BMA Probabilistic Quantitative Precipitation Forecasting over the Huaihe Basin Using TIGGE Multimodel Ensemble Forecasts

JIANGUO LIU

State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, and High Performance Computing Center, Department of Mathematics and Applied Mathematics, Huaihua University, Huaihua, Hunan, and University of Chinese Academy of Sciences, Beijing, China

ZHENGHUI XIE

State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

(Manuscript received 24 January 2013, in final form 23 November 2013)

ABSTRACT

Bayesian model averaging (BMA) probability quantitative precipitation forecast (PQPF) models were established by calibrating their parameters using 1–7-day ensemble forecasts of 24-h accumulated precipitation, and observations from 43 meteorological stations in the Huaihe Basin. Forecasts were provided by four single-center (model) ensemble prediction systems (EPSs) and their multicenter (model) grand ensemble systems, which consider exchangeable members (EGE) in The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE). The four single-center EPSs were from the China Meteorological Administration (CMA), the European Centre for Medium-Range Weather Forecasts (ECMWF), the National Centers for Environment Prediction (NCEP), and the Met Office (UKMO). Comparisons between the raw ensemble, logistic regression, and BMA for PQPFs suggested that the BMA predictive models performed better than the raw ensemble forecasts and logistic regression. The verification and comparison of five BMA EPSs for PQPFs in the study area showed that the UKMO and ECMWF were a little superior to the NCEP and CMA in general for lead times of 1–7 days for the single-center EPSs. The BMA model for EGE outperformed those for single-center EPSs for all 1–7-day ensemble forecasts, and mostly improved the quality of PQPF. Based on the percentile forecasts from the BMA predictive PDFs for EGE, a heavy-precipitation warning scheme is proposed for the test area.

1. Introduction

Rainfall is one of the most important weather phenomena, and improvement of quantitative precipitation forecasts (QPFs) is a primary goal of operational prediction centers and a major challenge facing the research community (Fritsch et al. 1998; Gourley and Vieux 2005). To understand the limits of deterministic prediction of the atmospheric state by setting initial state conditions, ensemble forecasting methods have been developed to improve the capabilities of QPFs and probabilistic quantitative precipitation forecasts (PQPFs; Mullen and Buizza 2001; Gneiting and Raftery 2005). The ensemble prediction system (EPS) forecasts from a single forecast center (model) addressed only the uncertainties inherent in the initial conditions (including different initial condition generation methods and different assimilation systems) in numerical weather prediction (NWP; Zhu 2005). By contrast, the grand ensemble, which incorporates the EPS forecasts from multiple forecast centers, can improve the accuracy of numerical weather and climate forecasts by capturing several of the uncertainties in the initial conditions, lateral boundary conditions, and model physics (dynamics and physical parameterizations), and are potentially able to provide a better representation of the true predictive probability distribution (Molteni

Corresponding author address: Professor/Dr. Zhenghui Xie, State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, P.O. Box 9804, Beijing, 100029, China.
E-mail: zxie@lasg.iap.ac.cn

DOI: 10.1175/MWR-D-13-00031.1

et al. 1996; Grimit and Mass 2002; Barnston et al. 2003; Palmer et al. 2004; Hagedorn et al. 2005; Goswami et al. 2007).

Realizing the full potential of ensemble forecasts requires statistical postprocessing of the model output. In tandem with statistical postprocessing, ensembles can give flow-dependent probabilistic forecasts in the form of probability distributions for the predictive variables (Gneiting and Raftery 2005). Bayesian model averaging (BMA) proposed by Raftery et al. (2005) is a statistical postprocessing method, which produces probabilistic forecasts in the form of a predictive probability density function (PDF) from ensembles of dynamic models, and the PDF is a weighted average of PDFs centered on the bias-corrected forecasts from a set of ensemble members. This approach was further developed by Sloughter et al. (2007, 2010) and Fraley et al. (2010). Studies applying the BMA method to a range of predictive variables have demonstrated that the BMA postprocessed PDFs outperformed the unprocessed ensemble forecast, and were well calibrated and accurate (Wilson et al. 2007; Duan et al. 2007; Vrugt et al. 2008; Yang et al. 2009; Schmeits and Kok 2010; Lee et al. 2012; Liu et al. 2013).

To date, BMA has mainly been used for single-center (model) EPSs in probabilistic weather forecasts. However, little previous work has been done on the application of BMA in a multicenter (model) grand ensemble system. A key component of The Observing System Research and Predictability Experiment (THORPEX), the THORPEX Interactive Grand Global Ensemble (TIGGE), provides a grand ensemble that incorporates the EPS forecasts from 10 forecast centers: the Australian Bureau of Meteorology (BoM), the China Meteorological Administration (CMA), the Meteorological Service of Canada (CMC), the Brazil Centro de Previsao Tempo e Estudos Climaticos (CPTEC), the European Centre for Medium-Range Weather Forecasts (ECMWF), the Japan Meteorological Agency (JMA), the Korea Meteorological Administration (KMA), Météo-France, the National Centers for Environment Prediction (NCEP), and the Met Office (UKMO). This provides a very good basis for the production of PQPFs (Richardson 2005; Park et al. 2008; Matsueda and Tanaka 2008; Zhao et al. 2011). The primary aim of this work is to explore the advantage of the BMA strategy for PQPFs based on the TIGGE multicenter EPS. For the multicenter grand ensemble system, the members from the same center are statistically indistinguishable, because they are initial value perturbed forecasts and derived from a single model, and should be treated as exchangeable. In the present study, BMA methods were applied to four single-center EPSs (CMA, ECMWF, NCEP, UKMO), and the multicenter grand ensemble system that incorporates CMA, ECMWF, NCEP, and UKMO after considering exchangeable members (EGE), to verify and evaluate the performances of the five BMA prediction systems for PQPFs. Finally, a heavy-precipitation warning scheme based on the BMA PQPFs was proposed.

The paper is organized as follows. In section 2 we briefly describe the BMA ensemble postprocessing procedure; we give a brief summary of the study area and datasets in section 3. In section 4 we present the results of the BMA PQPFs for TIGGE multimodel ensemble forecasts of 24-h accumulative precipitation for lead times of 1–7 days. We conclude with a summary and discussion in section 5.

2. Ensemble postprocessing using Bayesian model averaging

BMA is a way of combining inferences and predictions from multiple statistical models (Leamer 1978; Kass and Raftery 1995; Hoeting et al. 1999). Raftery et al. (2005) applied the BMA method to ensembles of dynamic models for surface air temperature and sea level pressure. The BMA predictive PDF is a mixture of the component PDFs:

$$p[\mathbf{y} \mid (f_1, \ldots, f_K, \mathbf{y}^{\mathrm{T}})] = \sum_{k=1}^{K} w_k p_k[\mathbf{y} \mid (f_k, \mathbf{y}^{\mathrm{T}})], \quad (1)$$

where $\mathbf{y}$ is the predictive variable, $p_k[\mathbf{y} \mid (f_k, \mathbf{y}^{\mathrm{T}})]$ is the component forecast PDF based on model f_k alone, w_k is the posterior probability of forecast k such that nonnegative and $\sum_{k=1}^{K} w_k = 1$, and K is the number of models being combined.

