Journal of Hydrology 519 (2014) 2890–2905

Contents lists available at ScienceDirect

Journal of Hydrology

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

An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai river basin

Yumeng Tao [a,b], Qingyun Duan [a,*], Aizhong Ye [a], Wei Gong [a], Zhenhua Di [a], Mu Xiao [a,c], Kuolin Hsu [b]

[a] State Key Laboratory of Earth Surface Processes and Resource Ecology, College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China
[b] Department of Civil and Environmental Engineering, University of California, Irvine, CA 92697-2175, United States
[c] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195-2700, United States

ARTICLE INFO

Article history:
Available online 24 April 2014

Keywords:
TIGGE
Ensemble forecast
Ensemble post-process
Precipitation
Huai river basin
Ensemble verification

SUMMARY

This paper evaluates how post-processing can enhance raw precipitation forecasts made by different numerical weather prediction (NWP) models archived in TIGGE (THORPEX Interactive Grand Global Ensemble) database. Ensemble Pre-Processor (EPP), developed at U.S. National Weather Service, is used to post-process raw precipitation forecasts. EPP involves several major steps: (1) deriving the joint distributions of raw forecasts and observations corresponding to different canonical events; (2) obtaining the probability distributions of observations given the raw forecasts; and (3) constructing ensemble forecasts from the conditional probability distributions given the raw forecasts. Raw precipitation forecasts from five NWP models (CMA, ECMWF, JMA, NCEP and UKMO) during the summer-fall period (rainy season) from 2007 to 2011 were evaluated over the Huai river basin in China. The lead time for the precipitation forecasts is set to 9 days, which are divided into 11 canonical events (defined as daily precipitation events or aggregate precipitation events over a period of several consecutive days). Our experiments show that post-processed precipitation forecasts shows substantial improvement over the raw forecasts. Post-processing reduces both the biases and the root mean squared error of the raw forecasts significantly. In terms of ensemble spread, both the Brier skill scores and continuous ranked probability skill score are improved appreciably after post-processing. Reliability diagrams and rank histograms also confirm that post-processed ensemble forecasts possess better ensemble spread property compared to the raw forecasts. Among the five NWP models, ECMWF and JMA have the best overall performance in both raw and post-processed forecasts. The raw and post-processed UKMO and NCEP forecasts outperform other models in certain events. Post-processing can improve the CMA raw forecasts substantially, but still its performance is consistently worse than that of the other models.

1. Introduction

Currently many numerical weather prediction (NWP) centers in the world are providing ensemble weather forecasts. Ensemble forecasts are generated by running the NWP models in forecast mode with perturbed initial conditions and/or physical parameterization schemes. Compared to single-value deterministic weather forecasts, ensemble forecasts provide not only the forecasts of the most likely events, but also the uncertainty information (Park et al., 2008). They have the added advantage of extended forecast lead times by considering the uncertainty in initial conditions and in model physics (Toth and Kalnay, 1997; Buizza et al., 2005; Duan et al., 2012). Realizing the potential benefits of ensemble forecasts, the World Meteorological Organization (WMO) sponsored the Observing System Research and Predictability Experiment (THORPEX) to further improve the ensemble forecasts of severe meteorological events with 1 day to 2-weeks lead times (Shapiro and Thorpe, 2004). As part of THORPEX program, the THORPEX Interactive Grand Global Ensemble (TIGGE) database was established to help evaluate and improve the accuracy of short and medium range of ensemble forecasts by collecting the ensemble forecasts generated by different NWP centers since the end of 2006 (Park et al., 2008; Bougeault et al., 2010).

Based on some promising results from Hydrological Ensemble Prediction Experiment (HEPEX), ensemble forecasts data collected by TIGGE database can potentially benefit the decision making of hydrologists and water resources managers (Thielen et al., 2008; He et al., 2010). However, to use the ensemble forecasts effectively, post-processing is necessary for multiple reasons: (1) although ensemble meteorological forecasts have improved significantly, the accuracy of the raw ensemble meteorological forecasts are still

* Corresponding author. Tel.: +86 010 5880 4191.
E-mail address: qyduan@bnu.edu.cn (Q. Duan).

http://dx.doi.org/10.1016/j.jhydrol.2014.04.040
0022-1694/

limited and not suitable for direct applications to hydrological forecasting; (2) the spread of the raw meteorological ensembles may be unreliable as the uncertainty range derived from the ensemble spread may not contain statistically consistent number of observations; and (3) the spatial scale of meteorological ensemble forecasts is incompatible with those required for generating hydrological forecasts, as hydrological models are usually run over catchments, while meteorological forecasts are normally over grids (Rayner et al., 2005).

One important way to resolve the aforementioned problems is to calibrate the raw weather forecasts using statistical methods, rather than to use dynamical models to downscale them (Wood and Schaake, 2008). Therefore, statistical methods are widely used for post-processing and interpreting raw meteorological ensemble forecasts (Glahn et al., 2009). Numerous approaches were established and developed, including superensemble techniques (Krishnamurti et al., 2000; Yun et al., 2005), "Ensemble dressing" methods (Roulston and Smith, 2003), Bayesian model averaging (Raftery et al., 2005), and analog techniques (Hamill et al., 2006). As an integral part of the U.S. National Weather Service River Forecast System (NWSRFS), Ensemble Pre-Processor (EPP) is designed to prepare ensemble meteorological forecasts as inputs to streamflow prediction models (Wu et al., 2011). The methodology has been used in multiple river forecast centers (RFCs) for years and demonstrated to improve precision and effectiveness of ensemble weather forecasts (Wood and Schaake, 2008; Voisin et al., 2010; Wu et al., 2011; Liu et al., 2013). It is functioned to adjust raw weather forecasts to construct ensemble forecasts from a retrospective verification period by correlating forecasts and observed quantities and applying the "Schaake Shuffle" method. The "Schaake Shuffle" method is used to reconstruct space–time variability of precipitation and temperature by reordering ensemble forecasts based on historical events (Clark et al., 2004).

To apply EPP to forecasts data from TIGGE database, one major issue is lack of long historical archive of weather forecasts. As Hamill et al. (2008) indicated, it is important to use a large sample size of hindcast dataset to include unusual and rare events. However, as mentioned above, TIGGE database contains data only after 2007 with multiple model forecasts available. Therefore, 5 years precipitation forecasts data of five NWP models collected by TIGGE database are applied to study the effectiveness of EPP to TIGGE forecasts.

The objective of this paper is to evaluate the skill of TIGGE ensemble precipitation forecasts over Huai river basin, China. The paper covers: (1) verification of potential skills of precipitation forecasts; (2) demonstration of the effectiveness of EPP to prepare inputs for hydrological models from raw precipitation forecasts with limited data length; and (3) analysis and comparison of the relative skills of five NWP models over different sub-basins and different canonical events. The five NWP models evaluated in the study are CMA, ECMWF, JMA, NCEP and UKMO. The raw forecast data were downloaded from TIGGE database and were post-processed using EPP version 3.

The paper is organized as follows. Section 2 reviews EPP methodologies. Section 3 describes the setups of the experiments, including description of study basin, data used, experimental design, and verification methods selected. Comparisons between raw models and post-processed models are presented in Section 4. Finally, the main conclusions are summarized in Section 5.

2. Methodology

Ensemble Pre-Processor (EPP) is an integral part of the NWSRFS developed by the Hydrology Laboratory of the NWS in the U.S. and has been used experimentally in different RFCs to post-process the raw quantitative precipitation forecasts (QPFs) and quantitative temperature forecasts (QTFs) needed for hydrological river forecasting (Schaake et al., 2007; Wu et al., 2011). EPP is basically a statistical model that relates single value QPFs (or QTFs) to corresponding observations. For EPP to work properly, it requires a long historical hindcast database of QPFs/QTFs and corresponding observations. The method involves: (1) establishing joint probability distributions of forecasts and observations, (2) generating probabilistic predictions of the observations conditioned on given forecasts, and (3) constructing ensemble members based on the conditional probabilistic predictions. A brief description of the EPP methodology for post-processing QPFs is presented as follows.

2.1. Canonical events

The joint probability distributions of QPFs and observations are built on the basis of canonical events, which correspond to precipitation events with specific lead times and durations. For example, a canonical event can be the next-day precipitation total, or the average daily precipitation from day 6 to day 10 in the forecast period The purpose of designing canonical events is to fully extract information in the raw forecasts. Due to the chaotic nature of the weather and climate system, any individual precipitation forecast beyond 5 days is deemed not reliable (Lorenz, 1963). But this does not mean that there is no skill for precipitation forecasts beyond 5 days. With ensemble forecasts that account for uncertainty in initial conditions and/or model physics, the aggregate (or average) precipitation for a future period beyond 5 days (say day 6–day 10) may still contain meaningful skill. To make use of this skill, we can construct a joint probability distribution of the aggregate precipitation forecasts during this period (i.e., for the canonical event) and the corresponding observations and derive the conditional probability of precipitation given the raw forecast. When we construct the canonical events, for short-term forecasts (say with a lead time between day 1 and day 4), the sub-daily or daily forecasts should have relatively high skill. Accordingly, the canonical events can be consisted of sub-daily (e.g., 6-hourly) or daily events. On the other hand, when the lead time is beyond 4 days, the skill of the forecasted precipitation for any particular sub-daily or daily period may be not meaningful. In this case, the canonical events should be consisted of aggregate precipitation over certain periods (e.g., day 5–7 or day 5–10). A canonical event can also be constructed based on user needs. In practice, the forecast users may be only interested in the total amount of precipitation during a period. For example, the water manager of a large reservoir may be only interested in the total reservoir inflow during the next 2 weeks, which is strongly related to the total precipitation during this period.

2.2. Conditional probability distribution of canonical events

The construction of conditional probability distribution given the forecasts of a specific canonical event involves setting up the data pool, constructing the marginal distributions of the forecasts and observations, obtaining the joint distribution, and then generating conditional distribution given the forecasts.

Let X denote the raw single-value precipitation forecasts of a given canonical event for a specific forecast date (say July 1st), and Y the corresponding observations. If the length of the hindcast database used to calibrate the statistical parameters is 10 years, then $\{X, Y\}$ should contain 10 data pairs. To increase the sample size, a time window is chosen so the forecasts of the canonical event on the days before and after the specific forecast date from the hindcast database are also included in the data pool. The forecasts (X) and observations (Y) in the data pools can be then expressed as: $X = (X_{ij})_{y\times d}$, $Y = (Y_{ij})_{y\times d}$, where y denotes the number

Table 1
Illustration of the "Schaake Shuffle" procedure for single-time step canonical events (SCEs).

