}^{M}\left(q_m^T - \sum_{j=1}^{J}\sum_{m=1}^{M}\frac{\gamma_{mj}^T}{n_{mj}^T}\right)^2 \tag{7}$$

where p_m is the forecast probability of threshold m; and q_m^T is the observed relative frequency corresponding to the threshold m of EPS T. The larger RES is, the higher the resolution of PQPFs, and the smaller REL is, the better the reliability.

BS skill (BSS) normalizes the mean square error of PQPFs of dichotomous events by reference forecast. For PQPFs of EPS T, the BSS is given by the following equation:

$$BSS^T = \frac{BS_{ref} - BS^T}{BS_{ref}} \tag{8}$$

where BS^T and BS_{ref} represent the BS of EPS T and reference forecast, respectively, and BS_{ref} is calculated by the observed sample frequency of each station (Thomas and Josip 2006; Wilks 2009):

$$BS^T = \frac{1}{N \times J}\sum_{j=1}^{J}\sum_{i=1}^{N}\left(p_{ij}^T - \gamma_{ij}\right)^2 \tag{9}$$

$$BS_{ref} = \frac{1}{N \times J}\sum_{j=1}^{J}\sum_{i=1}^{N}\left(p_{ref,j} - \gamma_{ij}\right)^2 \tag{10}$$

$$\gamma_{ij} = \begin{cases} 1 & o_{ij} \in I \\ 0 & o_{ij} \notin I \end{cases} \tag{11}$$

where p_{ij}^T is the forecast probability of the event of T EPS on i day at j station; $p_{ref,j}$ is the observed frequency at j station. Similar to the CRPSS, the perfect score is 1, and the lower the BSS is, the worse the skill of PQPFs. A lower limit of atmospheric predictability is a prediction that the future will be like the past climatology (Qingyun et al. 2019). Climatology is a forecast of the climatological outcome and is often used as an important reference for the forecast skill (Qingyun et al. 2019). This paper obtains observed samples to calculate climatology, that is, the average precipitation and precipitation distribution of observed samples are taken as the forecast of the climatological outcomes in this study (Konstantinos et al. 2019).

2.3.1.2 Verification metrics of QPFs

The output of EPS is a set of possible values, which not only provides PQPFs (ensembles) but also provides relatively robust QPFs by using the mean of all ensemble members (mean ensemble)

(WMO 2012). As an important output of EPS, multiple deterministic verification metrics are used to demonstrate different aspects of QPFs. Scatter plots (Fig. 3a) and Pearson correlation coefficients are used to measure the linear relationship between forecasts and observations (Qingyun et al. 2019). The RMSE and discrimination diagram are used to evaluate the accuracy and discrimination of QPFs. Accuracy refers to the average difference between individual forecasts and observations, while discrimination represents differences in forecasts for different outcomes.

In the scatter plots (Fig. 3a), the lower correlation between forecasts and observations results in scatter about the one-to-one line. The Pearson correlation coefficient is a measure of the degree of linear correlation between QPFs and observations. A Pearson correlation coefficient of 1 (− 1) indicates a perfect positive (negative) linear correlation between QPFs and observations, while the absence of such a relationship leads to 0.

RMSEs are often used to measure the accuracy of deterministic predictions. The RMSE evaluates the standard deviation of the error between deterministic predictions and the observations. For RMSE, a lower value indicates better accuracy.

The discrimination diagram divides predictions into three types: correct prediction, false positive and false negative (Fig. 3b).

2.3.2 Bayesian model average method

The BMA, developed by the University of Washington (Raftery et al. 2005), is now recognized as one of the best statistical postprocessing methods for constructing multimodel superensemble forecasts (Sloughter et al. 2007). By combining data from different EPSs, BMA generates a single probabilistic prediction in the form of a predictive probability density function (PDF) (Vogel et al. 2017). Given that y is the predictive variable, the output corresponding to model $M_1,\ldots,M_K$ is $f_1,\ldots,f_K$, and for the training dataset, (y^T,f^T),

$$p\left[y\middle|(f_1,\ldots,f_K,y^T)\right]=\sum_{k=1}^{K}\omega_k g_k(y|(f_K,y^T)) \tag{12}$$

$$\sum_{k=1}^{K}\omega_k=1 \tag{13}$$

where $g_k(y|f_k,y^T)$ is the PDF of the M_k EPS and ω_k is the BMA weight of the M_k EPS, reflecting the overall performance of the M_k EPS during the training period.

The default distribution of the variable is a normal distribution in BMA. The accumulated precipitation is zero in many cases; however, the distribution will be highly skewed for cases in which it is not zero. Therefore, a modified condition PDF of BMA is applied to extend BMA. In addition, the BMA variable in this study is taken as the cube root of precipitation to yield a good distribution. (Jianguo 2014; Sloughter et al. 2007).

The modified conditional PDF comprises two parts. The first part calculates the probability distribution of zero precipitation by a logistic regression model:

$$\begin{aligned}\log it\left\{p\left[y=0|(f_k,y^T)\right]\right\}&=\log\frac{p[y=0|(f_k,y^T)]}{p[y>0|(f_k,y^T)]}\\&=a_{0k}+a_{1k}f_k^{1/3}+a_{2k}\delta_k\end{aligned} \tag{14}$$

where a_{0k}, a_{1k}, and a_{2k} are computed by logistic regression.

The second part is the PDF when the precipitation is nonzero, which is represented by a gamma distribution (Sloughter et al. 2007):

$$h_k\left[y|(f_k,y^T)\right]=\frac{1}{\beta_k^{\alpha_k}\Gamma(\alpha_k)}y^{\alpha_k-1}\exp\left(-y/\beta_k\right) \tag{15}$$

where the shape parameters α_k and scale parameters β_k are expressed as follows:

$$\mu_k=\alpha_k\beta_k=b_{0k}+b_{1k}f_k^{1/3} \tag{16}$$

$$\sigma_k^2=\alpha_k\beta_k^2=c_0+c_1 f_k \tag{17}$$

where μ_k and σ_k^2 are the mean and variance in the gamma distribution, respectively; b_{0k} and b_{1k} are calculated by generalized linear regression; and c_0 and c_1 are obtained by using the maximum likelihood method.