Accumulated precipitation is zero in many cases, and for cases in which it is not zero its distribution is highly skewed, thus the Gaussian distribution does not fit this kind of data. Sloughter et al. (2007) gave an extensive treatment of the BMA method for PQPFs, and applied this extended BMA method to daily 48-h forecasts of 24-h accumulated precipitation in the North American Pacific Northwest in 2003–04 using the Washington mesoscale ensemble. It was found that the extended BMA method yielded a well-calibrated and sharp precipitation distribution. We follow Sloughter et al. (2007) in using the cube root of accumulated precipitation as the predictive variable, and the conditional PDF $p_k[\mathbf{y} \mid (f_k, \mathbf{y}^{\mathrm{T}})]$ in Eq. (1) includes two parts. The first part computes the probability of zero precipitation as a function of f_k by a logistic regression model:

$$\text{logit}\{p[\mathbf{y}=0\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]\}=\log\frac{p[\mathbf{y}=0\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]}{p[\mathbf{y}>0\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]}=a_{0k}+a_{1k}f_k^{1/3}+a_{2k}\delta_k,\quad(2)$$

where δ_k is equal to 1 if $f_k=0$, otherwise, it is equal to 0. Here a_{0k}, a_{1k}, a_{2k} are parameters to be estimated. The second part computes the PDF of the precipitation amount $h_k[\mathbf{y}\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]$, given that it is nonzero, by gamma distribution:

$$h_k[\mathbf{y}\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]=\frac{1}{\beta_k^{\alpha_k}\Gamma(\alpha_k)}\mathbf{y}^{\alpha_k-1}\exp(-\mathbf{y}/\beta_k).\quad(3)$$

The shape parameter α_k and scale parameter β_k of the gamma distribution are given as

$$\mu_k=\alpha_k\beta_k=b_{0k}+b_{1k}f_k^{1/3},\quad(4)$$

$$\sigma_k^2=\alpha_k\beta_k^2=c_0+c_1f_k,\quad(5)$$

where μ_k and σ_k^2 are the mean and variance of the gamma distribution, and b_{0k}, b_{1k}, c_0, c_1 are parameters to be estimated. Thus, the BMA predictive PDF of the cube root of the accumulated precipitation $\mathbf{y}$ for a K-member ensemble is

$$p[\mathbf{y}\,|\,(f_1,\ldots,f_K,\mathbf{y}^{\mathrm{T}})]=\sum_{k=1}^{K}w_k\{p[\mathbf{y}=0\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]I[\mathbf{y}=0]+p[\mathbf{y}>0\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]h_k[\mathbf{y}\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]I[\mathbf{y}>0]\},\quad(6)$$

where the general indicator function $I[\ ]$ is unity if the condition in brackets holds; otherwise, it is zero. The $p[\mathbf{y}=0\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]$ is specified by Eq. (2), and the gamma distribution $h_k[\mathbf{y}\,|\,(f_k,\mathbf{y}^{\mathrm{T}})]$ is specified by Eq. (3). Parameter estimation is based on training data for all stations in the study area from a training period, which we take here to be the N days of the forecast and corresponding observation data preceding initialization, following Sloughter et al. (2007). The parameters a_{0k}, a_{1k}, a_{2k} are estimated by logistic regression as in Eq. (2). The parameters b_{0k}, b_{1k} are resolved by generalized linear regression as in Eq. (4), and the parameters w_k, c_0, c_1 are estimated by the maximum likelihood function (Fisher 1922), which is

$$l(w_1,\ldots,w_K;c_0;c_1)=\sum_{s,t}\log p[\mathbf{y}_{s,t}|(f_{1,s,t},\ldots,f_{K,s,t},\mathbf{y}^{\mathrm{T}})],\quad(7)$$

where the s and t index observations in the training set correspond to indices of time and space. This function is maximized numerically using the expectation–maximization (EM) algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997). The details of parameter estimations can be found in Sloughter et al. (2007) and are not repeated here.

In BMA for ensemble forecasting with exchangeable members (Fraley et al. 2010), assuming that there are I groups, and there are m_i exchangeable members in the ith exchangeable group, then $m=\sum_{i=1}^{I}m_i$ is the total number of ensemble members. Let $f_{i,j}$ denote the jth member of the ith group. Then the BMA predictive PDF in Eq. (1) can be rewritten to accommodate groups of exchangeable members:

$$p[\mathbf{y}\,|\,(\{f_{i,j}\}_{i=1,\ldots,I,j=1,\ldots,m_i},\mathbf{y}^{\mathrm{T}})]=\sum_{i=1}^{I}\sum_{j=1}^{m_i}w_ip_i[\mathbf{y}\,|\,(f_{i,j},\mathbf{y}^{\mathrm{T}})].\quad(8)$$

Combining Eqs. (6) and (8), we get the following model:

$$p[\mathbf{y}\,|\,(\{f_{i,j}\}_{i=1,\ldots,I,j=1,\ldots,m_i},\mathbf{y}^{\mathrm{T}})]=\sum_{i=1}^{I}\sum_{j=1}^{m_i}w_i\{p[\mathbf{y}=0\,|\,(f_{i,j},\mathbf{y}^{\mathrm{T}})]I[\mathbf{y}=0]+p[\mathbf{y}>0\,|\,(f_{i,j},\mathbf{y}^{\mathrm{T}})]h_i[\mathbf{y}\,|\,(f_{i,j},\mathbf{y}^{\mathrm{T}})]I[\mathbf{y}>0]\}.\quad(9)$$

The parameter estimation in Eq. (9) is similar to that in Eq. (6), and requires modifying the probability maximum likelihood estimation method. The details of parameter estimation can be found in Fraley et al. (2010).

3. The study area and datasets

The Huaihe Basin is located about midway between the Yellow River and Yangtze River, the two largest rivers in China, with a basin area of about 270 000 km^2. Figure 1 is a map showing the location of the Huaihe Basin and the 43 meteorological stations in the basin. The main reason for choosing the Huaihe Basin was its susceptibility to intense precipitation events in summer and frequent severe flooding. Additionally, there is a dense rain gauge network in the study area.

The datasets used in this study were of observed precipitation from the rain gauge network in the Huaihe Basin (Fig. 1) and predicted precipitation data from the TIGGE ensemble forecasts. The 43 rain gauges in the study area recorded hourly accumulations of precipitation. The 24-h (2000–2000 Beijing local time) accumulated precipitation data from the rain gauges were calculated with quality control implemented by examining the coherence in time and space, and consistency with the synoptic situations. The ECMWF, NCEP, UKMO, and CMA EPS multimember 24-h accumulated precipitation forecasts with

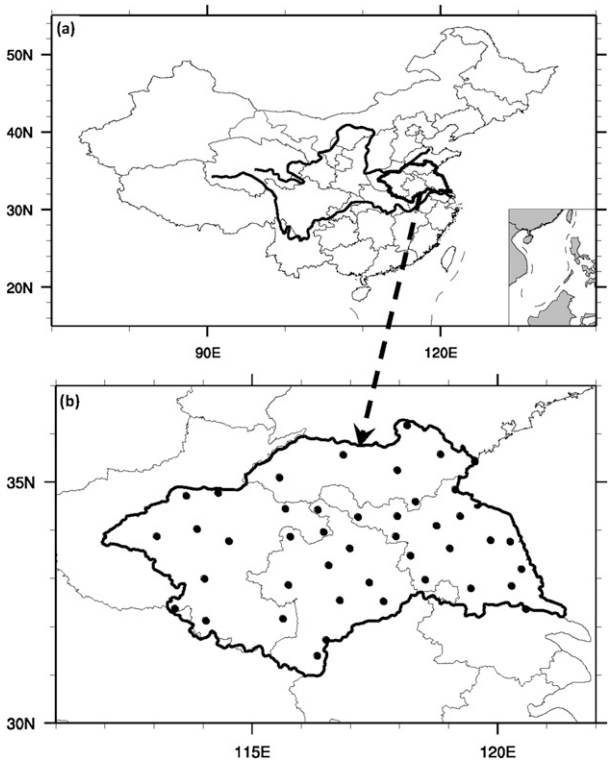

FIG. 1. The study domain: (a) location and (b) study area, showing the locations of 43 meteorological stations.

lead times of 1–7 days with initial time 1200 UTC obtained from the TIGGE–ECMWF portal were used in this study. The test period lasted from 1 June to 31 August 2007. Detailed information of the four single-center EPSs are listed in Table 1. The predicted precipitation of each EPS was interpolated to the 43 individual stations by the bilinear interpolation method. Both the observation and prediction of 24-h accumulated precipitation amounts less than 0.01 mm day^{-1} were deemed ''no precipitation.''