Observation matrix			Sampled ensemble members for SCEs in ascending order			Observation matrix for SCEs with ranks in parentheses			Shuffled ensemble members with the same ranks as the observations		
Year	Time Step 1	Time Step 2	Ens. #	Time Step 1	Time Step 2	Year	Time Step 1	Time Step 2	Ens. #	Time Step 1	Time Step 2
2004	9.2	5.8	1	1.1	9.7	2004	9.2(4)	5.8(4)	1	4.9(4)	12.2(4)
2005	10.2	4.4	2	2.8	10.8	2005	10.2(5)	4.4(3)	2	5.1(5)	11.1(3)
2006	3.7	1.0	3	3.2	11.1	2006	3.7(1)	1.0(1)	3	1.1(1)	9.7(1)
2007	5.5	12.3	4	4.9	12.2	2007	5.5(2)	12.3(7)	4	2.8(2)	14.0(7)
2008	15.0	8.2	5	5.1	12.5	2008	15.0(7)	8.2(6)	5	10.2(7)	13.3(6)
2009	6.9	7.7	6	9.5	13.3	2009	6.9(3)	7.7(5)	6	3.2(3)	12.5(5)
2010	11.0	2.3	7	10.2	14.0	2010	11.0(6)	2.3(2)	7	9.5(6)	10.8(2)
Step 1			Step 2			Step 3			Step 4		

Table 2
Illustration of the "Schaake Shuffle" procedure for composite-time step canonical events (CCEs).

Observation matrix			Sampled ensemble members for CCEs in ascending order			Observation matrix for CCEs with ranks in parentheses			Shuffled ensemble members for CCEs with the same ranks as the observations			Shuffled ensemble members for CCEs redistributed into individual time steps according to observed ratios		
Year	Time Step 1	Time Step 2	Ens. #	Time Step 1	Time Step 2	Year	Time Step 1	Time Step 2	Ens. #	Time Step 1	Time Step 2	Ens. #	Time Step 1	Time Step 2
2004	9.2	5.8	1	9.7		2004	(9.2 + 5.8)/2 = 7.5(5)		1	12.5(5)		1	12.5 * 9.2/7.5 = 15.33	12.5 * 5.8/7.5 = 9.67
2005	10.2	4.4	2	10.8		2005	(10.2 + 4.4)/2 = 7.3(4)		2	12.2(4)		2	12.2 * 10.2/7.3 = 17.05	12.2 * 4.4/7.3 = 7.35
2006	3.7	1.0	3	11.1		2006	(3.7 + 1.0)/2 = 2.4(1)		3	9.7(1)		3	9.7 * 3.7/2.35 = 15.27	9.7 * 1.0/2.35 = 4.13
2007	5.5	12.3	4	12.2		2007	(5.5 + 12.3)/2 = 8.9(6)		4	13.3(6)		4	13.3 * 5.5/8.98 = 8.15	13.3 * 12.3/8.98 = 18.38
2008	15.0	8.2	5	12.5		2008	(15.0 + 8.2)/2 = 11.6(7)		5	14.0(7)		5	14.0 * 15.0/11.6 = 18.10	14.0 * 8.2/11.6 = 9.90
2009	6.9	7.7	6	13.3		2009	(6.9 + 7.7)/2 = 7.3(3)		6	11.1(3)		6	11.1 * 6.9/7.3 = 10.49	11.1 * 7.7/7.3 = 11.71
2010	11.0	2.3	7	14.0		2010	(11.0 + 2.3)/2 = 6.7(2)		7	10.8(2)		7	10.8 * 11.0/6.65 = 17.86	10.8 * 2.3/6.65 = 3.74
Step 1			Step 2			Step 3			Step 4			Step 5		

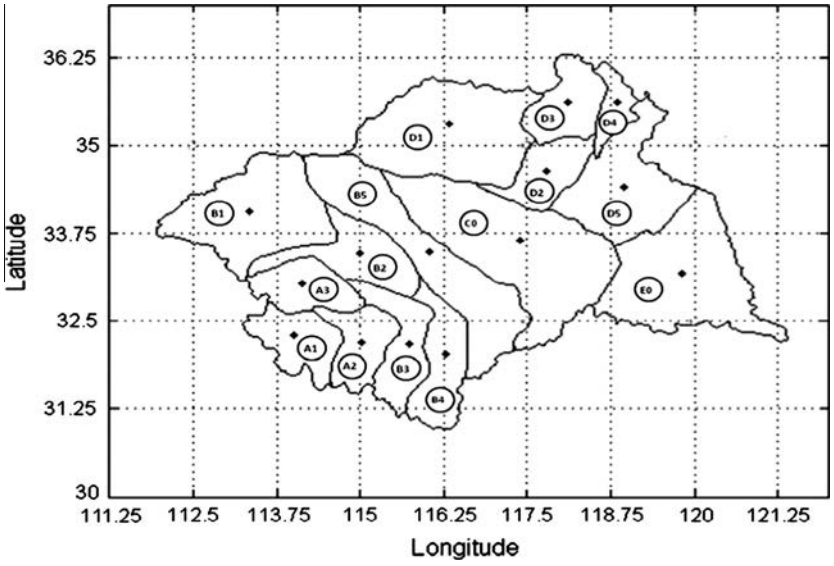

Fig. 1. Illustration of the catchments in Huaihe river basin. The black dots are the centers of the catchments.

of years in the hindcast database; d denotes the number of days in the time window.

With the data pool established, we can then derive the marginal distributions of X and Y and fit them to some forms of distribution functions. The marginal distribution of X can be expressed as:

$$F_X(x) = (1 - P_x) + P_x \times F_{X|X>0}(x|x > 0) \quad (1)$$

where $F_X(x)$ denotes the fitted cumulative distribution function (CDF) of the forecasts; P_x denotes forecasted probability of precipitation (PoP), which can be estimated by the ratio between the

Table 3
The information of Huai river basin's sub-basins.

ID	Catchment name	Centre (longitude, latitude)	Area (10^3 km^2)	Annual mean precipitation (mm)
A1	Dapoling upstream of huaihe to Xixian catchment	114.01°E, 32.31°N	16.5	1063.96
A2	Xixian upstream of huaihe to Wangjiaba catchment	115.02°E, 32.21°N	8.8	1009.00
A3	Ruhe and upstream of Honghe catchment	114.12°E, 33.04°N	9.5	904.64
B1	Upstream of Yinghe to Zhoukou catchment	113.33°E, 34.07°N	27.4	687.60
B2	Midstream of Yinghe and Zhoukou to Fuyang catchment	114.99°E, 33.47°N	14.3	824.94
B3	Shihe catchment	115.73°E, 32.19°N	10.6	1130.30
B4	Pihe, downstream of Huaihe and Huaigan catchment	116.29°E, 32.03°N	11.4	1103.45
B5	Wohe, midstream of Huaihe and Huaigan catchment	116.04°E, 33.48°N	28.7	781.07
C0	Bangbu to Hungtse, midstream and downstream of Huaigan and Huihe catchment	117.4°E, 33.65°N	42.3	859.87
D1	Nansihu catchment	116.32°E, 35.31°N	30.8	634.28
D2	Zaozhuang and Xuzhou catchment	117.78°E, 34.63°N	9.2	781.11
D3	Upstream of Yihe catchment	118.12°E, 35.62°N	10.1	719.30
D4	Upstream of Shuhe catchment	118.84°E, 35.61°N	4.4	717.89
D5	Downstream of Yihe and Shuhe catchment	118.95°E, 34.41°N	26.9	864.71
E0	Hungtse to downstream of huaihe catchment	119.82°E, 33.18°N	30.6	947.35

Table 4
Basic description of five models used in the study.

Models	Sources	Forecast length (days)	No. ensembles	Starting date of TIGGE archive
CMA	China Meteorological Administration	10	15	15th May, 2007
ECMWF	European Centre for Medium–range Weather Forecasts	15	51	1st October, 2006
JMA	Japan Meteorological Agency	9	51	1st October, 2006
NCEP	National Centers for Environmental Prediction	16	21	5th May, 2007
UKMO	UK Met Office	15	24	1st October, 2006

numbers of the rainy events and the total events; $F_{X|X>0}(x|x>0)$ denotes the CDF of forecasted rainfall amount given that the rain has occurred. When estimating P_x, a threshold is introduced such that the hourly forecasted precipitation amount below this threshold is treated as zero. The functional form of the CDF can either be Gamma, Weibull, or Exponential distribution function to account for potential extreme precipitation forecast. The use of the extreme value distribution functions allows the EPP to account for extreme events outside the data range. In this study, Gamma distribution is chosen to represent the CDF of rainfall amount, i.e., $X|X>0 \sim Gamma(\alpha_x, \beta_x)$, where α_x, β_x denotes the Gamma parameters to be estimated based on the hindcast data. In EPP, α_x, β_x are estimated by the sample mean and variance according to the Method of Moments (MoM). Altogether, there are three parameters, $X \sim (\widehat{P_x}, \widehat{\alpha_x}, \widehat{\beta_x})$, to be estimated in order to calculate $F_X(x)$, where $\widehat{P_x}$, $\widehat{\alpha_x}$ and $\widehat{\beta_x}$ denote the estimates of P_x, α_x and β_x, respectively.

The functional form of marginal distribution for Y is similar to that of X:

$$F_Y(y) = (1 - P_y) + P_y \times F_{Y|Y>0}(y|y>0) \quad (2)$$

where P_y denotes observed PoP; $F_{Y|Y>0}(y|y>0)$ denotes the CDF of observed rainfall amount given that the rain has occurred. The parameters of $F_Y(y)$, $Y \sim (\widehat{P_y}, \widehat{\alpha_y}, \widehat{\beta_y})$, are determined also by the MoM.

To obtain the joint distribution of X and Y, $F_{X,Y}(x, y)$, we need to transform X and Y into the normal space because obtaining joint distribution of two non-Gaussian variables is very difficult. In EPP, the Normal Quantile Transform (NQT) method (Kelly and Krzysztofowicz, 1997; Krzysztofowicz, 1997) is used to map X and Y into standard normal random variables U and V, respectively.

The NQT procedure works as follows. Below we used X as an example. The same procedure works for Y. Let

$$u = t(x) = \Phi^{-1}(F_X(x)) \quad (3)$$

where $u \sim N(0, 1)$ is standard normal random variable; t denotes NQT; $\Phi^{-1}(\cdot)$ denotes the inverse of CDF of the standard normal distribution. Note that the marginal distribution is not strictly monotonic when non-rainy events exist. In this case, the corresponding u to $x = 0$ in the standard normal space is an interval $(-\infty, u_0]$, where

$$u_0 = t(x = 0) = \Phi^{-1}(F_X(x = 0)) = \Phi^{-1}(1 - P_x) \quad (4)$$

In normal space, the joint distribution $F_{U,V}(u, v)$ can be constructed by assuming it as a bivariate normal distribution:

$$(U, V) \sim N\left(\vec{0}_2, \Sigma_{2\times 2}\right)$$

where covariance matrix Σ is expressed as $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, in which ρ denotes the correlation coefficient between U and V. ρ represents the correlation between the forecasted precipitation and corresponding observation. EPP offers four options on how ρ can be estimated by the sample Pearson correlation coefficient: (1) between raw forecasts X and observations Y for all events including dry and wet events, (2) between only raw forecasts X and observations Y for only the rainy days, (3) between the transformed variables U and V for all events, and (4) between the transformed variables U

Table 5
Canonical events designed to extract information from the raw ensemble forecasts.