In summary, the BMA predictive PDF of the cube root of the accumulated precipitation y is

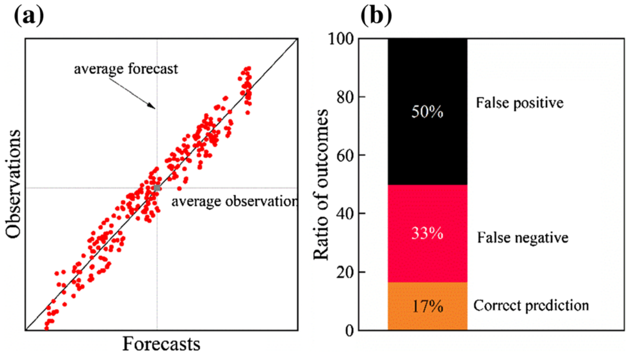

Fig. 3 **a** Schematic diagram of scatter plot; **b** Schematic diagram of discrimination diagram

$$p\left[y|(f_1,\ldots,f_k,y^T)\right] = \sum_{k=1}^{K}\omega_k\{p\left[y=0|(f_k,y^T)\right]I[y=0] + p\left[y>0|(f_k,y^T)\right]h_k\left[y|(f_k,y^T)\right]I[y>0]\} \tag{18}$$

where the general indicator function $I[]$ is 1 if the condition in brackets holds; otherwise, it is zero. Using the maximum likelihood method to calculate ω_k,

$$l(\omega_1,\ldots,\omega_K;c_0;c_1) = \sum_{j=1}^{J}\sum_{i=1}^{n}\log p\left[y_{i,j}|(f_{1,i,j},\ldots,f_{K,i,j},y^T)\right] \tag{19}$$

where i and j represent time and station, respectively; J is the number of total stations; and n is the days of training period. The equation above is maximized numerically by the expectation–maximization (EM) algorithm (Dempster 1977; McLachlan and Krishnan 1988).

3 Results and discussion

3.1 The performances of EPSs for a lead time of 24 h

The precipitation forecast for the lead time of 24 h receives the most attention during flood control. For the lead time of 24 h, the overall performances, the performances at different precipitation thresholds and the spatial distribution of the performances of five EPSs are examined.

3.1.1 The overall performances of EPSs

Figure 4 demonstrates the scatter plots of QPFs versus observations. The abscissa of the black box in Fig. 4 represents the average QPFs, and the ordinate represents the average observations. The Pearson correlation coefficients of QPFs are given as the numbers in the figure. The QPFs of JMA EPS have the best correlation with observations, while CMA EPS has the worst correlation. Table 2 lists the mean RMSEs of five typical EPSs during the verification period, which reflects the QPF accuracies. In line with the correlation results, followed by ECMWF EPS, QPFs of JMA EPS show the best accuracy (i.e., lowest RMSE), and CMA EPS shows the worst accuracy.

Figure 5 illustrates the PQPF skills of each EPS relative to climatology, and a value of CRPSS greater than 0 indicates more forecasting skills than climatology. The PQPFs of all EPSs have positive CRPSSs, which indicates that they are more skillful than climatology (i.e., observed sample). Followed by ECMWF and UKMO EPSs, the mean CRPSS of JMA EPS is the highest, and the

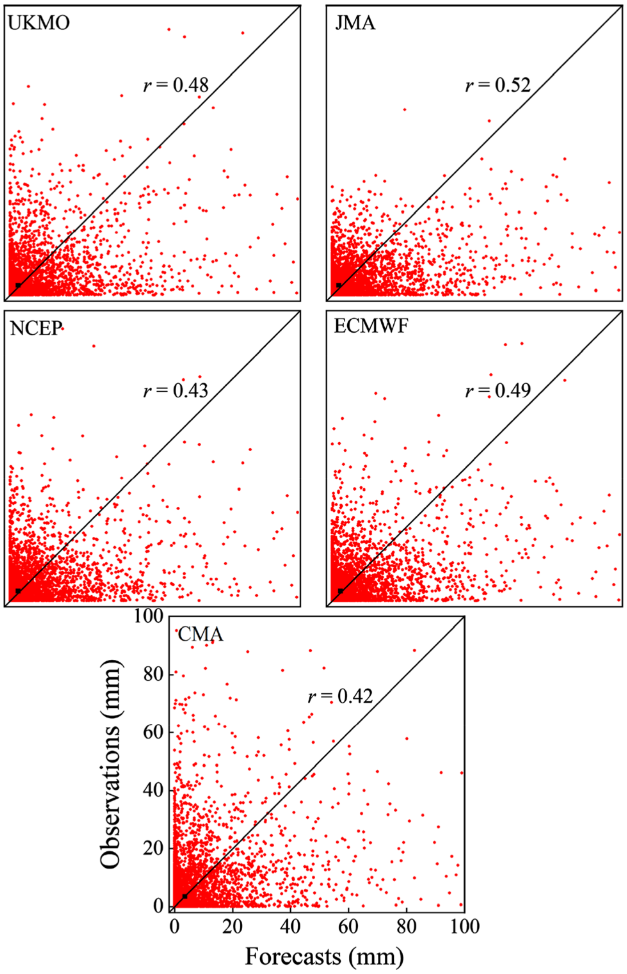

Fig. 4 Scatter plots of QPFs versus observations

Table 2 Basin mean QPF RMSEs of the five EPSs

EPS	RMSE (mm/days)
UKMO	9.30
CMA	10.38
ECMWF	9.12
JMA	8.85
NCEP	9.43

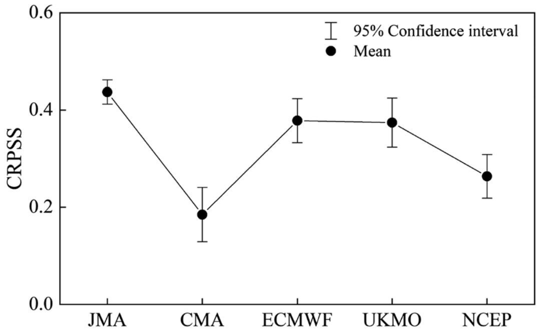

Fig. 5 Basin mean PQPF CRPSSs of EPSs

confidence interval is the narrowest, indicating the better PQPF skill. CMA EPS has the worst PQPF skill.

ECMWF has been proven to be a superior EPS in multiple regions, but the adaptability of other EPSs varies from region to region. For instance, ECMWF and JMA EPSs show the best skills in China's Huai River basin (Tao et al. 2014). Along the coasts of the northern Indian Ocean, ECMWF, UKMO and NCEP EPSs produce more skillful forecasts (Bhomia et al. 2017). In West Africa, the forecasts of ECMWF and UKMO EPSs are the best (Louvet et al. 2016).