4. BMA probabilistic quantitative precipitation forecasting

a. Experiment design and verification method

The experiments for BMA PQPFs were conducted in a recursive mode: the BMA was retrained each day throughout the verification period for each weather station in the study area, using a training sample period of N previous days (the length of the BMA training period), which is associated with the predictive variable and study area and needed to be determined (section 4b). Thus the BMA model was established dynamically (i.e., the training was carried out separately for each day in the verification period for each station in the study area). In this way, verification data were accumulated to verify the performance of the technique during the verification period.

In the BMA we developed and used, we first determined the length of the BMA model training period over the study area in the study time period. Next, four single-center ensembles and the multicenter EGE were calibrated using BMA separately. For the TIGGE single-center ensemble forecasts, the ensemble members were regarded as exchangeable members because they were derived from a single model, and the BMA weights were constrained to be equal and have a value of $1/n$ (n is the number of the ensemble members; Raftery et al. 2005; Sloughter et al. 2007; Schmeits and Kok 2010). Meanwhile, the parameters in Eqs. (2) and (4) were the same for each ensemble member [i.e., $a_{ik} = a_i(i = 0, 1, 2), b_{jk} = b_j(j = 0, 1)$]. Hence, only a_0, a_1, a_2, b_0, and b_1 have to be estimated using the regression method and c_0, c_1 have to be estimated using the maximum likelihood technique (Schmeits and Kok 2010). For the TIGGE multicenter grand ensemble forecast with exchangeable members (EGE), the BMA model in Eq. (9) was employed to calibrate the grand ensemble. In this case, the four distinct BMA weights and sets of parameter values need to be estimated first for four different models (centers) in the EGE, and then each model's weight was split to the ensemble members within that EPS. The ensemble members from the same EPS have equal BMA weights and parameters. The BMA weights ω_k and parameters a_{0k}, a_{1k}, a_{2k}, b_{0k}, b_{1k}, c_0, c_1 in this case were estimated using regression and the maximum likelihood technique in which the values were obtained iteratively using the EM algorithm, as described in section 2. Finally, each forecast lead time (1–7 days) was calibrated separately.

To examine the performance of the BMA predictive model in both a deterministic forecast and a probabilistic

TABLE 1. Comparison of single-center ensemble prediction systems used in this study.

Center	Country/domain	Model spectral resolution	Ensemble members (perturbed)	Spatial resolution	Forecast length (in days)
CMA	China	T213L31	14	0.5625° × 0.5625°	10
ECMWF	Europe	T399L62/T255L62	50	1° × 1°	15
NCEP	United States	T126L28	20	1° × 1°	16
UKMO	United Kingdom		23	1° × 1°	15

forecast, the mean absolute error (MAE) was chosen here to measure the deterministic forecast skill. In this study, the deterministic forecasts of BMA models and raw ensemble forecasts are specified as their medians according to Sloughter et al. (2007). We also computed the MAE values for the deterministic forecasts based on the respective means, and reached a similar conclusion with Sloughter et al. (2007), that is, the median is a more "robust" choice for deterministic prediction in the case of a high skewed distribution, leading to smaller MAE than would be the case with the mean. The average width of the lower 90% prediction intervals, Brier score (BS; Wilks 2006, section 7.4.2) and continuous ranked probability score (CRPS) were chosen to measure probabilistic forecast. The MAE, which evaluates the accuracy of deterministic forecast, is specified as (Wilks 2006, section 7.3.2)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |f_i - o_i|, \tag{10}$$

where o_i is the value of observation i, f_i is the deterministic forecast value at the time and locations of observation i, and N is the total number of observations. The average width of the lower 90% prediction intervals, which assesses the sharpness of the probabilistic forecast, is defined as the average of the 90th percentile forecast at observation times. The BS, which is the mean square error of the probabilistic forecast and evaluates the accuracy of the probabilistic forecast at different thresholds (in a specific weather event), is defined as

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{N} (f_k - o_k)^2, \tag{11}$$

where the index k denotes a numbering of the N forecast–event pairs, f_k is the forecasting probability that was forecast, o_k is the actual outcome of the event at instance k, if the event occurs $o_k = 1$; otherwise, $o_k = 0$. The CRPS, which evaluates the accuracy of probabilistic forecast distribution, is specified as (Wilks 2006, section 7.5.1)

$$\mathrm{CRPS} = \frac{1}{N}\sum_{i=1}^{N} \int_{-\infty}^{+\infty} [F_i(x) - H(x - o_i)]\, dx, \tag{12}$$

where $F_i(x)$ is the forecast cumulative distribution function (CDF) at the observation i, $H(x - o_i)$ is a Heaviside function that jumps from 0 to 1 at the observed value, if $x < o_i$, $H(x - o_i) = 0$, otherwise, $H(x - o_i) = 1$. The four verification statistics are all negatively oriented, that is, worse forecasts receive higher values, where the BS can take values in the range $0 \le \mathrm{BS} \le 1$.

b. The length of the BMA training period

The question arose as to what length of sliding-window training period would be adequate to arrive at a reasonable estimate of the BMA parameters. The answer involved a compromise, since the length of training periods also varied for the various areas and datasets; thus the decision could not be made automatically. We chose MAE, the average width of lower 90% prediction intervals and CRPS to examine the performance of BMA predictive models using different lengths of training period, and to determine the length of the training period. This is because the BMA methods for the single-center EPS and EGE EPS we developed and used are different, as described in section 4a. We need to determine the length of the BMA training period for the single-center EPS and EGE EPS. The 1-day ensemble forecast data of 24-h accumulated precipitation in this subsection comes from the UKMO EPS and EGE EPS. The BMA model was retrained each day for each station throughout the verification period, using a training sample period of the N previous days, where N = 10, 15, 20, 25, 30, 35, 40, 45, or 50. The verification period lasted from 22 July to 31 August 2007. The mean of the verification metrics was taken for all stations in the study area and for each day in the verification period.

Table 2 shows the MAE, CRPS value, and average width of the lower 90% prediction intervals from the BMA forecast for different training period lengths and raw ensemble forecasts from the UKMO EPS. It is shown that all the BMA models for different training period lengths had better performances than the raw ensemble forecasts. In addition, the MAE of the BMA deterministic forecasts decreased as the number of training days increased up to 25 days, then little change was observed beyond 25 days. The CRPS also decreased as the training period increased from 10 to 25 rainfall days, then stabilized, and the average width of BMA lower 90% prediction intervals increased with the number of training days, becoming stable after 25 days. In summary, it seemed that there are some advantages in increasing the training period to 25 days, and that there is little improvement beyond that time.

Table 3 shows the MAE, CRPS value, and average width of the lower 90% prediction intervals from the BMA forecast for different training period lengths and raw ensemble forecasts from EGE EPS. We reach similar conclusions with Table 2, finding that there are some advantages in increasing the training period to 25 days, and that there is little improvement beyond that time.

TABLE 2. Mean verification metrics for BMA models of UKMO EPS for different training periods and raw ensemble (Ens) for accumulated precipitation (unit: mm).

	Length of BMA training period (days)								
Metric	10 BMA (Ens)	15 BMA (Ens)	20 BMA (Ens)	25 BMA (Ens)	30 BMA (Ens)	35 BMA (Ens)	40 BMA (Ens)	45 BMA (Ens)	50 BMA (Ens)
MAE	6.99 (8.46)	6.95 (8.46)	6.92 (8.46)	6.91 (8.46)	6.91 (8.46)	6.90 (8.46)	6.91 (8.46)	6.89 (8.46)	6.90 (8.46)
CRPS	5.78 (6.98)	5.74 (6.98)	5.71 (6.98)	5.70 (6.98)	5.70 (6.98)	5.70 (6.98)	5.70 (6.98)	5.68 (6.98)	5.68 (6.98)
Width of lower 90% prediction interval	19.95	20.76	22.28	24.36	24.43	24.77	24.88	24.87	24.89

TABLE 3. Mean verification metrics for BMA models of EGE EPS for different training periods and raw ensemble (Ens) for accumulated precipitation (unit: mm).