Canonical events	Lead time combinations
1	Day 1
2	Day 2
3	Day 3
4	Day 4
5	1–2 Days
6	1–3 Days
7	1–4 Days
8	1–9 Days
9	5–6 Days
10	5–9 Days
11	7–9 Days

Table 6
Description of common verification measures used in the study.[a]

Verification measures	Formulas	Descriptions	Perfect/ no skill	Remarks
Bias	$Bias = \frac{\bar{x}-\bar{y}}{(\bar{x}+\bar{y})/2}$	Differences between forecasts and observations over long time period	0/n/a	
Root mean square error skill score (RMSE-SS)	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2}$; $RMSE\text{-}SS = \left(1-\frac{RMSE}{RMSE_{ref}}\right)$	Closeness of forecast and observation over a long time period	1/⩽0	$RMSE_{ref}$: reference
Pearson correlation coefficient	$r = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}}$	Linear dependency between forecasts and observations	1/⩽0	
Brier skill score (BSS)	$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i-o_i)^2$; $BSS = 1-\frac{BS}{BS_{ref}}$	Accuracy of PoP comparing to climatology	1/⩽0	$BS_{ref} = \bar{o}(1-\bar{o})$
Continuous ranked probability skill score (CRPSS)	$CRPS = \int_{-\infty}^{\infty}(F_{x_i}(t)-F_{y_i}(t))^2 dt$; $CRPSS = 1-\frac{CRPS}{CRPS_{ref}}$	Integrated squared error between the cumulative distribution function (CDF) of the forecasts and the CDF of the observations	1/⩽0	$F_{x_i}(t)/F_{y_i}(t)$: CDF of forecasts/observations
Reliability diagram (RD)	$f_i = \frac{1}{N_k}\sum_{i\in I_k}F_{y_i}(t)$; $p_{f_i}(x=1\|f_i) = \frac{1}{N_k}\sum_{i\in I_k}p(t\geqslant x_i)$	Closeness between forecast probability and observed frequency	Diagonal/ no skill	t: threshold; $F_{y_i}(t)$: PoP; N_k: # in the kth bin; I_k: index for kth bin
Rank histogram	$E[P(e_{i-1}\leqslant o\leqslant e_i)] = \frac{1}{n+1}$; $r_j = \overline{P(e_{i-1}\leqslant o\leqslant e_i)}$	Errors in ensemble forecasts' mean and spread relative to observations	Even/n/a	

[a] Denote x_i, y_i, p_i and o_i as the forecast, corresponding observation, probability of precipitation and observed frequency respectively at time i, while N is the amount of pairs of forecast and observation. Similarly, denote $\bar{x}$, $\bar{y}$, $\bar{o}$, are denoted as forecast average, observation average, and average of observed frequency.

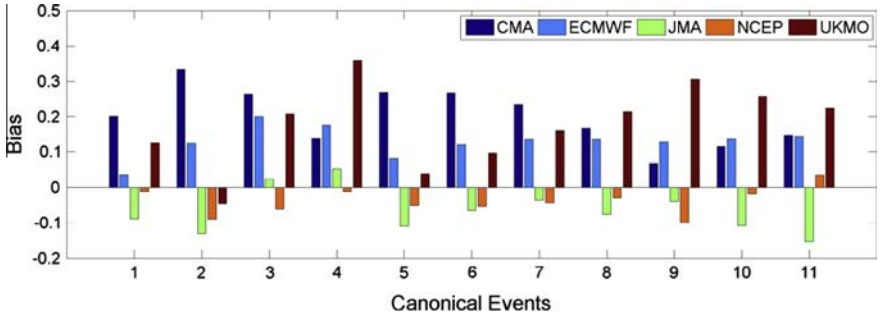

Fig. 2. Bias of the raw mean ensemble forecasts relative to the corresponding observations of different canonical events for five models. The different colors of bars indicate different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

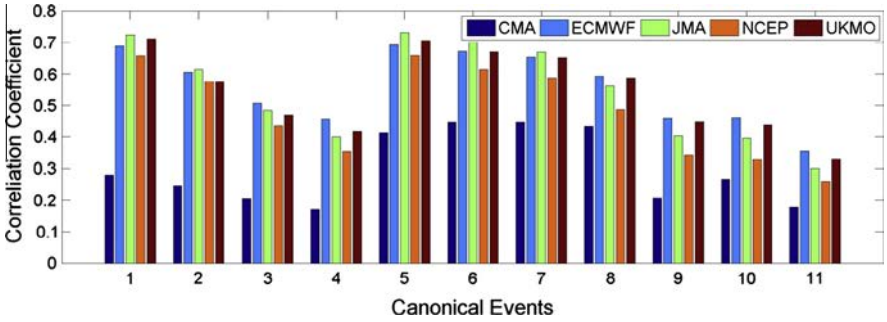

Fig. 3. Correlation coefficients between the post-processed mean ensemble forecasts and the corresponding observations of different canonical events for five models. The different colors of bars indicate different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and V for only the rainy events. In this study, the sample Pearson correlation coefficient between all raw forecasts and observations is used. Robertson et al. (2013) offered a more nuanced approach for determining ρ, which can better deal with the transformation and zero value problems in raw precipitation forecasts. Future version of EPP may incorporate their approach.

After the NQT procedure is carried out, the joint distribution of (X, Y), $F_{X,Y}(x,y)$, is now substituted by the bivariate normal distribution of (U,V), $F_{U,V}(u,v)$, the exact form of which is determined by parameters P_x, α_x, β_x, P_y, α_y, β_y and ρ.

Once $F_{U,V}(u,v)$ is determined, the conditional distribution of observations given the raw single-value precipitation forecast, $F_{Y|X=x}(y|x)$, is:

$$F_{Y|X=x}(y|x) = F_{V|U=u}(v|u) \tag{5}$$

where $u = t(x)$. Based on the property of the bivariate normal distribution, the conditional distribution of V given $U = u$ can be obtained as

$$V|U = u \sim N(\rho^{-1}u, 1-\rho^2)$$

If $X = 0$, the conditional distribution is computed as:

$$F_{Y|X=0}(y|x=0) = F_{V|U\leqslant u_0}(v|u\leqslant u_0) = \frac{F_{U,V}(u_0,v)}{F_U(u_0)} \tag{6}$$

where $u_0 = t(x=0) = \Phi^{-1}(F_X(x=0))$; $(U,V) \sim N(\vec{0}_2, \Sigma_{2\times 2})$; $U \sim N(0,1)$ and $F_{U,V}(\cdot,\cdot)$ and $F_U(\cdot)$ denote CDF of (U, V) and U respectively.

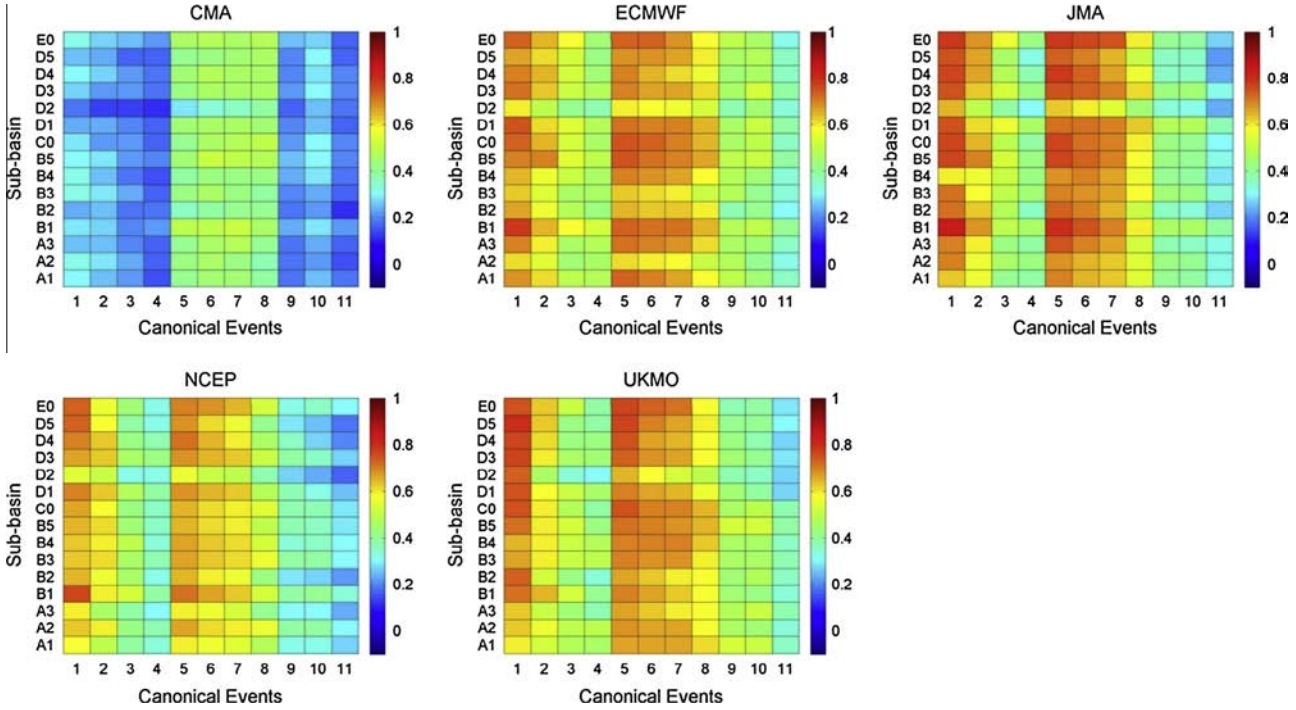

Fig. 4. Correlation coefficients between the post-processed mean ensemble forecasts and the corresponding observations of different canonical events and 15 sub-basins for five models.

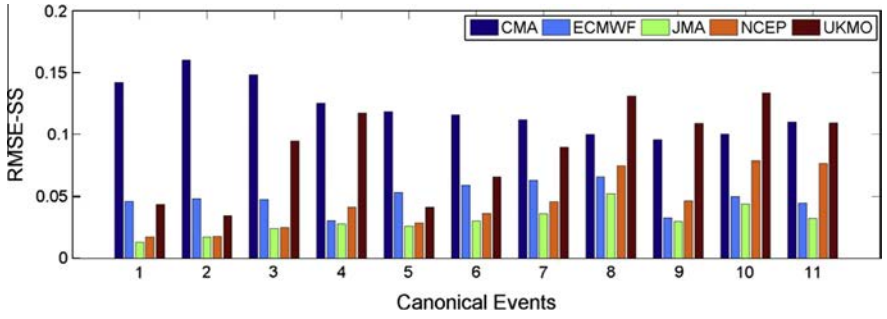

Fig. 5. Root mean square error skill score (RMSE-SS) of the post-processed mean ensemble forecasts relative to the raw mean ensemble forecasts of different canonical events for five models. The reference of the skill score is the raw mean ensemble forecasts. The different colors of bars indicate different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

After $F_{Y|X=x}(y|x) = F_{V|U=u}(v|u)$ is solved based on the hindcast data and corresponding observations, the conditional probabilistic forecasts given the raw forecasts can be approximated by a sequence of stratified samples of random variable V given U. The corresponding conditional probabilistic forecasts in the original space can be approximated by remapping the values from the Normal space to the original space by using the inverse of NQT.