3.1.2 EPS performances at different precipitation thresholds

In general, drought relief focuses on the forecast quality at a low precipitation threshold, while flood control concerns the forecast quality at a large precipitation threshold. Therefore, a large threshold and a low threshold are selected in this paper to evaluate the EPS capacity for drought relief and flood control, respectively. The precipitation between the large threshold and low threshold is not considered here.

There are few data points at the threshold of more than 50 mm/days (Table 3) during the verification period, and it is difficult to meet the needs of the reliability diagram. Therefore, 10 mm/days and 25 mm/days are selected as the low threshold and large threshold for dichotomous events, respectively, whereby the quality of QPFs and PQPFs from five EPSs are estimated at two thresholds.

Figure 6 plots the reliability curves of PQPFs for different events. The closer the curve is to the diagonal, the more reliable the PQPF is. For clarity, the EPS with more ensemble members has more probability bins. The BSS, reliability (REL) and resolution (RES) of the BS are shown as numbers in the figure. The horizontal dashed line is the observed sample frequency (i.e., climatology). When the reliability curve is lower than the dashed line, the forecasting skills are inferior to climatology at this forecast probability.

For the dichotomous event at a low precipitation threshold (< 10 mm/days), it largely deviates from the diagonal line as the prediction probability decreases, which is a severe false negative case. For the dichotomous event at a large precipitation threshold (> 25 mm/days), severe false negatives occur with increasing prediction probability. However, in contrast to false negatives, false positives are more advantageous for flood control safety.

Table 3 The proportion of data for different dichotomous events (%)

Threshold	CMA	ECMWF	JMA	NCEP	UKMO
< 10 mm/days	89.75	91.58	93.08	91.78	92.63
> 25 mm/days	3.38	2.37	1.36	1.87	2.38
> 50 mm/days	1.08	0.46	0.08	0.22	0.56

All EPSs have superior PQPFs skill at low precipitation thresholds due to the higher BSS value at low precipitation thresholds. CMA and NCEP EPSs have relatively poor PQPF reliabilities at both thresholds, presenting poor PQPF skills at both thresholds. UKMO EPS is sharper and has the best PQPF skill at a low threshold (the largest BSS), which is mainly attributed to its best reliability and resolution (the smallest REL and the largest RES). For the large precipitation threshold, ECMWF and JMA EPSs have better PQPF skills, where ECMWF has better resolution and is sharper, and JMA is more reliable.

It is easy to calibrate the reliability term through postprocessing, while the resolution term is difficult to postprocess because it is intrinsic to the model (Xiang et al. 2014). Therefore, for flood warnings, the ECMWF EPS is relatively more promising and is expected to further acquire skill through postprocessing.

Figure 7 reveals the discriminations of QPFs at two thresholds. For a low threshold (< 10 mm/days), the ratio of correct prediction is approximately 90% for each EPS, representing the superior QPF discrimination of all EPSs at the low threshold, which is of reference value for drought warnings. For a large threshold (> 25 mm/days), no EPSs can discriminate well. The CMA EPS has almost the same ratio of correct prediction with others, while its false negative ratio is inferior to others. Therefore, in the case of flood control, the QPFs of the CMA EPS are preferred when adopting deterministic forecasts among these EPSs.

Overall, the EPS forecasting skill for precipitation with a large threshold is far lower than that with a low threshold. This result relates to the main precipitation types over the HB and the characteristics of EPS. Typhoons and plum rains are the main sources of precipitation in the HB. The typhoon is a tropical cyclone, and the plum rain belongs to the East Asian monsoon (Wang et al. 2011; Chen et al. 2018). EPS is good at predicting precipitation generated by the above two types (Lan et al. 2011; Olson et al. 1995), so EPS has good forecast quality for low thresholds. However, the accurate forecast of heavier precipitation is a challenge to EPS (Lan et al. 2011). It is obvious that EPSs can play an effective role in drought predictions for the HB. However, when EPSs are used to force the hydrological model to produce a flood forecast, they should be used carefully.

3.1.3 Spatial distribution of EPS performances

The spatial distribution of precipitation is more realistic and accurate in mountainous terrain when elevation