	Length of BMA training period (days)								
Metric	10 BMA (Ens)	15 BMA (Ens)	20 BMA (Ens)	25 BMA (Ens)	30 BMA (Ens)	35 BMA (Ens)	40 BMA (Ens)	45 BMA (Ens)	50 BMA (Ens)
MAE	6.99 (8.50)	6.95 (8.50)	6.93 (8.50)	6.91 (8.50)	6.91 (8.50)	6.90 (8.50)	6.91 (8.50)	6.90 (8.50)	6.89 (8.50)
CRPS	5.77 (6.98)	5.73 (6.98)	5.71 (6.98)	5.69 (6.98)	5.68 (6.98)	5.69 (6.98)	5.68 (6.98)	5.67 (6.98)	5.68 (6.98)
Width of lower 90% prediction interval	19.65	20.39	22.32	23.71	23.91	24.02	24.05	24.16	24.13

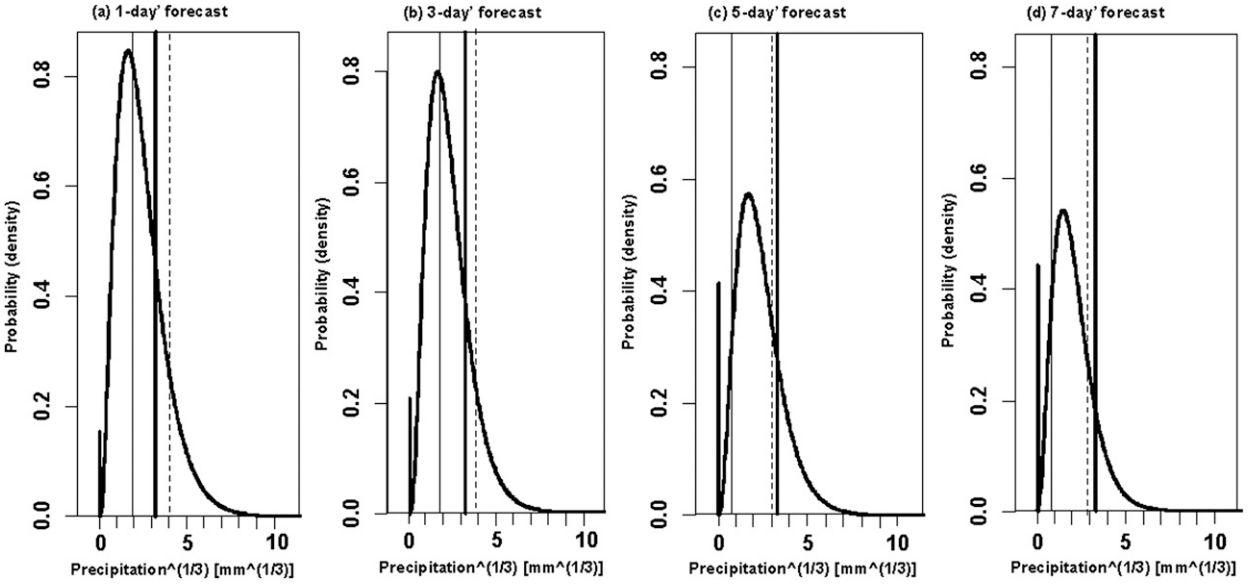

FIG. 2. BMA-fitted PDFs of 24-h accumulated precipitation from UKMO at station 57297 on 2 Aug 2007 with the lead time of (a) 1 day, (b) 3 days, (c) 5 days, and (d) 7 days. The thick vertical line at zero represents the BMA estimate of the probability of no precipitation, and the solid curve is the BMA PDF of the precipitation amount given that it is nonzero. The dashed vertical line represents the 90th percentile upper bound of the BMA PDF, the thin vertical line represents the deterministic forecast (median forecast), and the thick vertical solid line represents the verifying observation.

Accordingly, considering the number of calculations and comparison between UKMO BMA EPS and EGE BMA EPS, we chose 25 days as the length of the BMA training period for the single-center EPS and EGE EPS in this study area. This is different from the length of the BMA training period in Sloughter et al. (2007), because the length of the BMA training period is specific to different datasets and regions (Sloughter et al. 2007).

c. *BMA PQPFs for a heavy-rainfall event*

A heavy-rainfall event occurred on 2 August 2007 during the test period in the study area. The example in Fig. 2 illustrates how the BMA method works for PQPFs. It shows the observations, the deterministic forecast and probability of zero precipitation, along with the BMA predictive PDF, which were the averaged contributions from the bias-corrected ensemble members from the UKMO for station number 57297 with lead times of 1, 3, 5 and 7 days. Because the 23 UKMO ensemble members come from a single model, here the BMA weights were constrained to be equal and have a value of 1/23. The probability of exceeding a given amount is shown in Fig. 2 as the proportion of the area under the BMA PDF (top black curve) to the right, multiplied by the probability of the nonzero precipitation. The figure shows that the observation is far outside the deterministic forecast for a heavy-rainfall event, and the 90th percentile of forecasts is a better prediction than that of the deterministic forecasts. Furthermore, the BMA predictive PDF had a decreasing predictive skill as lead times increased, both for the deterministic forecast and the probabilistic forecast.

Figure 3 shows the percentile forecasts by the BMA predictive PDF of EGE for the 24-h accumulated precipitation at weather stations 57297 and 57083 on 2 August 2007. It is seen in the figure that the observations were located near the 90th percentile of forecasts at station 57297 (25–50 mm) and above the 95th percentile at station 57083 (>100 mm) for this particular heavy-precipitation event within the 1–7-day lead time. From Fig. 3, we can see that the predictive ability of deterministic forecasts (50th percentile forecasts) is somewhat lower than that for probabilistic forecasts for a heavy-precipitation event. Acknowledging that the outcome of a single case does not verify BMA PQPFs, the objective and detailed verification results are presented in the section 4e.

d. *BMA probabilistic quantitative precipitation forecasting*

This subsection describes the application of BMA to 1–7-day forecasts of 24-h accumulated precipitation for the summer in 2007 over the Huaihe Basin using the CMA EPS, ECMWF EPS, NCEP EPS, UKMO EPS, and EGE EPS in TIGGE, a simple comparison between the BMA model, logistic regression and raw ensemble