2.3. Ensemble construction

Section 2.2 describes how to generate the conditional probabilistic forecasts given the raw forecasts, which are approximated by an ensemble of random samples from the conditional distribution for a given canonical event. The ensemble members for all canonical events need to be processed together to create ensemble forecasts in the time series format needed for generating hydrological forecasts. There is a key requirement for any individual ensemble member: each time series represented by the ensemble member must have space–time statistical properties consistent with that of historical observations. To construct ensemble forecasts with consistent space–time statistical properties of historical observations, the "Schaake Shuffle" methodology for canonical events is used in EPP.

The "Schaake Shuffle" procedure was first described in Clark et al. (2004) for single-time step events. Here a description of the "Schaake Shuffle" procedure based on canonical events is presented. Two illustrations are given in Tables 1 and 2, involving single-time step canonical events (SCEs, the same as in Clark et al., 2004) and the composite time step canonical events (CCEs), respectively.

For a specific lead time on a given forecast date, a historical observation matrix corresponding to the same time period of the QPF time series is constructed as:

$$O = (O_{ij})_{\#years\times m} = (\vec{O}_1, \ldots, \vec{O}_m)_{\#years\times m}$$

where $\#years = \#ens$, $\#years$ denotes number of years in historical data used to construct the ensembles, and $\#ens$ is the number of ensemble members. Each $\vec{O}_j$, $j = 1,\ldots,m$, is a vector of historical observation data for jth time step.

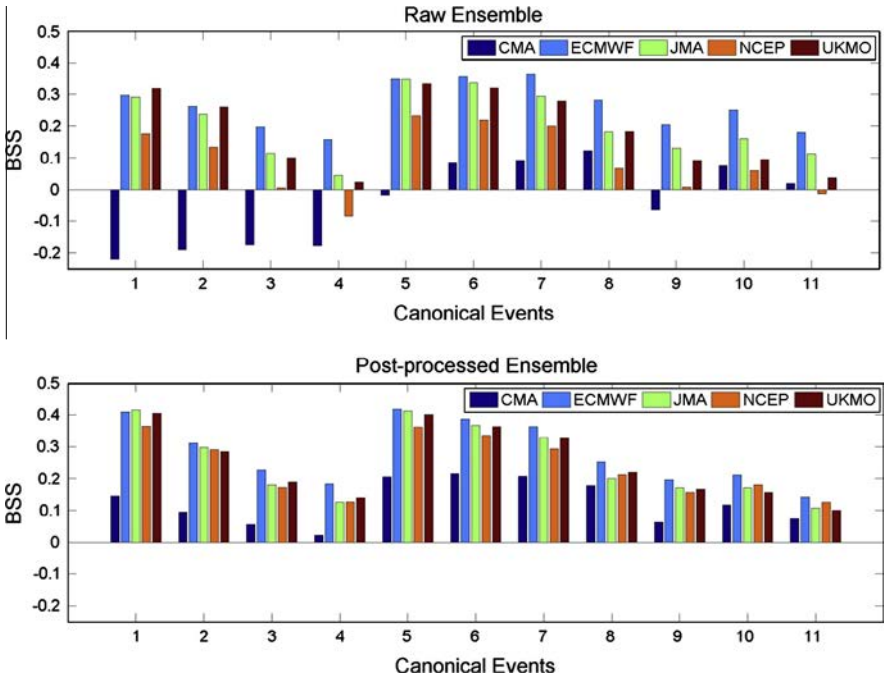

Fig. 6. Brier skill score (BSS) of the raw and post-processed ensemble forecasts of different canonical events for five models. The different colors of bars indicate different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

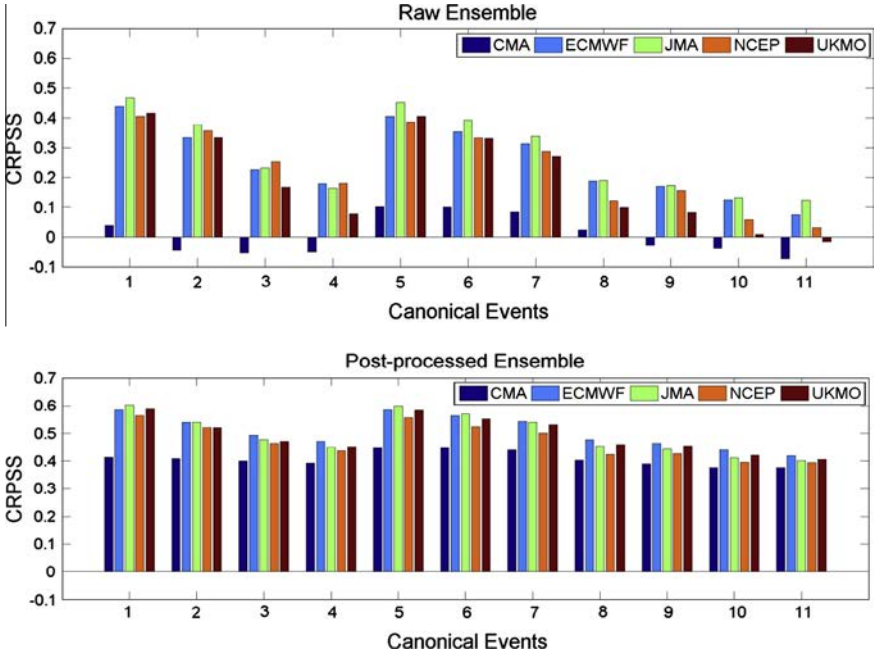

Fig. 7. Continuous ranked probability skill score (CRPSS) of the raw and post-processed ensemble forecasts of different canonical events for five models. The different colors of bars indicate different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In Table 1, the "Schaake Shuffle" procedure for SCEs is illustrated. In Step 1, the observation matrix is obtained. In Step 2, the ensemble members, which are sampled *#year* times from the conditional probability distributions for a specific canonical event, are shown in an ascending order. In Step 3, the observation matrix is marked with the ranks from the smallest (1) to the largest (7). In Step 4, the ensemble members in Step 2 are rearranged according to the observed ranks shown in Step 3, thus completing the "Schaake Shuffle". Now we have generated the ensemble members, with ensemble member #1 having the values of {4.9, 12.2} and ensemble member #2 having the values of {5.1, 11.1} and so on. If you examine the ranks of those ensemble members, each of them corresponds to the ranks of the observations in a historical year. Meanwhile, for each time step, the collection of ensemble members forms an empirical probability distribution that is the same as the conditional probability distribution of the corresponding canonical event.

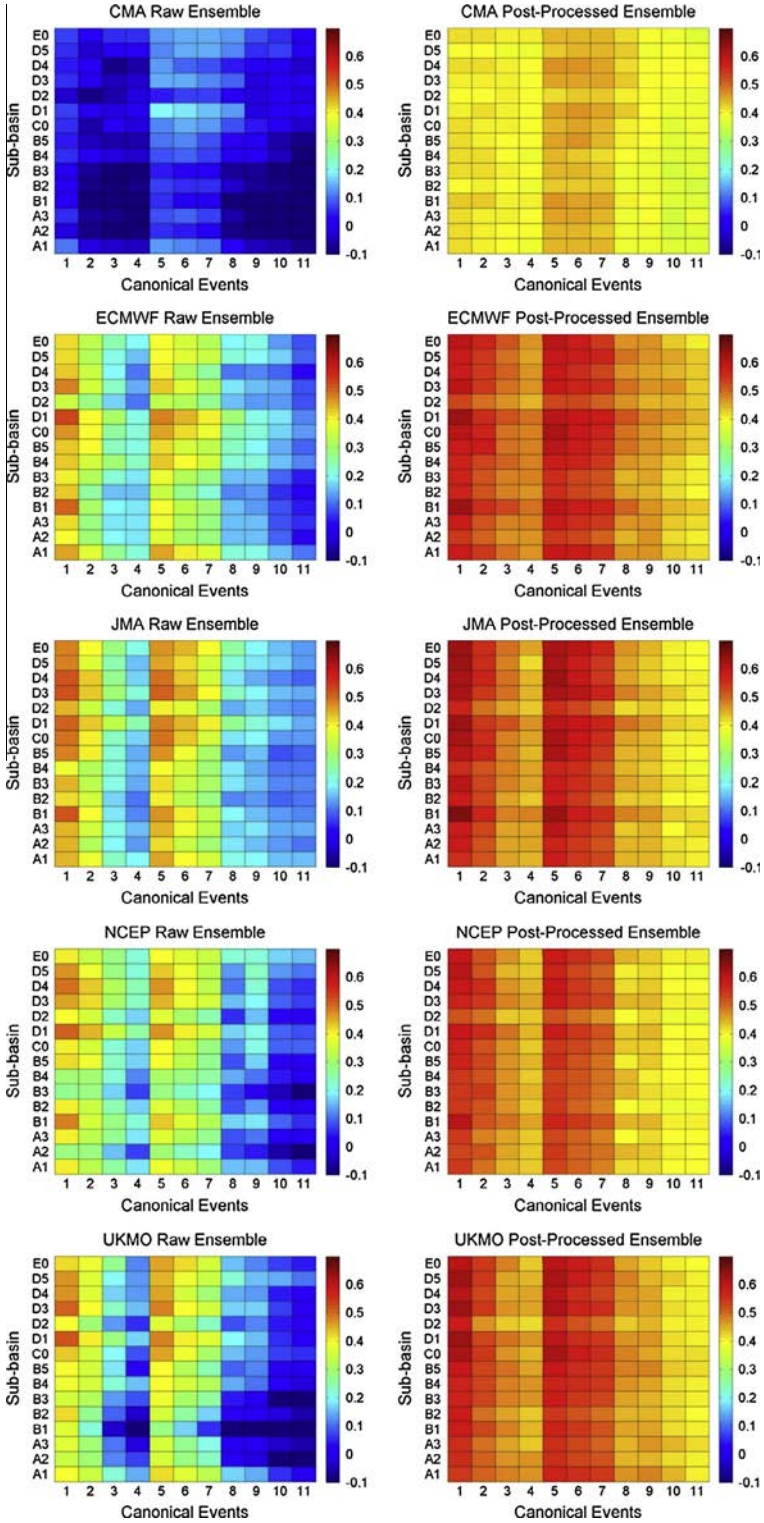

Fig. 8. Continuous ranked probability skill score (CRPSS) of the raw and post-processed ensemble forecasts of different canonical events and 15 sub-basins for five models.