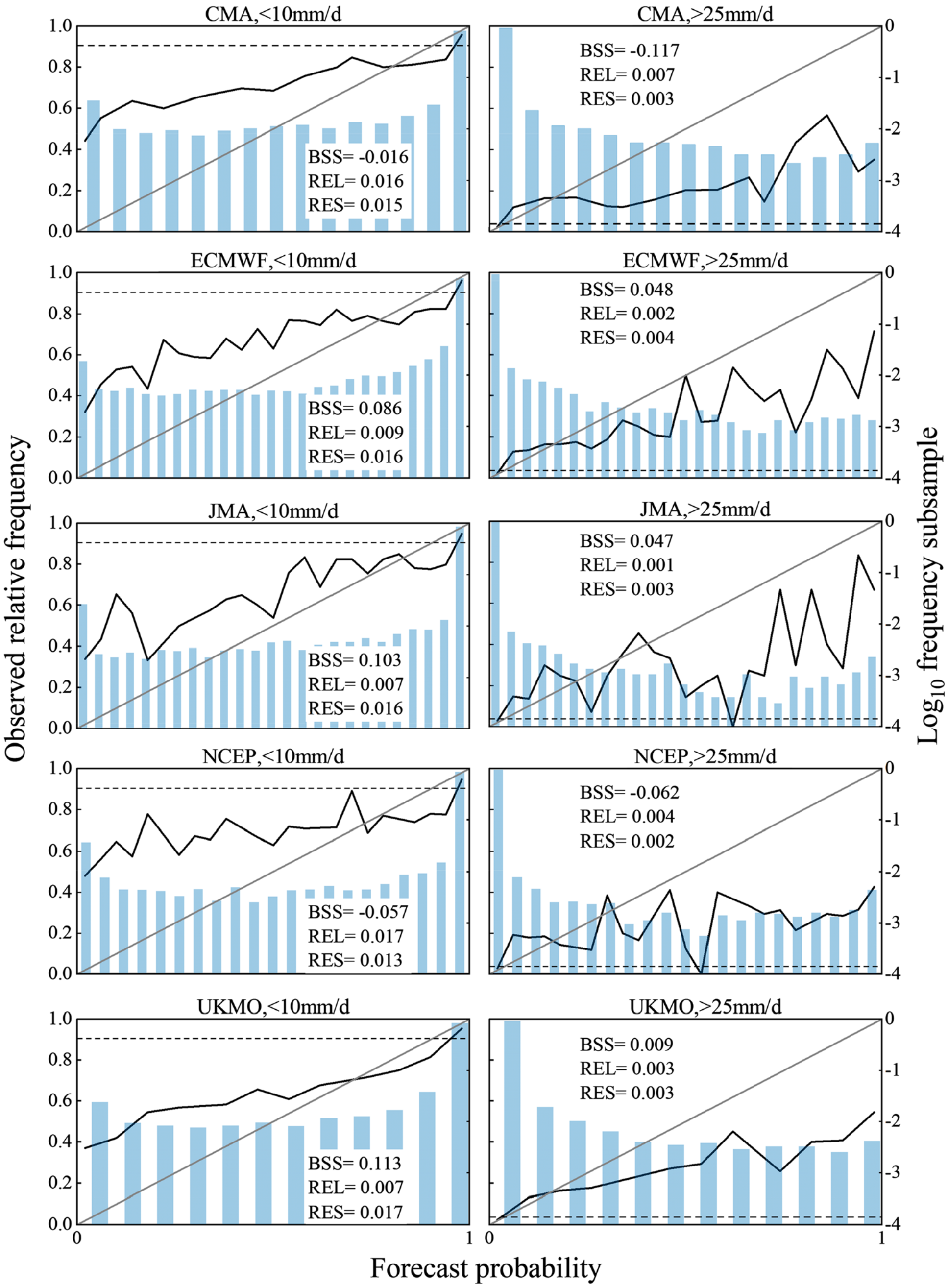

Fig. 6 Reliability diagrams of PQPFs at two thresholds. The bar graphs show the subsample frequencies on the logarithm scale. The horizontal dashed line is the observed sample frequency (i.e., climatology)

dependence is considered (Song et al. 2019). Thus, the interpolation method used in this paper is Gradient plus Inverse-Distance-Square (GIDS) (Price et al. 2000), which can consider the influence of elevation. Affected by the different climates over the HB, the average daily precipitation distribution decreases from South to North during the verification period (Fig. 8). The precipitation distribution in the HB is not only affected by climate but also by topography and geographical location. The precipitation in mountainous areas and coastal areas increases.

In this study, two approaches are applied to evaluate the spatial differences in EPS performances in the HB. First, the mainstream is taken as the dividing line between the northern and southern HB (the Qinling Mountains-Huai River line is the North–South boundary line of China). The mean verification metrics of EPSs in the northern and southern HB are calculated and displayed in Table 4 to study the adaptability of EPSs in the climatic transition zone. Second, the spatial distribution of the verification metrics of EPSs is carried out by the GIDS (Figs. 9, 10), which intuitively describes the spatial changes in the prediction quality of EPSs.

The PQPF skills and QPF accuracies of all EPSs at a large threshold are worse than those at a low threshold

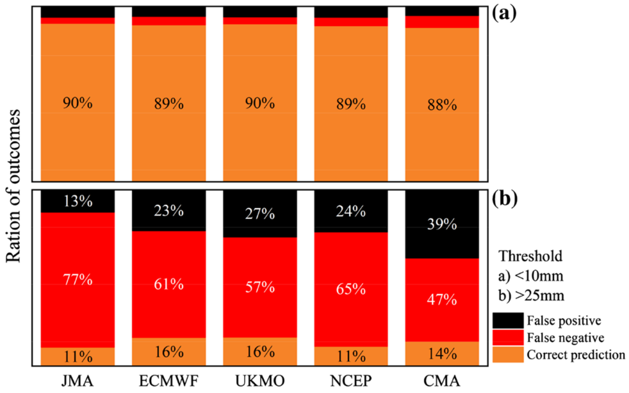

Fig. 7 Discrimination diagrams of QPFs at two-thresholds

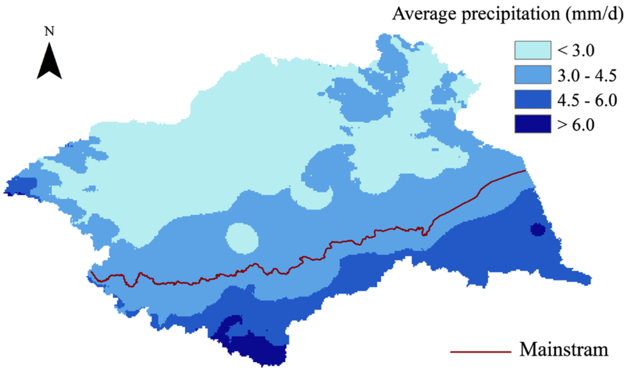

Fig. 8 The average daily precipitation at stations

(Figs. 6, 7); thus, the PQPF skills and QPF accuracies are better in the northern HB than in the southern HB. UKMO EPS has the best PQPF skill in the northern HB due to its better skill at a low threshold, while the PQPF skill of JMA EPS is superior in the southern HB. In terms of QPF accuracy, the ECMWF and JMA EPSs perform the best in the southern and northern HB, respectively. The PQPF skill and QPF accuracy of CMA EPS are the worst in both the northern and southern HB.

The PQPF skill is significantly decreased in the mountainous area (the yellow and green parts in Fig. 9). The PQPF skill distribution is not completely consistent with the precipitation distribution, which is because atmospheric predictability varies with precipitation formation and type (Qingyun et al. 2019). There are many factors affecting precipitation formation and type, such as atmospheric circulation, topography, and geographical location (including lake and ocean effects) (Chen et al. 2018). For PQPF, the forecast of precipitation caused by ocean effects is skillful. However, the precipitation caused by the complex terrain is still very difficult to forecast because the original resolution of EPSs is not adequate (Kaufman et al. 2003).