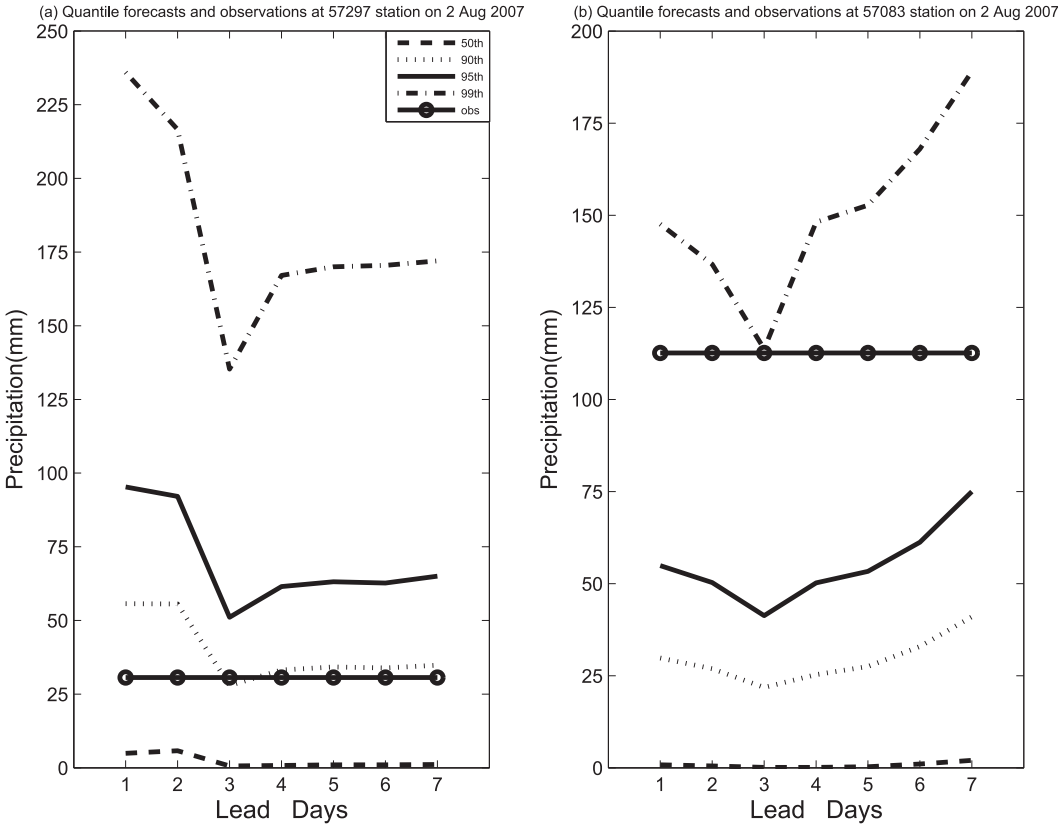

FIG. 3. Percentile forecasts and observed 24-h accumulated precipitation on 2 Aug 2007 from the EGE at (a) station 57297 and (b) station 57083.

forecasts, and the assessment of the five BMA forecasting systems for PQPFs. The verification period lasted from 22 July to 31 August 2007 and was the same as that used in section 4a. Four verification metrics—MAE, average width of lower 90% prediction intervals, BS, and CRPS—were employed to examine the performances of BMA predictive models. The means of these verification metrics were taken for all stations in the study area, and for each day of the verification period.

For 1-day forecasts, Table 4 shows the mean verification metrics of five BMA EPSs and raw ensemble forecasts. It is shown that all the BMA models for PQPFs performed better than raw ensemble forecasts in the five EPSs, the BMA model of the UKMO EPS performed best in four single-center EPSs, and the BMA model of EGE EPS performed best in all five BMA forecasting systems. The six-category precipitation forecasts obtained from the BMA predictive PDFs of the five EPSs were further assessed by the BS value (Fig. 4). For the single-center EPS, the CMA EPS performed better for rainfall in the region of 10 and 25 mm, but performed poorly for high-precipitation events

TABLE 4. Mean verification metrics for BMA models and raw ensemble (Ens) for different ensemble systems with lead time of 1 day for accumulated precipitation (unit: mm).

	Center				
Metric	NCEP BMA (Ens)	ECMWF BMA (Ens)	CMA BMA (Ens)	UKMO BMA (Ens)	EGE BMA (Ens)
MAE	6.96 (9.44)	6.95 (8.82)	6.93 (8.59)	6.91 (8.46)	6.91 (8.50)
CRPS	5.78 (8.59)	5.75 (7.98)	5.71 (7.09)	5.70 (6.98)	5.69 (6.98)
Width of lower 90% prediction interval	24.59	24.52	24.46	24.36	23.71

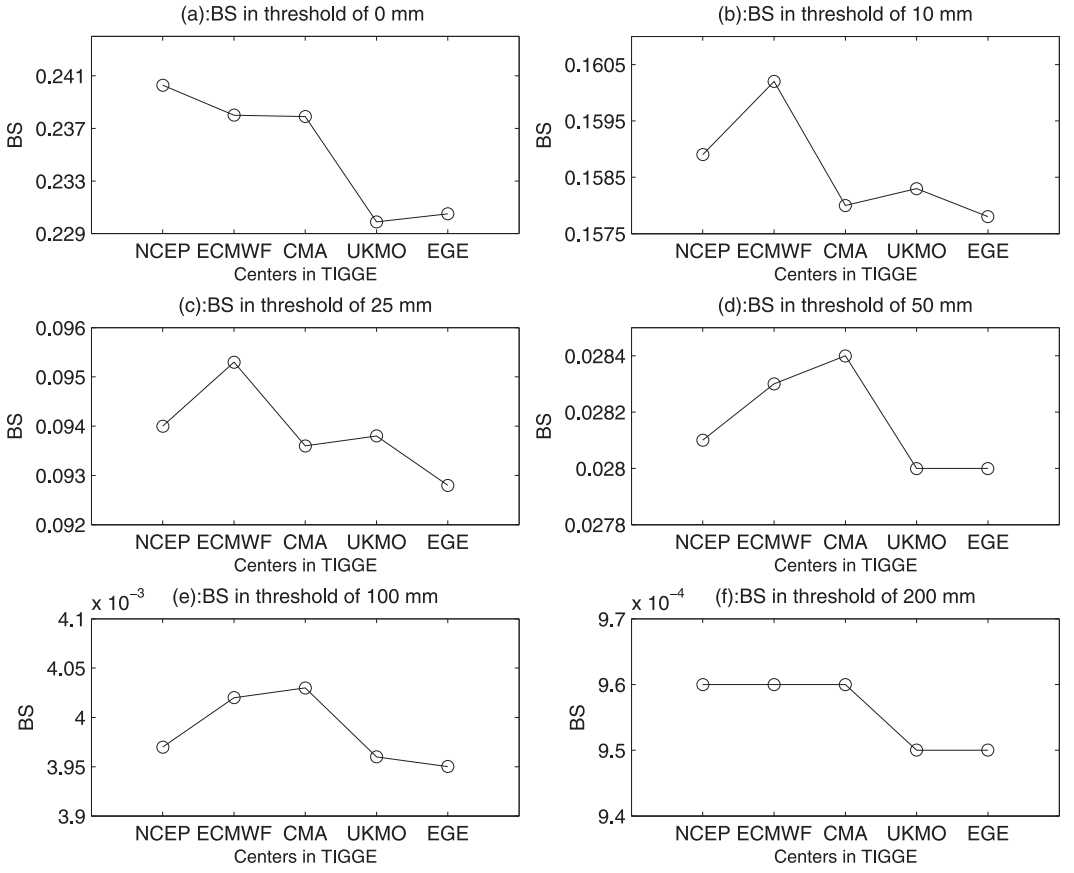

FIG. 4. Mean BS value of the six category precipitation forecasts obtained from the BMA of different EPSs with a lead time of 1 day: (a) 0, (b) 10, (c) 25, (d) 50, (e) 100, and (f) 200 mm.

(>25 mm); the UKMO EPS performed best in the other four categories of precipitation forecasts; and the BMA models of EGE had the best prediction skill (the lowest BS value) in all six-category precipitation forecasts.

Table 5 gives the BMA mean weights for EGE EPS. It is shown that the UKMO EPS had the highest weight and their ensemble members were the major contributing components of the BMA predictive PDF, which agreed with the previous conclusion in Table 4. BMA weights are not always indicative of forecast quality because of correlations between ensemble members (Raftery et al. 2005). In the EGE EPS, the four single-center EPSs come from four different models, so the BMA weights reflect the ensemble members' overall performance in the training period in this study.