For a CCE, the observation matrix is again shown in Step 1 of Table 2. In Step 2, the ensemble members in ascending order are sampled #*year* times from the conditional probability distributions for the CCE, which is the average precipitation forecasts for time steps 1 and 2. In Step 3, the observation matrix in Step 1 is used to obtain average observations corresponding to the CCE, which are the averages of observations for time steps 1 and 2. These average values and their ranks are recorded in Step 3 of Table 2. In Step 4, the ensemble members shown in Step 2 are shuffled so their ranks are the same as the observed ranks. In Step 5, the ensemble members for the CCE are repartitioned into individual time steps according to the ratios of the corresponding observations.

In practice, there are many canonical events (which can be SCEs or CCEs) that are included in post-processing process. In performing the "Schaake Shuffle" procedure, we start with the canonical event which has the lowest skill, which usually corresponds to

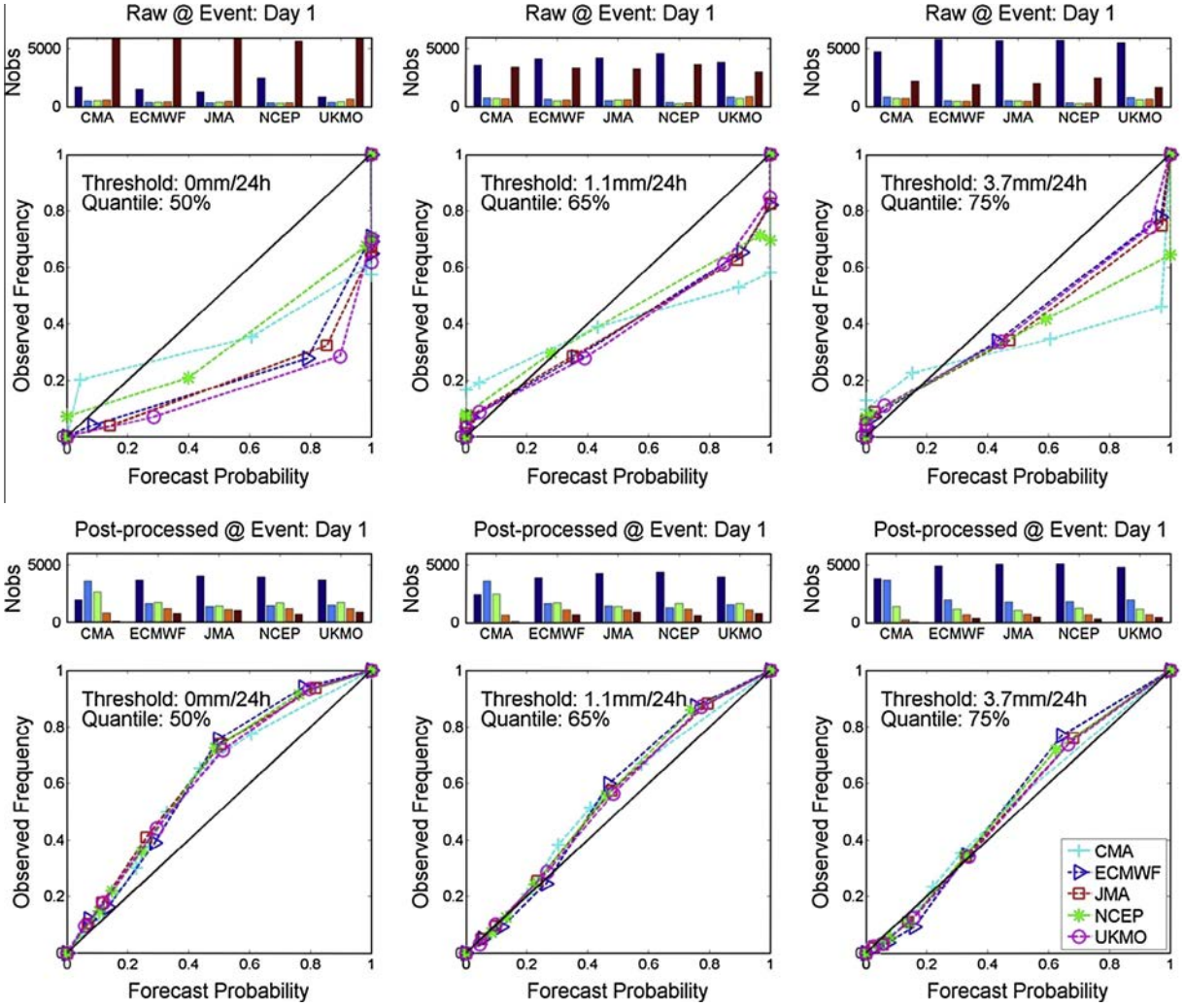

Fig. 9. Reliability diagram (RD) of the raw and post-processed ensemble forecasts of canonical event 1 (forecast period: day 1) for five models. The different colors of lines denote different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the canonical event that has the longest lead time and work sequentially to the event with the highest skill score, which usually correspond to the event with the shortest lead time.

3. Study basin, data description and experimental design

3.1. Study basin

The Huai river basin (30°55′–36°36′N, 111°55′–121°25′E), the 7th largest river system in China, is chosen as the study basin. Huai river basin is located between the Yellow river and the Yangtze river and runs from the west to the east (Zhang et al., 2011). With an approximate total drainage area of 270,000 km^2, Huai river basin involves five provinces with a dense of population of 185 million people and contributes over 17% of the national agriculture production (Liu et al., 2013). The Huai river basin is a natural boundary that separates China into warm temperate zone to the north and sub-tropical humid climate zone to the south. The basin has an annual mean temperature between 11–16 °C and an annual mean precipitation of approximately 900 mm. The precipitation shows a significant time and space variation. Rainy season (June–September) precipitation accounts for 50–80% of the total annual precipitation. The amount of precipitation decreases from the south to the north, the mountainous region to the plain region, and the coast to inland. According to historical data, both floods and droughts occur frequently in Huai river basin. In the most recent 50 years, 9 major floods and 12 major droughts ravaged the basin, causing severe losses of lives and properties. Better precipitation forecasting is critical to the flood forecasting and water resources management in the Huai river basin (Huai River Commission, 2009).

In this study, the Huai river basin is divided into 15 sub-basins to be consistent with operational meteorological forecast units of Huai river basin Meteorological Forecasting Center. Fig. 1 provides the boundaries of sub-basins while the geographic information and average annual precipitation is listed in Table 3 (Liu et al., 2013).

3.2. Data used

Data used in this study are the raw ensemble precipitation forecasts of five NWP models from the TIGGE database and the observed mean areal precipitation (MAP) data. The five NWP models are CMA, ECMWF, JMA, NCEP and UKMO. The raw precipitation forecasts are at a spatial resolution of 0.5° × 0.5°, which are spatially interpolated from the original NWP outputs by TIGGE database automatically. The study period is the rainy season (1st May–31st October) from 2007 to 2011 of the Huai river basin. Some other key information of the models, including model

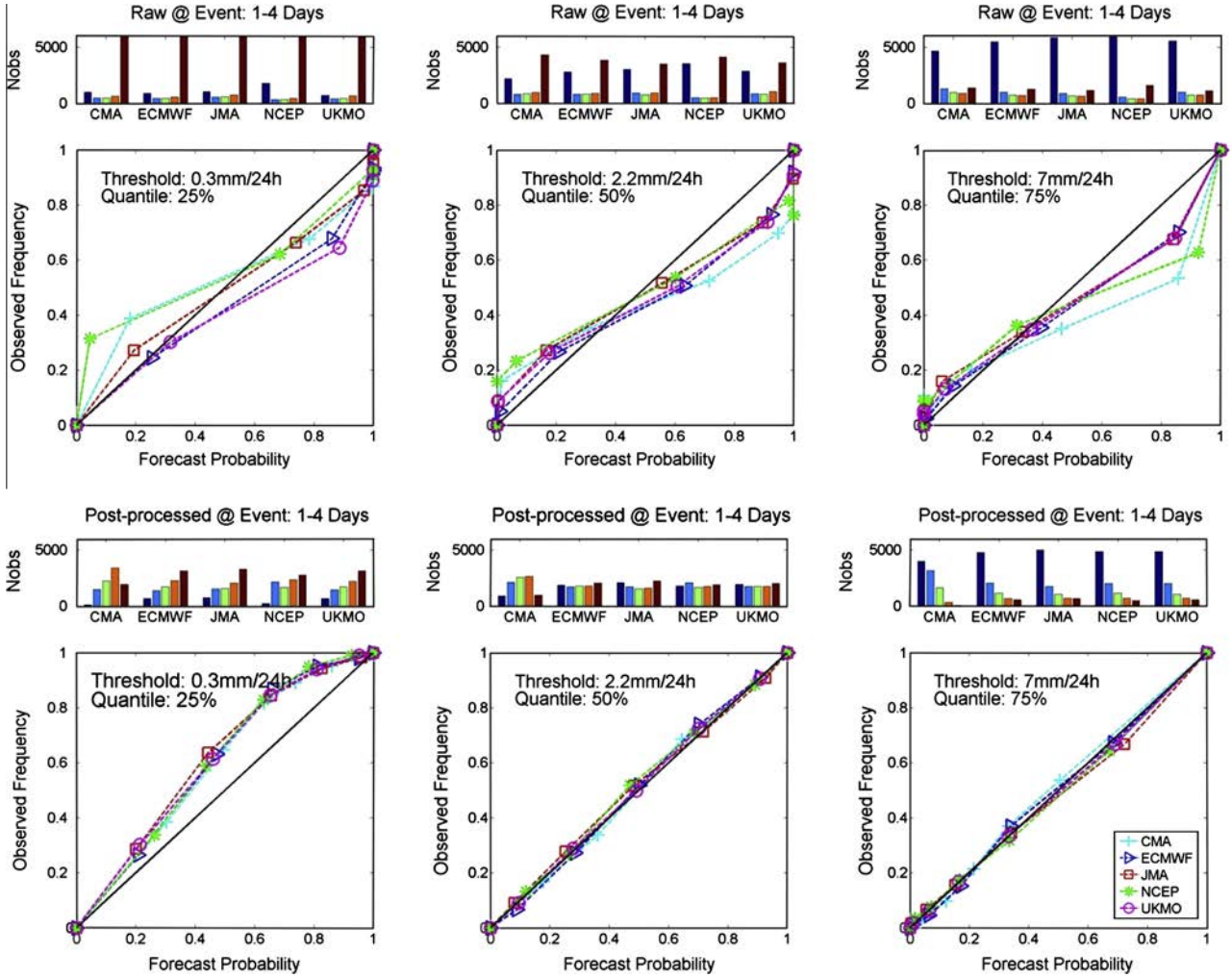

Fig. 10. Reliability diagram (RD) of the raw and post-processed ensemble forecasts of canonical event 7 (forecast period: 1–4 days) for five models. The different colors of lines denote different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sources, forecast length, number of ensembles and initial date of TIGGE operational model, as shown in Table 4. The ensemble mean is used as the input to EPP to derive joint distributions between forecasts and observations for different canonical events. The rationale is that mean ensemble forecasts have better skill than any individual ensemble member (Atger, 1999). Further the spread of raw ensemble forecasts is still not skillful enough to be included in deriving the joint probability between forecasts and observations (Hamill and Colucci, 1998).