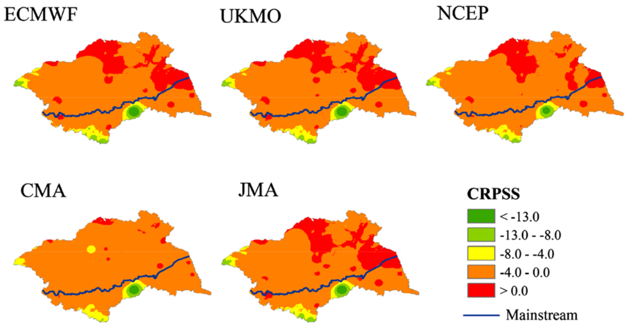

Fig. 9 Spatial distributions of PQPF CRPSSs in the HB

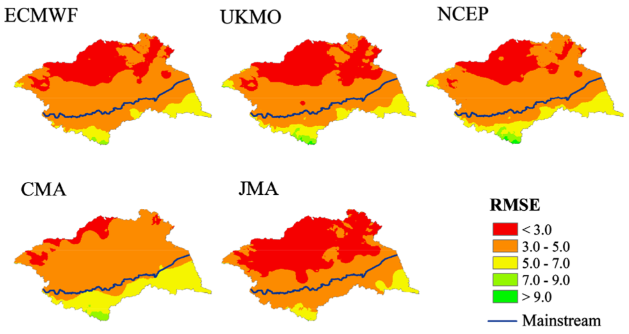

Fig. 10 Distributions of QPF RMSEs in the HB

For QPF, the ensemble mean process eliminates the above skill, giving rise to a QPF accuracy distribution that is similar to the precipitation distribution. In addition, regardless of the amount of precipitation, the QPF accuracies are always low for mountainous areas.

Table 4 Mean verification metrics of EPSs in northern and southern HB

EPS	Southern HB			Northern HB		
	CRPSS	95% confidence intervals for CRPSS	RMSE	CRPSS	95% confidence intervals for CRPSS	RMSE
ECMWF	0.42	0.37–0.47	10.20	0.45	0.41–0.48	6.88
UKMO	0.37	0.31–0.43	10.86	0.47	0.44–0.51	6.81
NCEP	0.30	0.24–0.37	11.18	0.32	0.27–0.36	7.25
CMA	0.21	0.14–0.29	12.85	0.28	0.24–0.33	8.87
JMA	0.45	0.40–0.49	10.51	0.44	0.40–0.47	6.59

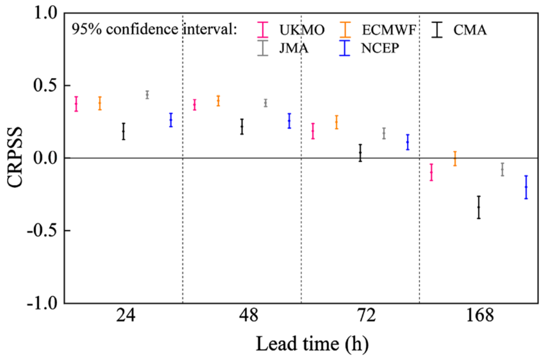

Fig. 11 PQPF CRPSSs of EPSs for different lead times

3.2 The performances of EPSs for different lead times

The longer lead time is more favourable for flood control operation in the future, but it is disadvantageous to forecast accuracy. Hence, it is necessary to analyse the forecast quality of EPSs for different lead times. In this study, we select four lead times of 24 h, 48 h, 72 h and 168 h, and verify the accumulative precipitation forecast for each lead time.

Figure 11 demonstrates the PQPF CRPSSs of EPSs for different lead times, and the 95% confidence intervals are also provided. As the lead time increases, the PQPF skill consistently decreases, and the confidence interval becomes wider. The PQPF skill is poor for the lead time of 168 h; as a result, the forecast for a lead time of 168 h has no practical value. As the lead time increases, the PQPF skill advantage of ECMWF EPS gradually appears. The PQPF of CMA EPS has poor performances for all lead times.

Figure 12 shows the QPF accuracies for different lead times. With increasing lead time, the accuracies of QPFs decrease. For a lead time of 24 h, the QPF accuracy of ECMWF EPS ranks second best following JMA EPS. The QPF accuracy of CMA EPS lags behind the others for all lead times.

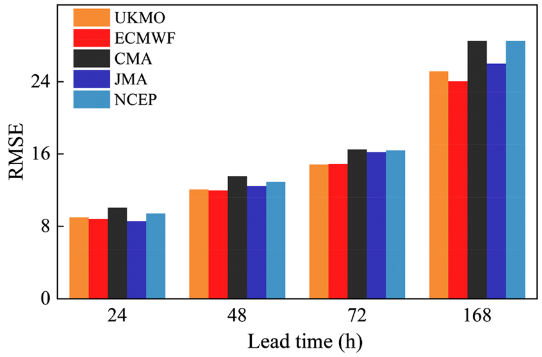

Fig. 12 QPF RMSEs of EPSs for different lead times

In particular, ECMWF EPS shows a good PQPF skill and QPF accuracy for a long lead time because a long lead time requires more ensemble members to obtain the maximum forecast skill (Clark et al. 2011; Richardson 2001). CMA EPS has the least ensemble members among the five EPSs, and it may be a reason for its poor performance. Thus, for long lead times, an important consideration for the selection of EPSs is the number of ensemble members.

3.3 The performance of the multimodel superensemble

The multimodel superensemble is obtained from all members of five EPSs by BMA. For the members of an individual EPS, their weights are constrained to be equal because they are derived from the same model (Robert 2018; Xiang et al. 2014). This section focuses on the flood control support capacity of a multimodel superensemble. Since the flood season in the HB lasts from June to September each year, July 31 to August 31 is selected as the verification period for the multimodel superensemble, and the performances of five individual EPSs during the same period are also verified for comparison.

(1) The length of the BMA training period

The BMA model is reconstructed each day for each station throughout the verification period. The training period is a sliding window, and the parameters are calibrated using the training period of n previous days. In this study, following the references (Bo et al. 2017; Wu et al. 2014), 35 days, 4 days, 45 days, 50 days, 55 days and 60 days are selected as the training sample periods to train the model. The means of CRPSS and RMSE are taken for all stations and for each day in the verification period.

Table 5 lists the results of the training sample periods for a lead time of 24 h. The multimodel superensemble shows the lowest RMSE for 35 days,

Table 5 The mean verification metrics of the multimodel superensemble by different training sample periods for a lead time of 24 h

Training sample period (days)	RMSE	CRPSS
35	13.65	0.43
40	14.24	0.42
45	14.24	0.40
50	14.16	0.44
55	14.20	0.45
60	14.19	0.45

Table 6 The lengths of the BMA training period for different lead times

Lead time (h)	24	48	72	168
Training period (days)	35	55	55	60

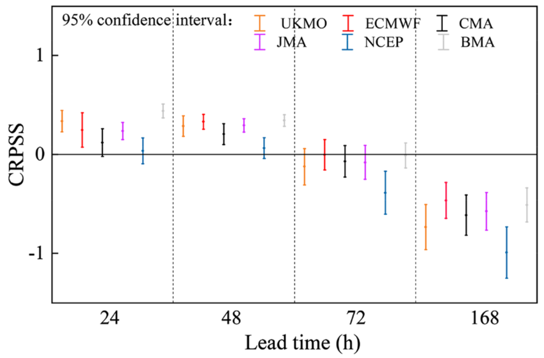

Fig. 13 PQPF CRPSSs of EPSs and multi-model super-ensemble for different lead times

and the discrepancy of CRPSS between 35 days and other training sample periods is quite small. Thus, 35 days is chosen as the length of the training period for the lead time of 24 h. Table 6 lists the training period lengths for different lead times.