Table 6 shows the mean BS value of probabilistic precipitation forecasts for multicenter EPS using different ensemble methods at various thresholds: sample climatology, ensemble consensus voting, logistic regression based on the cube root of the ensemble mean, and BMA. Here "climatology" refers to the empirical distribution of the verifying observations for all stations in the study area, the "logistic regression" is similar to the Eq. (2), and the ensemble consensus voting takes the probability of precipitation to be equal to the proportion of ensemble members that predicted precipitation. The raw ensemble forecasts had poor forecast skill especially at lower thresholds. The BMA model and logistic regression forecasts based on the cube root of the ensemble mean were superior to raw ensemble forecasts and comparable in BS (i.e., they showed comparable forecast skill for the specific weather event, but BMA has the advantage of estimating the complete PDF). A similar outcome was found for the 1–7-day forecasts (not shown in Table 6).

TABLE 5. Mean weights of BMA for EGE with lead time of 1 day.

Center (members)	Weight
UKMO (23)	$0.46 = 0.02 \times 23$
CMA (14)	$0.21 = 0.015 \times 14$
ECMWF (50)	$0.2 = 0.004 \times 50$
NCEP (20)	$0.13 = 0.0065 \times 20$

TABLE 6. Mean BS value for probabilistic precipitation forecasts using different ensemble methods with lead time of 1 day.

Score	Threshold (mm)	Climatology	Ensemble	Logistic	BMA
BS	0	0.2489	0.4179	0.2228	0.2229
BS	10	0.1595	0.1947	0.1512	0.1508
BS	25	0.0933	0.1013	0.0917	0.0922
BS	50	0.0274	0.0285	0.0272	0.0271
BS	100	0.0037	0.0038	0.0037	0.0037
BS	200	0.0009	0.0009	0.0009	0.0009

For the 1–7-day forecasts, the performances of the BMA PQPFs for the five EPSs are compared in Fig. 5, using the three verification metrics of MAE, CRPS, and average width of the lower 90% predictive interval. The mean verification metrics is an arithmetic mean for all stations without missing data over the Huaihe Basin (i.e., any missing data for a station were excluded on that day). In general, the BMA models of the UKMO and ECMWF EPSs were a little superior to the other two for single-center EPSs, and the BMA model of the EGE EPS demonstrated the best predictive skill. Notable exceptions were that the BMA model of the CMA EPS was superior to both the NCEP and ECMWF EPSs for lead times of 1 day, but showed the lowest predictive skill for lead times of 3–7 days.

e. *A heavy-rainfall warning scheme*

The BMA method generated calibrated and sharp predictive PDFs, providing a reliable description of the total predictive uncertainty, and extreme event information was obtained through an analysis of the PDF. In this subsection we propose a heavy-rainfall warning scheme based on a study of the BMA predictive PDF. The previous studies described in section 4d showed that the BMA model of the EGE EPS performed best in all EPSs. According to the BMA predictive PDF for EGE, forecasts appeared in all percentiles. Figure 6 shows that the deterministic forecasts predicted normal weather very well (Fig. 6a) but were almost completely unable to predict heavy rainfall (>25 mm, Fig. 6b). Only probabilistic predictions (such as 90th percentile forecasts in Fig. 6b) warned of extreme precipitation events and provided a basis for taking precautions against extreme precipitation weather. Figure 7, which illustrates the percentile forecasts and observations of the 24-h accumulated precipitation at station 57290 from 9 July to 8 August 2007, indicates that there were two heavy-rainfall events exceeding 50 mm. These observations

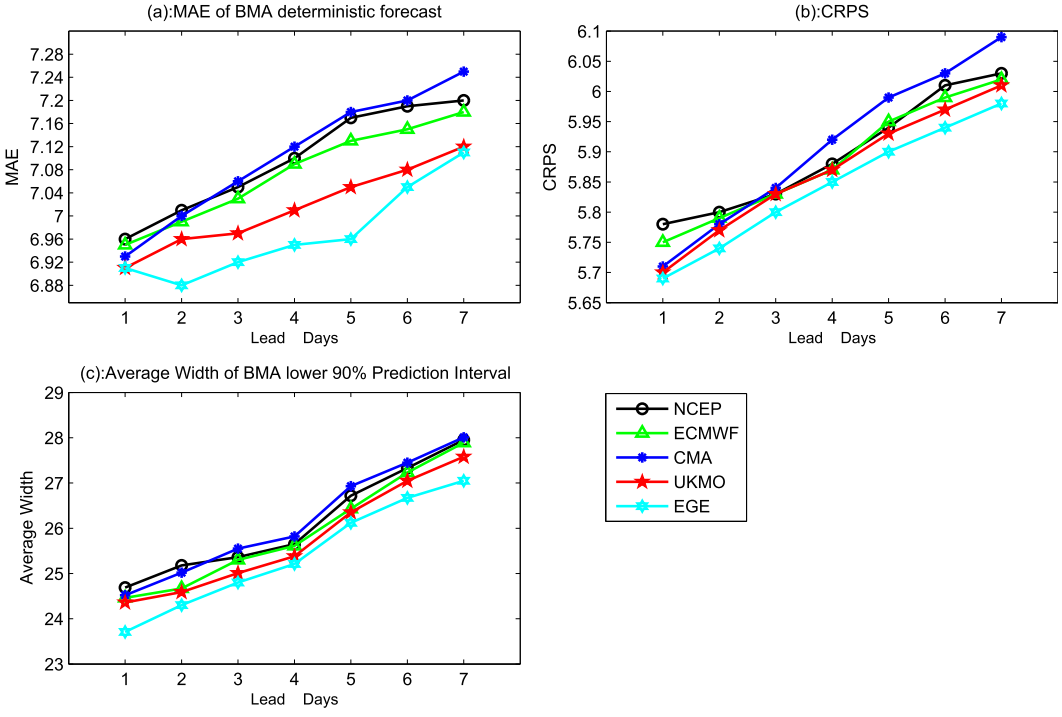

FIG. 5. Mean verification metrics for BMA models of 24-h accumulated precipitation with lead times of 1–7 days for different EPSs at all stations in the Huaihe Basin: (a) MAE of BMA deterministic forecasts, (b) CRPS, and (c) average width of lower 90% prediction intervals.

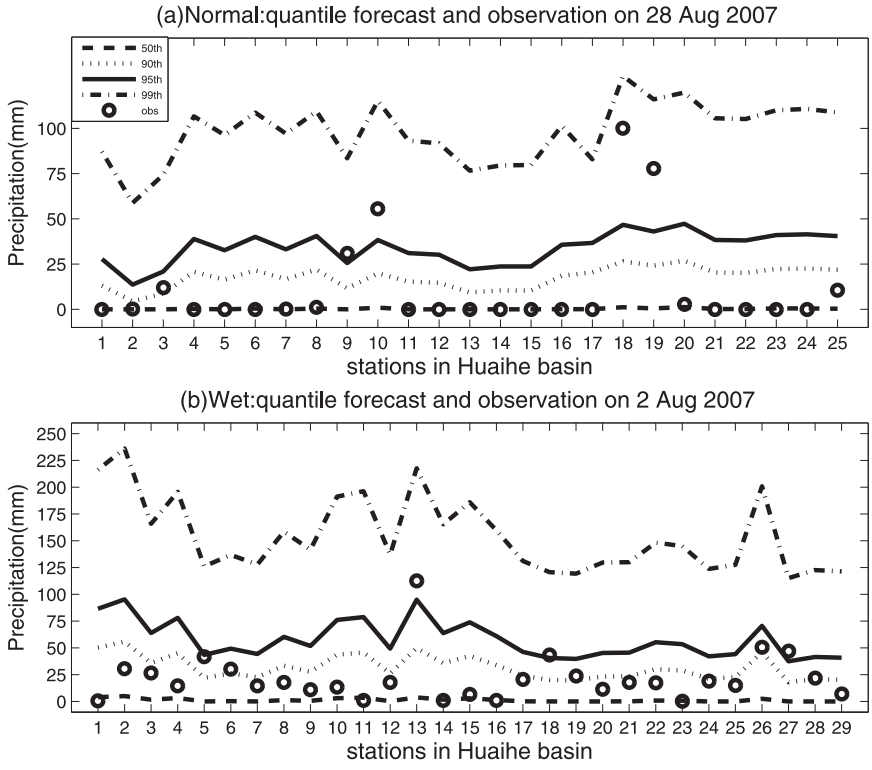

FIG. 6. Percentile forecasts and observed 24-h accumulated precipitation in the Huaihe Basin: (a) normal on 28 Aug 2007 and (b) wet on 2 Aug 2007.