The daily MAP data for 15 sub-basins of the Huai river basin are computed from observation data of 160 stations using the Thiessen Polygon method. The initial station data are obtained from the CMA data center. The observation data covers the data period from 1st January, 1981 to 30th November, 2011 to ensure an adequate amount of historical data for implementing the "Schaake Shuffle" procedure.

3.3. Experimental design

The MAP data of each sub-basin and the ensemble mean of the forecasts at the grid point nearest to the sub-basin center are used to estimate parameters and produce ensemble members. When generating the data pool for each canonical event, a fixed 61-days time window is used ($d = 61$), as this choice can guarantee approximately 30 wet days in the data pool for a 5-years hindcast database used in this study ($y = 5$). To compare the skill of the five models, the forecasts are processed separately with the same design of canonical events with lead time up to 9 days, which is the longest lead time for JMA. Eleven canonical events are designed to extract as much information from the raw forecasts as possible (see Table 5). The parameters of the joint distribution of each event are estimated every 5 days in the study period, from 1st May to 31st October. Then the parameters are linearly interpolated for the other days that are not computed. The calibration period is from 1st May to 31st October over 2007 to 2011. When applying the "Schaake Shuffle", the historical observations used to generate ensemble members are from 1st May to 31st October, 1981 to 2010. The results and verification presented in Section 4 is based on different canonical events to analyze the value of the designed events.

3.4. Verification method

The raw and post-processed forecasts are evaluated in two aspects using multiple statistical verification measures. One aspect is the accuracy of the ensemble forecast means, which measures the goodness of the ensemble forecast means relative to the observations. This kind of statistical measures includes bias, correlation coefficient and root mean square error skill score (RMSE-SS). The other aspect is the skill of the ensemble forecasts, including

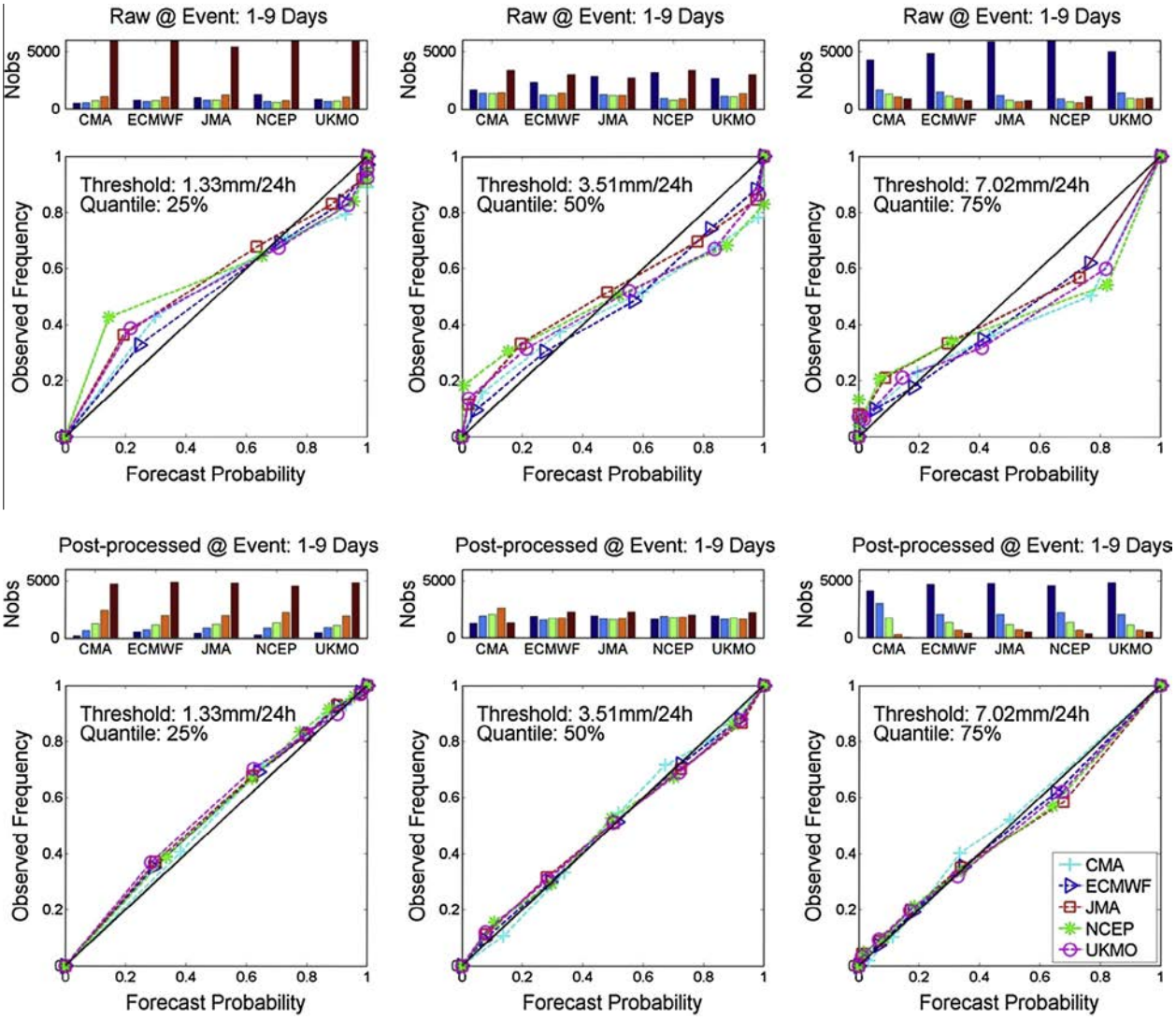

Fig. 11. Reliability diagram (RD) of the raw and post-processed ensemble forecasts of canonical event 8 (forecast period: 1–9 days) for five models. The different colors of lines denote different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reliability and calibration (Wilks, 1995). Brier skill score (BSS) measures improvements of the forecast of probability of precipitation (PoP) relative to climatology. Continuous ranked probability skill score (CRPSS) indicates the improvements of the probabilistic forecast of a continuous quantity to a reference forecast (Casati et al., 2008). Reliability diagram (RD) shows the closeness between the forecast probability and the observed frequency (Brocker and Smith, 2007). Verification by rank histograms assesses the calibration of the ensemble forecasts. Table 6 shows the specific definition of these statistical measures.

4. Results and analysis

The results in this section are post-processed ensemble precipitation forecasts of five individual NWP models by using EPP. Verification is done over the period from 1st June to 30th September, 2007 to 2011. The first (May) and last (October) months are not included for verification because of a lack of data before or after these months when calibrating the model parameters. Both the mean and spread of ensemble forecasts are verified. All results were compared to the raw forecasts and/or to the observations.

4.1. Verification of mean ensemble forecasts

Fig. 2 shows the biases of the raw mean ensemble forecasts relative to MAP daily observations averaged over 15 sub-basins and the verification period of all 11 canonical events for the five models. Biases of raw mean ensemble of all NWP models are quite significant and they do not show strong relationships to canonical events and lead times (the first four events: day 1–day 4). The biases of JMA and NCEP are the smallest and those of CMA and UKMO are the largest. The biases of the post-processed mean ensemble forecasts is not shown here, because they are very close to zero.

Fig. 3 shows the correlation coefficients between the post-processed mean ensemble forecasts and the corresponding observations averaged over 15 sub-basins and the verification period of 11 canonical events for the NWP models over the entire basin. The correlation coefficients of the raw forecasts and the post-processed mean ensemble forecasts are very similar and thus only post-processed ones are presented. Among the five models, the correlation of the post-processed results for all models except CMA is very similar and significant with values above 0.7 in some cases. The correlation coefficient is strongly related to the length of lead times, with shorter lead times having higher correlation.