(2) The performance of the multimodel superensemble

The multimodel superensemble is expected to show an improved performance compared with all individual EPSs. Figure 13 illustrates the PQPF skills of five EPSs and the multimodel superensemble for different lead times from July 31 to August 31, as well as the 95% confidence intervals. Only at lead times of 24 h and 48 h do EPSs and multimodel superensembles have better PQPF skills than climatology during the flood season. The multimodel superensemble has a slight improvement effect on individual EPSs for all lead times, except 168 h, which manifests a slightly higher CRPSS score and the narrowest confidence interval. However, for the lead time of 168 h, the PQPF skill of the multimodel superensemble ranks second after ECMWF EPS. It seems that 60 days does not meet the training requirement of 168 h.

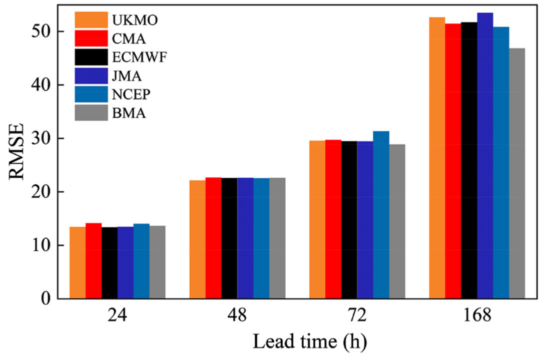

Fig. 14 QPF RMSEs of EPSs and multimodel superensemble for different lead times

At the same time, the QPF accuracy of the multimodel superensemble is expected to be improved compared with that from the five individual EPSs. Figure 14 plots the QPF accuracies of EPSs and multimodel superensembles for different lead times. Contrary to the result of PQPFs, for all lead times except for 168 h, the multimodel superensemble has a slightly improved QPF accuracy compared with individual EPSs, while the multimodel superensemble exhibits a remarkable performance with the highest QPF accuracy for the lead time of 168 h.

Compared with meteorological elements, such as air temperature and wind speed, the statistical postprocessing of precipitation is more difficult to conduct. The reasons are listed as follows by Scheuerer and Hamill (2015): (1) The skewed distribution of precipitation discontinuity is difficult to fit. (2) The difficulty of forecasting increases with increasing precipitation threshold. (3) The shortage of samples for heavy rainfall and rainstorms is also a major problem.

The results of research on the value of multimodel superensembles have been mixed. Hamill (2012) stated that PQPF based on multimodel superensembles has better reliability and prediction skills than PQPF based on individual EPSs. Peter Vogel et al. (2017) found that the BMA was not very good at improving the skill, and the BMA was not much more valuable than climatology for long lead times. This finding is consistent with the results of this paper. Renate Hagedorn et al. (2012) investigated the possibility of combining all available EPSs into a multimodel superensemble and found that ECMWF EPS was a major contributor to the performance improvement, and the multimodel superensemble did not improve much more than ECMWF, which may explain the CRPSS results of 168 h in this paper. Saedi et al. (2020) proved that BMA has a great influence on improving probabilistic prediction, but it is not very effective in deterministic predictions. Ji et al. (2019) further proved that the deterministic prediction constructed by the BMA is accurate for low precipitation thresholds but has limited accuracy for medium and high precipitation thresholds. It is evident that the QPF accuracy for a short lead time is in good agreement with the above two results.

The improvement in the forecast reliability of the multimodel superensemble mainly comes from the potential bias cancelation in different members (Duan et al. 2012). If the forecasting skills of the EPSs are markedly different from each other, then the deterministic prediction through postprocessing is better than the best EPS (Winter and Nychka 2010). If the QPFs released by different EPSs are highly correlated or the best EPS performs significantly better than others, then the QPF through postprocessing cannot always be better than the best individual EPS (Jeong and Kim 2009; Renate Hagedorn et al. 2012; Winter and Nychka 2010). Therefore, for the lead time of 168 h, EPSs have obviously different QPF accuracies (Fig. 12), eliciting a good performance of BMA for improving QPF accuracy.

4 Summary

This study provided a comprehensive verification of QPFs and PQPFs from five operational global EPSs in the HB from April to December 2015. Focusing on the lead time of 24 h, the forecast qualities are evaluated in terms of overall performance, different thresholds and spatial adaptability. The forecast qualities for different lead times are later assessed. In addition, for different lead times, BMA was used to integrate all members of the five EPSs, and the overall performance of the multimodel superensemble in the main flood season of the HB was verified. The main conclusions are listed as follows:

(1) As the ECMWF EPS has the largest ensemble members, the ECMWF EPS has the best forecast quality both in QPFs and PQPFs for longer lead times. CMA EPS has the least ensemble members, which may account for its poor forecast quality for all lead times.

(2) EPS has a reference value for drought warnings in the HB. The PQPF of the ECMWF EPS has a potential ability for the prediction of intense precipitation.

(3) For long lead times, a large number of ensemble members is valuable for high forecast quality, so computing resources should be allocated to increase the ensemble members. According to the spatial distribution of EPS performances, for a lead time of 24 h, resources should be focused on the development of higher resolution, which is conducive to increasing the forecasting skills for various types of precipitation.

(4) Owing to the climate transitional zone over the HB, EPS forecast quality is better in the northern HB than in the southern HB. Furthermore, the PQPF skill is also affected by the precipitation type. PQPF is skillful for forecasting the precipitation caused by the ocean effect but is poor for predicting the precipitation affected by mountain topography. The QPF accuracy is also influenced by the terrain, causing it to decrease in mountainous areas.

(5) The multimodel superensemble has slightly improved the PQPF skill for short lead times, and for long lead times, it is not much more valuable than climatology. When the QPF accuracy of each individual EPS is significantly different, the multimodel superensemble will obtain an improved QPF accuracy.