were both located above the 90th percentile; all the 90th percentile forecasts on those days exceeded the heavy-rainfall event threshold (50 mm), and also were located above the 90th percentile forecasts of the previous few days. If similarly regular patterns were found in the historical ensemble forecasts and observations at this station, it would be possible to propose a summer heavy-rainfall warning scheme for the area in which this station

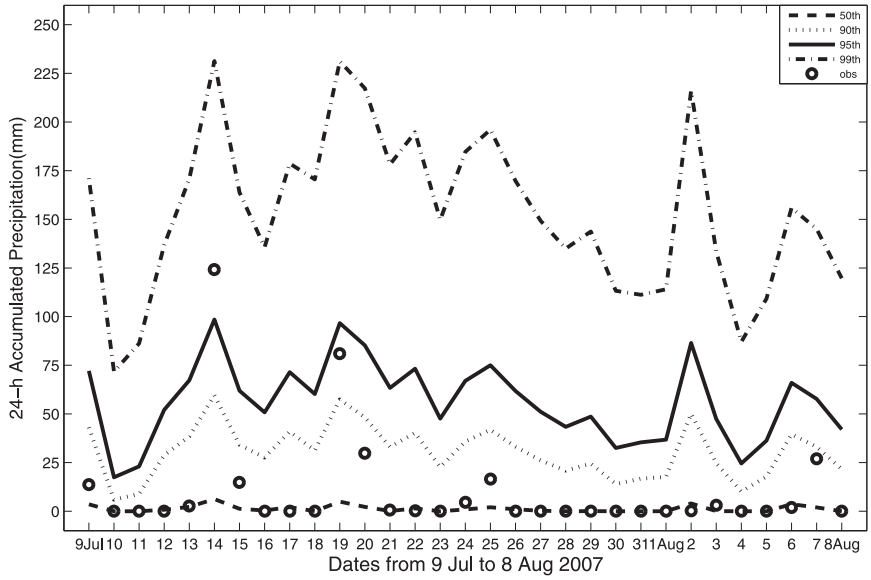

FIG. 7. Percentile forecasts and observed 24-h accumulated precipitation at station 57290 from 9 Jul to 8 Aug 2007.

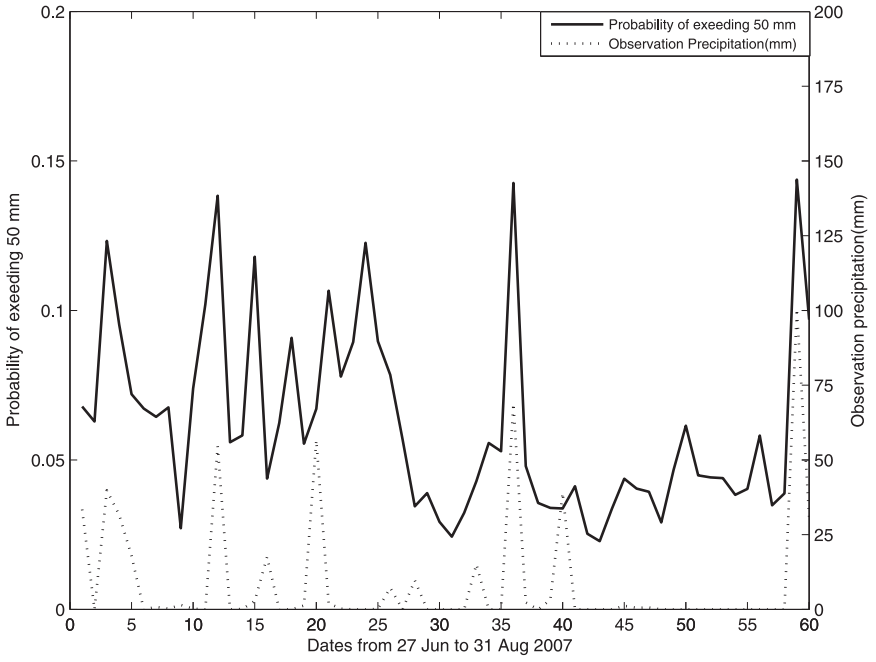

FIG. 8. Probability of exceeding 50 mm and observed 24-h accumulated precipitation at station 54909 from 27 Jun to 31 Aug 2007 (the days without missing data).

is located. The heavy-rainfall warning corresponds to the 90th percentile, so if the 90th percentile forecast for some day clearly lies above the heavy-rainfall event threshold (50 mm) and exceeds the 90th percentile forecasts for previous few days, we would trigger a heavy-rainfall warning in this area, and the decision-makers could then take precautions against the predicted extreme precipitation event. It is to be expected that there will be occasional incorrect and missed predictions in the percentile forecasts: in Fig. 7, the incorrect forecast on 2 August 2007, are examples. According to the above heavy-rainfall warning scheme, we should trigger a heavy-rainfall warning in this area on 2 August 2007, but the observations show that there was no rainfall at this station on 2 August 2007. Here, the percentile of the warning scheme is specific to different stations and times.

Figure 8 shows the variation of the probability of exceeding 50 mm and observation of 24-h accumulated precipitation from 27 June to 31 August 2007 at station 54909 (the days without missing data). Figure 9 shows the variation of the probability of exceeding 50 mm and observation of 24-h accumulated precipitation at all stations over the Huaihe Basin on 31 August 2007. The trends and patterns of BMA probabilistic forecasts were consistent with observations in most cases, particularly for the some of the heavy-rainfall events, which were also predicted well during the test period. These results suggest that a heavy-rainfall warning scheme based on BMA probabilistic forecasts is also feasible for a specific station, and can be extended to other stations through different percentiles.

5. Summary and discussion

In this study, we applied BMA methods to TIGGE single-center and multicenter EPSs, and obtained five BMA predictive systems for PQPFs. The performances of the five BMA predictive systems for PQPFs during the verification period over the study area were investigated using the 24-h accumulated precipitation data obtained from the TIGGE-ECMWF portal, and the observations of 43 rain gauges located in the study basin. It was shown that the BMA models for PQPFs performed better than raw ensemble forecasts for all five EPSs with lead times of 1–7 days. The BMA model and logistic regression forecasts based on the cube root of the ensemble mean showed comparable forecast skill for the specific weather event, but BMA has the advantage of estimating the complete PDF. In general, for single center EPSs, the UKMO, and ECMWF predictions were a little superior to the CMA and NCEP in the study area; however, the CMA showed good predictive skill for weak precipitation with a 1-day lead time, and showed poor predictive skill in other cases. The BMA model of the EGE performed better than those of the single-center EPSs with lead times of 1–7 days.