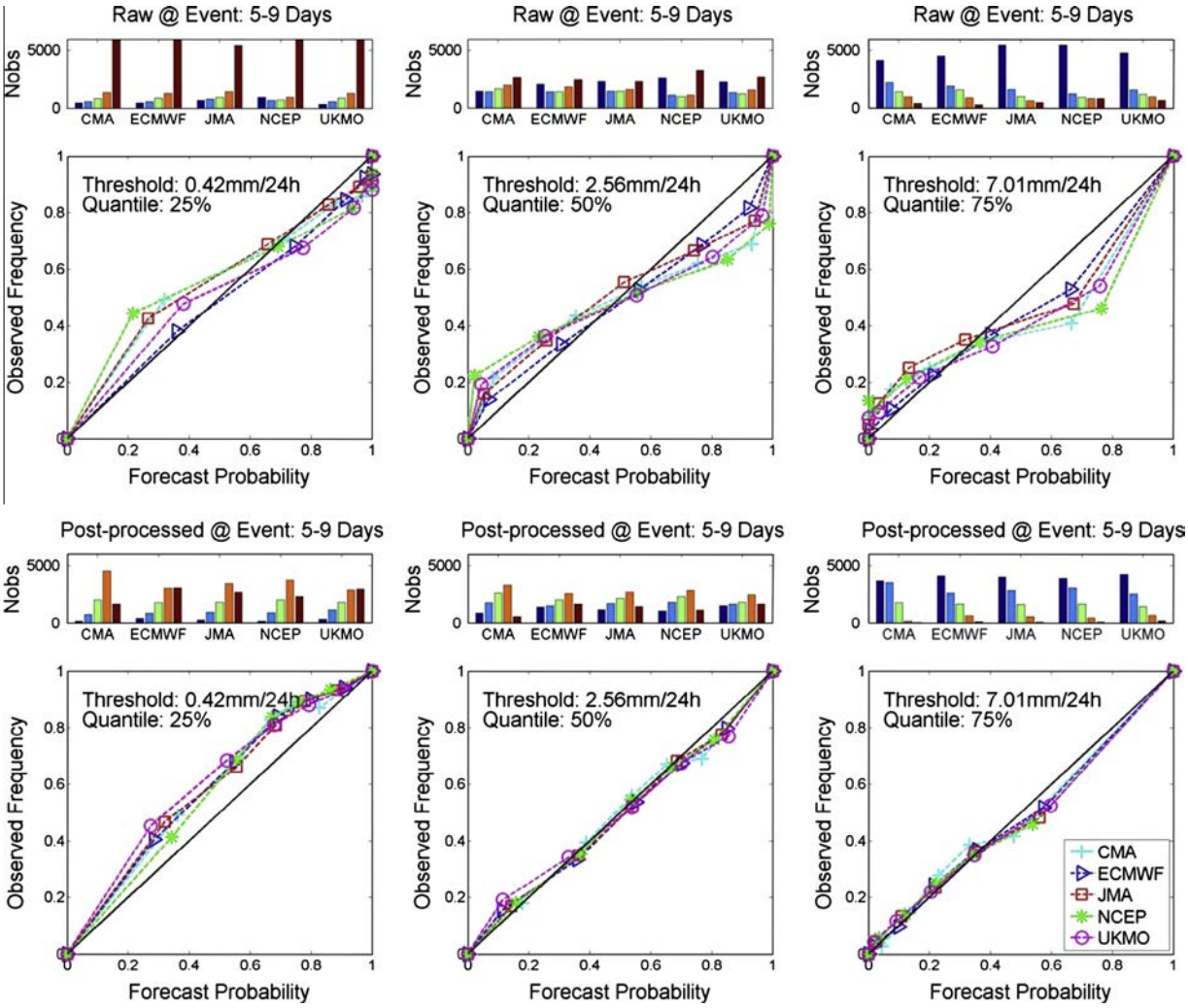

Fig. 12. Reliability diagram (RD) of the raw and post-processed ensemble forecasts of canonical event 10 (forecast period: 5–9 days) for five models. The different colors of lines denote different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Interestingly, the correlation for CCEs with the same lead time (i.e., events 5–7) is generally higher than those of SCEs. This has useful implications for water resources applications especially for cases in which exact timing of precipitation is not critical. It is notable that for CCEs with lead times over 6 days, the correlation is still meaningful at a value of 0.3 and above. Fig. 4 displays the correlation coefficients between the post-processed mean ensemble forecasts and the corresponding observations of different canonical events for 15 sub-basins. Warm color indicates high correlation while cold color indicates low correlation. Fig. 4 indicates that there are some slight differences between basins in terms of correlation coefficients. The overall patterns for all sub-basins in Fig. 4 show the same tendency as in Fig. 3. Again, the effect of lead times is obvious with shorter lead times having higher correlation. It also confirms that CCEs have higher correlation compared SCEs with the same lead times. The results here suggest that the design of canonical events should be considered carefully to maximize the information in precipitation forecasts. According to the chaos theory developed by Lorenz (1963), weather forecasts with lead times over 5 days are not to be reliable because a slight perturbation in initial conditions might result in large different in weather response. Because ensemble forecasts can account for the uncertainty in the initial conditions, mean ensemble forecasts with lead times over 6 days can still be meaningful.

Fig. 5 presents the root mean square error skill score (RMSE-SS) of the post-processed mean ensemble forecasts relative to the raw mean ensemble forecasts averaged over 15 sub-basins during the verification period for the five models. The larger the RMSS-SS values are, the more improvement is shown of the post-processed forecasts over the raw forecasts. For all models, RMSE-SS of the post-processed mean ensemble forecasts are quite significant, and thus the post-processing has improved the RMSE of raw forecasts substantially. This is especially true for CMA model, which is the worst in terms of RMSE for the raw forecasts. The same is true for UKMO.

4.2. Verification of ensemble forecasts

Fig. 6 displays the Brier skill score (BSS) of both the raw and post-processed ensemble forecasts relative to climatology for the 15 sub-basins and 11 canonical events. The raw forecasts already exhibit quite significant skills compared to climatology except for CMA (top panel in Fig. 6). The bottom panel in Fig. 6 shows that EPP can significantly improve all raw forecasts, even for CMA, whose value was increased to above 0.1.

Figs. 7 and 8 show the continuous ranked probability skill scores (CRPSS) for the raw and post-processed ensemble forecasts relative to climatology of 11 canonical events for the entire basin

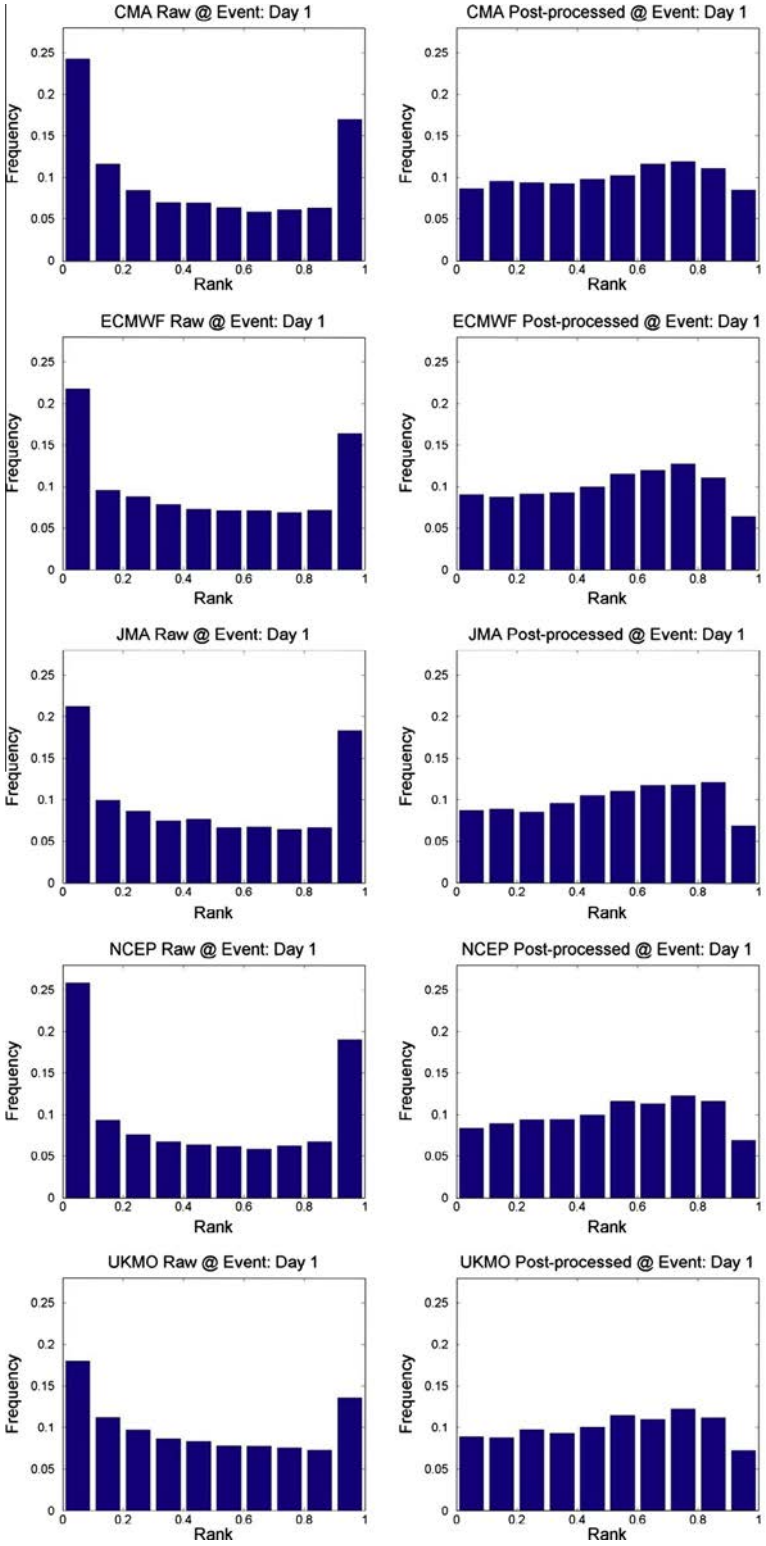

Fig. 13. Rank histogram of the raw and post-processed ensemble forecasts of canonical event 1 (forecast period: day 1) for five models.

and 15 individual basins. Fig. 7 shows that the CRPSS of the post-processed ensemble forecast has obviously been increased from that of the raw ensemble forecasts. The improvement is more obvious for CCEs with lead times over 6 days. Fig. 8 illustrates the improvement of the post-processed forecasts over the raw forecasts are true at individual basin level as well.

Figs. 9–12 present the reliability diagrams (RD) of different canonical events (Event 1 (day 1), 7 (1–4 days), 8 (1–9 days) and 10 (5–9 days)) for both raw and post-processed ensemble forecasts for the 5 models based on results from the entire basin for the verification period. The thresholds are decided by different quantiles of observations to show the performance of precipitation forecasts of different magnitudes. The interval of ensemble forecasts of each model are decided by the number of events in each bin to ensure an adequate amount of events falls into each interval. Compared to the raw ensemble forecasts, the post-processed ensemble

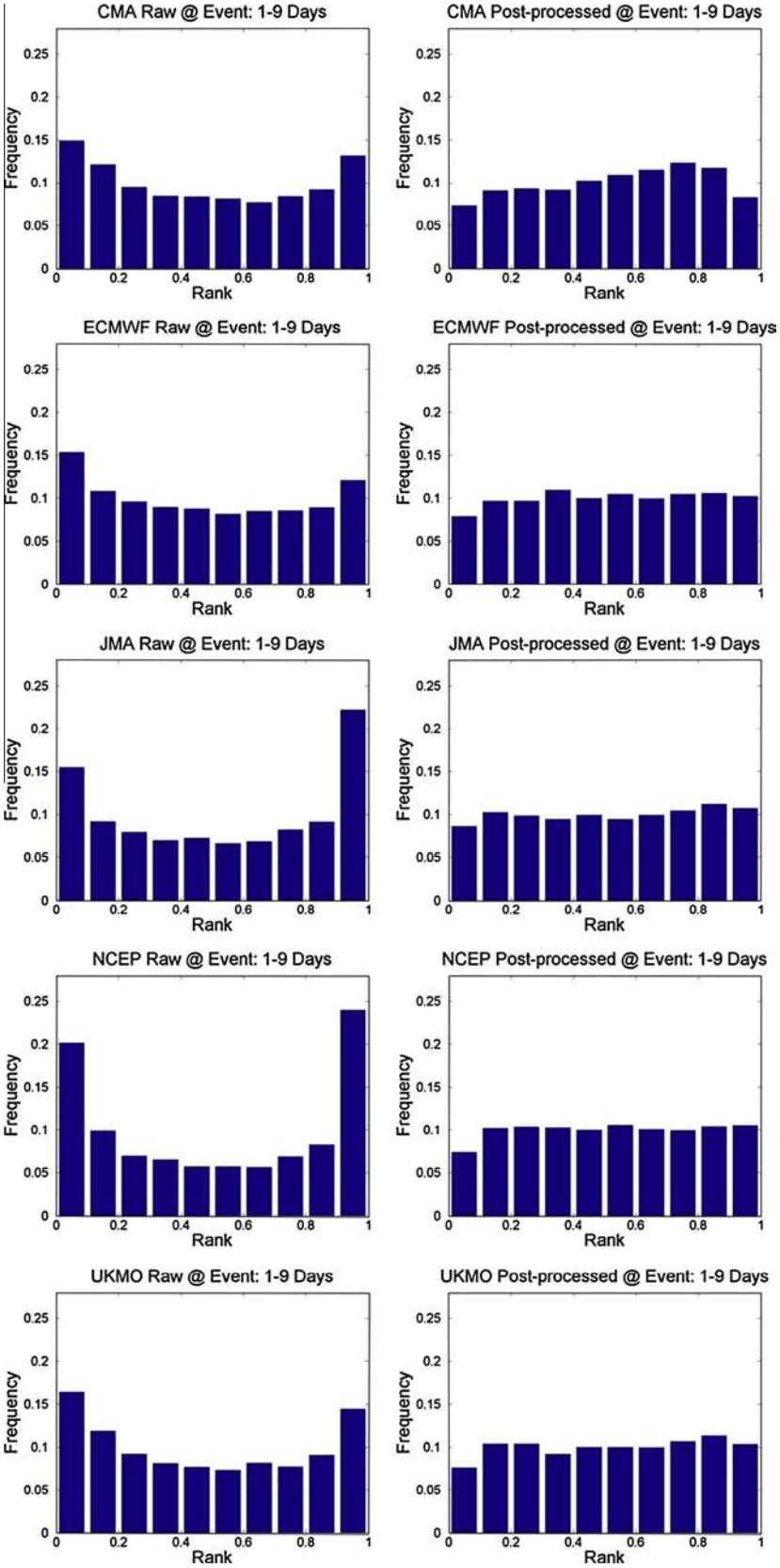

Fig. 14. Rank histogram of the raw and post-processed ensemble forecasts of canonical event 8 (forecast period: 1–9 days) for five models.