The results of this study are only applicable to a specific river basin, but the analytical method for the adaptability of ensemble forecasting over a river basin is generally applicable. This result not only provides a detailed feedback report for the precipitation ensemble forecast model but also provides information for the subsequent watershed flood forecast based on precipitation ensemble forecasts. Limited by the synchronization of the verification period of prediction data sets and observed data, only a portion of the 2015 period is used in this study. Therefore, future work should include a longer verification period to derive a more general conclusion. In future studies, many other postprocessing methods should be tested and compared (Aminyavari and Saghafian 2019; Huo et al. 2019; Shin et al. 2019). Furthermore, gridded datasets may be helpful to further improve the accuracy of the assessment of EPS.

Author's contribution Methodology, WH; conceptualization, ZPA; software, WH; validation, ZFL; formal analysis, WH, SX and MYF.; writing—original draft, WH; writing—review and editing, WH and ZFL; visualization, ZFL.; supervision, ZPA.; funding acquisition, ZPA.

Funding This study is supported by the National Key Technologies R&D Program of China (Grant No. 2017YFC0405606), the Fundamental Research Funds for the Central Universities (Grant No. B200202032), the China Postdoctoral Science Foundation (Grant No. 2019M661715), National Natural Science Foundation of China, China (Grant No. 51909062), National Natural Science Foundation of Jiangsu Province, China (Grant No. BK20180509).

Availability of data and materials All data used during the study are available from the first author by request.

Code availability Codes used during the study are available from the first author by request.

Compliance with ethical standards

Conflict of interest There is no conflict of interest regarding the publication of this article.

References

Aminyavari S, Saghafian B (2019) Probabilistic streamflow forecast based on spatial post-processing of TIGGE precipitation forecasts. Stoch Environ Res Risk Assess 33:1939–1950. https://doi.org/10.1007/s00477-019-01737-4

Aminyavari S, Saghafian B, Delavar M (2018) Evaluation of TIGGE ensemble forecasts of precipitation in distinct climate regions in Iran. Adv Atmos Sci 35(4):457–468

Bhomia S, Jaiswal N, Kishtawal CM (2017) Accuracy assessment of rainfall prediction by global models during the landfall of tropical cyclones in the North Indian Ocean. Meteorol Appl 24:503–511

Bischiniotis K, van den Hurk B, Zsoter E, Coughlan de Perez E, Grillakis M, Aerts JCJH (2019) Evaluation of a global ensemble flood prediction system in Peru. Hydrol Sci J 64:1171–1189. https://doi.org/10.1080/02626667.2019.1617868

Bo Qu, Xingnan Z, Florian P, Tao Z, Yuanhao F (2017) Multi-model grand ensemble hydrologic forecasting in the Fu river basin using Bayesian model averaging. Water 9:74. https://doi.org/10.3390/w9020074

Bonnardot F, Quetelard H, Jumaux G, Leroux MD, Bessafi M (2018) Probabilistic forecasts of tropical cyclone tracks and intensities in the southwest Indian Ocean basin. Q J R Meteorol Soc 145:675–686

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

Chen X, Yuan H, Xue M (2018) Spatial spread-skill relationship in terms of agreement scales for precipitation forecasts in a convection-allowing ensemble. Q J R Meteorol Soc 144:85–98. https://doi.org/10.1002/qj.3186

Clark AJ, Kain JS, Stensrud DJ, Xue M, Kong F, Coniglio MC, Thomas KW, Wang Y, Brewster K, Gao J, Wang X, Weiss SJ, Du J (2011) Precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. Mon Weather Rev 139:1410–1418

Cloke HL, Pappenberger F (2009) Ensemble flood forecasting: a review. J Hydrol 375:613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Dempster AP (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc Series B: Methodologic 39:1–38

Demargne J, Brown J, Liu Y, Seo DJ, Wu L, Toth Z, Zhu Y (2010) Diagnostic verification of hydrometeorological and hydrologic ensembles. Atmos Sci Lett 11:114–122

Duan Y, Gong J, Du J, Charron M, Chen J, Deng G, DiMego G, Hara M, Kunii M, Li X, Li Y, Saito K, Seko H, Wang Y, Wittmann C (2012) An overview of the Beijing 2008 olympics research and development project (B08RDP). Bull Am Meteorol Soc 93:381–403

Hagedorn R, Buizza R, Hamill TM, Leutbecher M, Palme TN (2012) Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. Q J R Meteorol Soc 138:1814–1827

Hamill TM (2012) Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. Mon Weather Rev 140:2232–2252. https://doi.org/10.1175/MWR-D-11-00220.1

Hemri S, Scheuerer M, Pappenberger F, Bogner K, Haiden T (2014) Trends in the predictive performance of raw ensemble weather forecasts. Geophys Res Lett 41:9197–9205

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15:559–570

Huo W, Li Z, Wang J et al (2019) Multiple hydrological models comparison and an improved Bayesian model averaging approach for ensemble prediction over semi-humid regions. Stoch Environ Res Risk Assess 33:217–238

Jeffrey SW, Thomas MH (2002) Ensemble data assimilation without perturbed observations. Mon Weather Rev 130:1913–1924

Jeong DI, Kim Y-O (2009) Combining single-value streamflow forecasts—A review and guidelines for selecting techniques. J Hydrol 377:284–299

Ji L, Zhi X, Zhu S, Fraedrich K (2019) Probabilistic precipitation forecasting over East Asia using Bayesian model averaging. Weather Forecast 34:377–392. https://doi.org/10.1175/WAF-D-18-0093.1

Jianguo XZL (2014) BMA probabilistic quantitative precipitation forecasting over the Huaihe Basin using TIGGE multimodel ensemble forecasts. Mon Weather Rev 142:1542–1555. https://doi.org/10.1175/MWR-D-13-00031.1

Karuna Sagar S, Rajeevan M, Vijaya Bhaskara Rao S, Mitra AK (2017) Prediction skill of rainstorm events over India in the TIGGE weather prediction models. Atmos Res 198:194–204. https://doi.org/10.1016/j.atmosres.2017.08.017

Kaufmann P, Schubiger F, Binder P (2003) Precipitation forecasting by a mesoscale numerical weather prediction (NWP) model: 8 years of experience. Hydrol Earth Syst Sci 7:812–832

Kirtman BP et al (2014) The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bull Am Meteor Soc 95:585–601