Based on the percentile forecasts from the BMA predictive PDFs for EGE, a heavy-rainfall warning scheme is

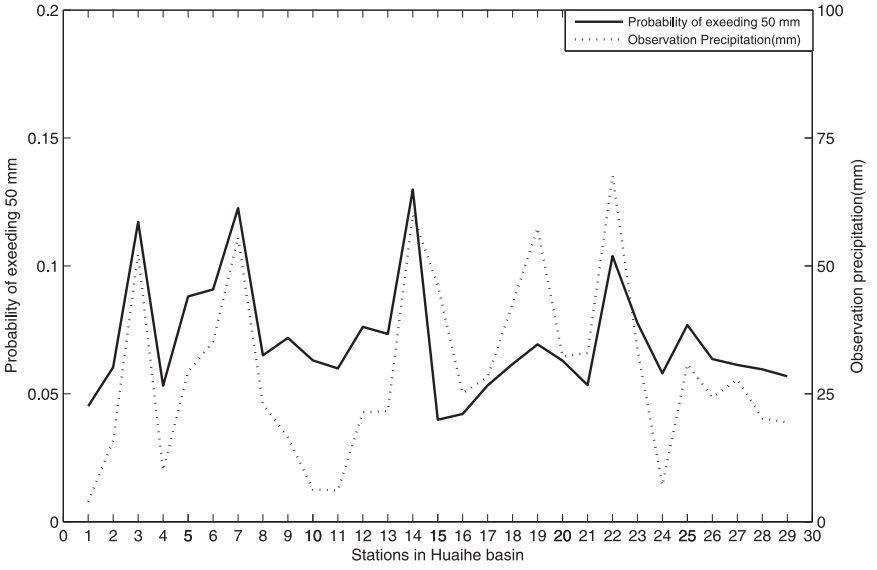

FIG. 9. Probability of exceeding 50 mm and observed 24-h accumulated precipitation at all stations in the Huaihe Basin on 31 Aug 2007.

proposed for the study area. The heavy-rainfall warning corresponds to a given percentile in the test area, if this percentile forecast for a particular day clearly lies above the heavy-rainfall event threshold and exceeds this percentile in forecasts for the previous few days, a heavy-rainfall warning will be issued in this area. The scheme has important implications for precautionary measures to be taken at times of heavy-precipitation weather and in early flood forecasting and warning.

As discussed above, the BMA PQPFs based on multicenter EGE EPSs with lead times of 1–7 days greatly improved the quality of PQPFs in most cases, demonstrating the potential of BMA methods and multicenter grand ensemble modeling. With the developing probabilistic forecasting methodology and the use of multimodel ensemble forecasting, extreme precipitation events forecasting and 3–10-day probability flood forecasting will become well developed, as pointed out by Thielen et al. (2009) and Zhao et al. (2011). Thus, an advanced statistical postprocessing approach and multicenter ensemble forecasting will greatly facilitate the development of the heavy-rainfall events forecasting and early probability flood forecasting. In addition, in this study, the predicted precipitation of each EPS was interpolated to the stations by the bilinear interpolation method, and the effect of topography was not considered, so the advanced interpolation methods may further improve the quality of PQPFs. Finally, Schmeits and Kok (2010) and Tian et al. (2011) indicated that the methods of bias correction and of estimating the parameters in BMA usually affect the predictive skill of the model, so more work needs to be done on comparing the BMA model performance using a range of bias correction methods and parameter estimation methods.

Acknowledgments. This research was supported by the National Basic Research Program of China (Grant 2010CB428403), the National Natural Science Foundation of China (Grants 91125016 and 41075062), and the Special Funds for Public Welfare of China (Grant GYHY201306045). The authors gratefully acknowledge the Department of Statistics of the University of Washington for making the ensemble BMA software available online, the TIGGE–ECMWF portal for providing the dataset of ECMWF EPS, NCEP EPS, UKMO EPS, CMA EPS in TIGGE, and CMA for providing precipitation observations. We thank the editor, Josh Hacker, and the two anonymous reviewers for constructive comments and suggestions, which helped to improve the paper.

REFERENCES

Barnston, A. G., S. J. Mason, L. Goddard, D. G. DeWitt, and S. E. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc.,* **84,** 1783–1796.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. A,* **39B,** 1–39.

Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian, 2007: Multimodel ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.,* **30,** 1371–1386.

Fisher, R. A., 1922: On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London,* **222A,** 309–368.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecasting ensembles with exchangeable and missing

members using Bayesian model averaging. *Mon. Wea. Rev.,* **138,** 190–202.

Fritsch, J. M., and Coauthors, 1998: Quantitative precipitation forecasting: Report of the eighth prospectus development team, U.S. Weather Research Program. *Bull. Amer. Meteor. Soc.,* **79,** 285–299.

Gneiting, T., and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science,* **310,** 248–249.

Goswami, M., K. M. O'Connor, and K. P. Bhattarai, 2007: Development of regionalisation procedures using a multi-model approach for flow simulation in an ungauged catchment. *J. Hydrol.,* **333** (2–4), 517–531.

Gourley, J., and B. E. Vieux, 2005: A method for evaluating the accuracy of quantitative precipitation estimates from a hydrologic modeling perspective. *J. Hydrometeor.,* **6,** 115–132.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting,* **17,** 192–205.

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

Hoeting, J. A., D. M. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial (with discussion). *Stat. Sci.,* **14,** 382–401.

Kass, R. E., and A. E. Raftery, 1995: Bayes factors. *J. Amer. Stat. Assoc.,* **90,** 773–795.

Leamer, E. E., 1978: *Specification Searches.* Wiley, 370 pp.

Lee, J. A., W. C. Kolczynski, T. C. McCandless, and S. E. Haupt, 2012: An objective methodology for configuring and down-selecting an NWP ensemble. *Mon. Wea. Rev.,* **140,** 2270–2286.

Liu, J. G., Z. H. Xie, L. N. Zhao, and B. H. Jia, 2013: BMA probabilistic forecasting for the 24-h TIGGE multi-model ensemble forecasts of surface air temperature (in Chinese). *Chin. J. Atmos. Sci.,* **37** (1), 43–53.

Matsueda, M., and H. L. Tanaka, 2008: Can MCGE outperform the ECMWF ensemble? *SOLA,* **4,** 77–80.

McLachlan, G. J., and T. Krishnan, 1997: *The EM Algorithm and Extensions.* Wiley, 274 pp.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECWMF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

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

Palmer, T. N., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Park, Y. Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.,* **134,** 2029–2050.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Richardson, D., 2005: The THORPEX Interactive Grand Global Ensemble (TIGGE). *Geophys. Res. Abstr.*, **7,** Abstract EGU05-A-02815.

Schmeits, M. J., and K. J. Kok, 2010: A comparison between raw ensemble output, (modified) Bayesian model averaging and extended logistic regression using ECMWF ensemble prediction reforecast. *Mon. Wea. Rev.,* **138,** 4199–4211.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.

——, T. Gneiting, and A. E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.,* **105,** 25–35.

Thielen, J., J. Bartholmes, M. H. Ramos, and A. de Roo, 2009: The European Flood Alert System. Part 1: Concept and development. *Hydrol. Earth Syst. Sci.,* **13,** 125–140.

Tian, X. J., Z. H. Xie, A. H. Wang, and X. C. Yang, 2011: A new approach for Bayesian model averaging. *Sci. China Earth Sci.,* **54,** 1–9.

Vrugt, J. A., C. G. H. Diks, and M. P. Clark, 2008: Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. *Environ. Fluid Mech.,* **8** (5–6), 579–595.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 627 pp.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1364–1385.

Yang, C., Z. W. Yan, and Y. H. Shao, 2009: Probabilistic quantitative precipitation forecasts for TIGGE ensemble forecasts (in Chinese). Water resources response and sustainable utilization under changing environment. *Extended Astracts, 2009 Annual Conf. of the Water Resources Professional Committee of Chinese Irrigation Works Society,* Dalian, China, Water Resources Professional Committee of Chinese Irrigation Works Society, 117 pp.

Zhao, L. N., F. Y. Tian, H. Wu, D. Qi, J.-Y. Di, and Z. Wang, 2011: Verification and comparison of probabilistic precipitation forecasts using the TIGGE data in the upriver of Huaihe Basin. *Adv. Geosci.,* **29,** 95–102.

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability. *Adv. Atmos. Sci.,* **22** (6), 781–788.