forecasts improved the spread significantly in all cases. The forecasts of low threshold events are generally bad in all cases except event 8 (1–9 days) for the post-processed results, suggesting that the forecast of exact precipitation amount is very difficult for small

storm events. Also, the estimation of PoP plays a very important role in predicting small storm events. Thus, by developing a more accurate way to estimate PoP in the future, we may improve the small storm prediction. In contrast, for high threshold events, the RDs of the raw forecasts in all cases are problematic, but the RDs of post-processed forecasts indicate excellent reliability.

Figs. 13 and 14 show the rank histogram of raw and post-processed ensemble forecasts of canonical events 1 (day 1) and 8 (1–9 days). From Fig. 13, the spread of raw ensemble forecasts of all models tends to be too narrow and thus many observations fall in out of range of 10–90%. The ranks of post-processed ensembles are much more uniform than that of raw ensembles though it slightly overestimates precipitation due to lack of observations fall into the last bin. For cumulative forecasts (Event 8), similar patterns to Fig. 13 exist for Fig. 14. Although raw forecasts still suffer from narrow spread, rank histograms of post-processed ensembles are relatively even.

5. Summary and conclusions

This paper aims to evaluate ensemble forecast skills of multiple models obtained from TIGGE database in Huai river basin, China. EPP is used as post-processer in 15 sub-basins over summer period (1st May–31st October) from 2007 to 2011. Common statistical verification metrics are used to verify the potential skill of the five typical models and EPP. Comparison and analysis are made between precipitation ensemble forecasts between models, as well as raw and post-processed forecasts.

Most measurements show that all models tend to be potentially skillful for use as inputs to hydrological models. ECMWF and JMA outperform other models according to multiple verification measurements, such as correlation coefficient, BSS and CRPSS. For canonical events with short lead times, JMA usually performs the best among all models, while ECMWF tends to have an advantage during longer lead times. UKMO also shows relatively high quality of results in multiple aspects, though its raw ensemble's skill is not as stable as the other models, as shown in BSS and CRPSS. NCEP generally shows lower reliability than ECMWF, JMA, and UKMO. However, it shows equally high quality in many canonical events, even outperforms some of those models in certain events. The skill of CMA shows an obvious gap to other models, though it gains reliability significantly after post-processing.

Comparing to raw ensemble forecasts, accuracy and reliability of post-processed ensemble forecasts have been significantly improved. For all NWP models, post-processed forecasts shows higher skill in both mean and spread of the ensemble members, especially in biases, BSS, CRPSS, reliability (shown as RD) and spread (shown as rank histogram) of raw ensembles. Therefore, EPP is demonstrated as an effective tool for hydrological forecasting.

The results show significant value of operational model data of TIGGE database and its potential skill using as hydrological inputs. On the other hand, it proves the necessity of post processing for precipitation forecasts from dynamic models to eliminate error and adjust ensemble spread. It is interesting that this study demonstrated that all countries can take advantages of the global coverage offered by advanced NWP models such as JMA and ECMWF, while continue to improve their own models. This study is done using only 5 years of hindcast data. Its ability to accurately capture events beyond the five historical records is still a question mark. As a next step of the research, we will examine effect of the length of hindcast data on forecasting skill. We will explore a better way to combine skillful NWP model forecasts to obtain improved multi-model precipitation ensembles for ensemble streamflow forecasting.

Acknowledgments

The authors would like to acknowledge the support provided by National Science and Technology Support Plan Program (No. 2013BAB05B04) and Natural Science Foundation of China (Grant No. 41375139). We thank Mr. Kai Tu in the Institute of Atmospheric Physics, the Chinese Academy of Sciences for helping prepare the TIGGE data. We further thank Professor Q.J. Wang for their valuable comments which improved the article significantly.

References

Atger, F., 1999. The skill of ensemble prediction systems. Mon. Weather Rev. 127, 1941–1953.
Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D., Ebert, B., Fuentes, M., Hamill, T.M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.-Y., Parsons, D., Raoult, B., Schuster, D., Dias, P.S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., Worley, S., 2010. The THORPEX Interactive Grand Global Ensemble (TIGGE). Bull. Am. Meteorol. Soc. 91, 1059–1072.
Brocker, J., Smith, L., 2007. Increasing the reliability of reliability diagrams. Weather Forecast. 22 (3), 651.
Buizza, R., Houtekamer, P.L., Toth, Z., Pellerin, G., Wei, M., Zhu, Y., 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Mon. Weather Rev. 133 (5), 1076.
Casati, B., Wilson, L., Stephenson, D., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E., Browne, B., Masonh, S., 2008. Forecast verification: current status and future directions. Meteorol. Appl. 15, 3–18.
Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., Wilby, R., 2004. The Schaake Shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. J. Hydrometeorol. 5 (1), 243–262.
Duan, M., Ma, J., Wang, P., 2012. Preliminary comparison of the CMA, ECMWF, NCEP, and JMA ensemble prediction systems. Acta Meteorl. Sin. 26 (1), 26–40.
Glahn, B., Peroutka, M., Wiedenfeld, J., Wagner, J., Zylstra, G., Schuknecht, B., Jackson, B., 2009. MOS uncertainty estimates in an ensemble framework. Mon. Weather Rev. 137 (1), 246.
Hamill, T., Colucci, S., 1998. Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. Mon. Weather Rev. 126 (3), 711–724.
Hamill, T., Whitaker, J., Mullen, S., 2006. Reforecasts: an important dataset for improving weather predictions. Bull. Am. Meteorol. Soc. 87 (1), 33–46.
Hamill, T., Hagedorn, R., Whitaker, J., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. Mon. Weather Rev. 136 (7), 2620–2632.
He, Y., Wetterhall, F., Bao, H., Cloke, H., Li, Z., Pappenberger, F., Hu, Y., Manful, D., Huang, Y., 2010. Ensemble forecasting using TIGGE for the July–September 2008 floods in the Upper Huai catchment: a case study. Atmos. Sci. Lett. 11, 132–138.
Huai River Commission, 2009. Huai river water resources bulletin for year 2009. Chin. Ministry Water Resour., 29, in Chinese.
Kelly, K.S., Krzysztofowicz, R., 1997. A bivariate meta–Gaussian density for use in hydrology. Stoch. Hydrol. Hydraul. 11 (1), 17–31.
Krishnamurti, T.N., Kishtawal, C., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S., Surendran, S., 2000. Multimodel ensemble forecasts for weather and seasonal climate. J. Clim. 13 (23), 4196–4216.
Krzysztofowicz, R., 1997. Transformation and normalization of variates with specified distributions. J. Hydrol. 197 (1–4), 286–292.
Liu, Y., Duan, Q., Zhao, L., Ye, A., Tao, Y., Miao, C., Mu, X., Schaake, J., 2013. Evaluating the predictive skill of post-processed NCEP GFS ensemble precipitation forecasts in China's Huai river basin. Hydrol. Process. HEPS Spec. Issue 27, 57–74.
Lorenz, E.N., 1963. Deterministic non-periodic flow. J. Atmos. Sci. 20, 130–141.
Park, Y., Buizza, R., Leutbecher, M., 2008. TIGGE: preliminary results on comparing and combining ensembles. Q. J. Roy. Meteorol. Soc. 134, 2029–2050.
Raftery, Adrian E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133 (5), 1155.
Rayner, S., Lach, D., Ingram, H., 2005. Weather forecasts are for wimps: why water resource managers do not use climate forecasts. Clim. Change 69, 197–227.
Robertson, D.E., Shrestha, D., Wang, Q., 2013. Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. Hydrol. Earth Syst. Sci. 17 (9), 3587–3603.
Roulston, M.S., Smith, L., 2003. Combining dynamical and statistical ensembles. Tellus A 55 (1), 16–30.
Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., Seo, D.J., 2007. Precipitation and temperature ensemble forecasts from single-value forecasts. Hydrol. Earth Syst. Sci. Discuss. 4, 655–717.
Shapiro, M., Thorpe, A., 2004. THORPEX International Science Plan WMO/TD.1246, WWRP/THORPEX. 2, 51.
Thielen, J., Schaake, J., Hartman, R., Buizza, R., 2008. Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007. Atmos. Sci. Lett. 9 (2), 29–35.

Toth, Z., Kalnay, E., 1997. Ensemble forecasting at NCEP and the breeding method. Mon. Weather Rev. 125, 3297–3319.
Voisin, N., Schaake, J., Lettenmaier, D., 2010. Calibration and downscaling methods for quantitative ensemble precipitation forecasts. Weather Forecast., 100804092600065.
Wilks, D.S., 1995. Statistical Methods in the Atmospheric Sciences: An Introduction. International Geophysics Series, vol. 59. Academic Press.
Wood, A., Schaake, J., 2008. Correcting errors in streamflow forecast ensemble mean and spread. J. Hydrometeorol. 9 (1), 132.
Wu, L., Seo, D.J., Demargne, J., Brown, J.D., Cong, S., Schaake, J., 2011. Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. J. Hydrol. 399, 281–298.
Yun, W.T., Stefanova, L., Mitra, A., Kumar, T., Dewar, W., Krishnamurti, T., 2005. A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts. Tellus A 57 (3), 280–289.
Zhang, Y., Arthington, A., Bunn, S., Mackay, S., Xia, J., Kennard, M., 2011. Classification of flow regimes for environmental flow assessment in regulated rivers: the Huai River Basin, China. River Res. Appl. 26, 1–17.