Krishnamurti TN, Kishtawal CM, LaRow TE, Bachiochi DR, Zhang Z, Williford CE, Gadgil S, Surendran S (1999) Improved weather and seasonal climate forecasts from multimodel superensemble. Science 285:1548–1550

Krishnamurti TN, Kumar V, Simon A, Bhardwaj A, Ghosh T, Ross R (2016) A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. Rev Geophys 54:336–377

Lan C, Pagano Thomas C, Wang QJ (2011) A Review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting. J Hydrometeorol 12:713–728. https://doi.org/10.1175/2011JHM1347.1

Louvet S, Sultan B, Janicot S, Kamsu Tamo PH, Ndiaye O (2016) Evaluation of TIGGE precipitation forecasts over West Africa at intraseasonal timescale. Clim Dyn 47:31–47. https://doi.org/10.1007/s00382-015-2820-x

McLachlan GJ, Krishnan T (1998) The EM algorithm and extensions. Stat Med 17:1187

Meng Zhiyong ZF (2011) Limited-area ensemble-based data assimilation. Mon Weather Rev 139:2025–2045. https://doi.org/10.1175/2011MWR3418.1

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

Olson DA, Junker NW, Korty B (1995) Evaluation of 33 years of quantitative precipitation forecasting at the NMC. Weather Forecast 10:498–511

Osinski R, Bouttier F (2018) Short-range probabilistic forecasting of convective risks for aviation based on a lagged-average-forecast ensemble approach. Meteorol Appl 25:105–118

Pappenberger F, Beven KJ, Hunter NM, Bates PD, Gouweleeuw BT, Thielen J, de Roo APJ (2005) Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). Hydrol Earth Syst Sci 9:381–393

Park YY, Buizza R, Leutbecher M (2008) TIGGE: preliminary results on comparing and combining ensembles. Q J R Meteorol Soc 134:2029–2050

Price DT, McKenney DW, Nalder IA, Hutchinson MF, Kesteven JL (2000) A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. Agric Forest Meteorol 101:81–94

Qingyun D, Florian P, Andy W, Cloke Hannah L, Schaake John C (2019) Handbook of hydrometeorological ensemble forecasting. Springer, Berlin Heidelberg

Raftery AE, Tilmann G, Balabdaoui F, Polakowski M (2005) Using bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev133 (5):1155–1174. https://doi.org/10.1175/MWR2906.1

Richardson DS (2001) Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. Q J R Meteorol Soc 127:2473–2489

Roberto B (2019) Introduction to the special issue on 25 years of ensemble forecasting. Q J R Meteorol Soc 145:1–11. https://doi.org/10.1002/qj.3370

Roebber Paul J, Schultz David M, Colle Brian A, Stensrud David J (2004) Toward improved prediction: high-resolution and ensemble modeling systems in operations. Weather Forecast 19:936–949

Saedi A, Saghafian B, Moazami S, Aminyavari S (2020) Performance evaluation of sub-daily ensemble precipitation forecasts. Meteorol Appl 27:6. https://doi.org/10.1002/met.1872

Scheuerer M, Hamill TM (2015) Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. Mon Weather Rev 143:4578–4596

Shin Y, Lee Y, Choi J, Park J-S (2019) Integration of max-stable processes and Bayesian model averaging to predict extreme climatic events in multi-model ensembles. Stoch Environ Res Risk Assess 33:47–57

Sloughter JML, Raftery AE, Gneiting T, Fraley C (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon Weather Rev 135:3209–3220. https://doi.org/10.1175/MWR3441.1

Song L, Chen M, Gao F, Cheng C, Chen M, Yang L, Wang Y (2019) Elevation influence on rainfall and a parameterization algorithm in the Beijing area. J Meteorologic Res 33(6):1143–1156

Tao Y, Duan Q, Ye A, Gong W, Di Z, Xiao M, Hsu K (2014) An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai river basin. J Hydrol 519:2890–2905

Taylor JW, Buizza R (2003) Using weather ensemble predictions in electricity demand forecasting. Int J Forecast 19:57–70

Thomas MH, Josip J (2006) Measuring forecast skill: is it real skill or is it the varying climatology? Q J R Meteorol Soc 132:2905–2923. https://doi.org/10.1256/qj.06.25

Tilmann G, Raftery EA (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102:359–378. https://doi.org/10.1198/016214506000001437

Trenberth KEE (1992) Climate system modeling. Cambridge University Press, Cambridge

Verlinden KLB (2017) Using the second-generation GEFS reforecasts to predict ceiling, visibility, and aviation flight category. Weather Forecast 32:1765–1780

Vogel P, Knippertz P, Fink AH, Schlueter A, Gneiting T (2017) Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. Statistics. https://doi.org/10.1175/WAF-D-17-0127.s1

Wang H (2017) Preface to the special issue on the forecast and evaluation of meteorological disasters (FEMD). Adv Atmos Sci 34(2):127

Wang B, Ding QH, Liu J (2011) Concept of global monsoon. In: Chang C-P, Ding Y, Lau N-C, Johnson RH, Wang B, Yasunari T (eds) The global monsoon system: research and forecast. World Scientific, Singapore, pp 3–14

Wilks DS (2009) Statistical methods in the atmospheric sciences, 2nd edition. International geophysics series, vol 91. Elsevier: Amsterdam

Winter CL, Nychka D (2010) Forecasting skill of model averages. Stoch Environ Res Risk Assess 24:633–638

WMO (2012) Guidelines on ensemble prediction systems and forecasting, Switzerland

Wu Juan Lu, Zhiyong GW (2014) Flood forecasts based on multi-model ensemble precipitation forecasting using a coupled atmospheric-hydrological modeling system. Nat Hazards 74:325–340. https://doi.org/10.1007/s11069-014-1204-6

Xiang Su, Huiling Y, Yuejian Z, Yan L, Yuan W (2014) Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012. J Geophys Res Atmos 119:7292–7310

Ye J, He Y, Pappenberger F, Cloke HL, Manful DY, Li Z (2014) Evaluation of ECMWF medium-range ensemble forecasts of precipitation for river basins. Q J R Meteorol Soc 140:1615–1628

Ying H, Yuan W, Hao W (2019) Evaluation of Multi-NWPs rainstorm forecasting performance in different time scales in Huaihe River basin and discussion on flood predictability. Meteorol Mon 45:989–1000

Zhang Xu, Qianjin D, Chen J (2019) Comparison of ensemble models for drought prediction based on climate indexes. Stoch Environ Res Risk Assess 33:593–606