HYDROLOGICAL PROCESSES
Hydrol. Process. **27**, 57–74 (2013)
Published online 7 September 2012 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.9496

Evaluating the predictive skill of post-processed NCEP GFS ensemble precipitation forecasts in China's Huai river basin

Y. Liu,[1] Q. Duan,[2]* L. Zhao,[3] A. Ye,[2] Y. Tao,[4] C. Miao,[2] X. Mu[4] and J. C. Schaake[5]

[1] *Department of Atmospheric Science, Chengdu University of Information Technology, Chengdu, China*
[2] *State Key Laboratory of Earth Surface Processes and Resource Ecology, College of Global Change and Earth System Science Beijing Normal University, Beijing, China*
[3] *Public Meteorological Service Center and National Meteorological Center, Chinese Meteorological Administration, Beijing, China*
[4] *College of Mathematical Science, Beijing Normal University, Beijing, China*
[5] *Retired, National Weather Service/NOAA, Silver Spring, MD, USA*

Abstract:

The National Center for Environmental Predictions (NCEP) has produced an ensemble meteorological reforecast product by using a fixed version of Global Forecast System (GFS) ensemble prediction system since 1 January 1979. The 15-member ensemble product, with a global coverage at a 2.5° × 2.5° spatial resolution and a 14-day lead time, has been used successfully by the River Forecast Centers of the National Weather Service (NWS) to produce basin scale precipitation and temperature ensemble forecasts in the US for several years now. This study evaluates the predictive skill of post-processed ensemble forecasts based on GFS precipitation reforecast in China's Huai river basin. The evaluation is carried out in 15 sub-areas of the Huai river basin and covers the 1/1/1981–31/12/2003 period. The Ensemble Pre-Processing system version 3 (EPP3), developed at NWS, is used to develop joint probability distributions between forecasted ensemble mean precipitation and corresponding observations and to generate individual ensemble members that preserve space–time correlation of the observed precipitation data. Several statistical verification measures are used to quantify the goodness of fit between post-processed (i.e. EPP3 processed) ensemble mean and observation and to assess the ensemble spread. Results indicate that the post-processed forecasts have meaningful predictive skill for the first few days for ensemble daily precipitation forecasts. Predictive skill of ensemble forecasts of cumulative precipitation for lead times up to 14 days are significant. The forecast skill is highly dependent on seasonality, with relatively lower skills seen for wet summer season, when convective storm patterns dominate, as compared with other seasons. The predictive skill of the post-processed ensemble precipitation is much better than the raw forecasts and the climatological ensemble forecasts. The results from this study suggest that the NCEP's GFS reforecasts can be a valuable resource for places other than the US.

KEY WORDS ensemble precipitation forecasting; hydrologic ensemble prediction; Huai river basin; ensemble verification GFS reforecast product

Received 28 November 2011; Accepted 11 July 2012

INTRODUCTION

Even though weather forecasts have become essential in people's life today, hydrologists have made limited use of weather forecasts in their hydrologic forecasts until recently. There are numerous reasons for this (Rayner *et al*., 2005). Inadequate accuracy and reliability, especially in early weather forecast products, was one reason preventing their wide use. Another reason is that weather forecast products were not well designed for hydrologic applications. Weather and hydrologic forecasts differ in both space and time scales. Typical hydrologic forecasts are developed over a hydrologic basin, whose drainage area ranges from 10 s to 1000 s of km^2, whereas typical weather forecasts are made on spatial grids that are usually much larger in spatial scales, especially for the global forecasts. Therefore, weather forecasts need to be downscaled to basin scales to be used for hydrologic applications (Schaake *et al*., 2010). The third reason that hydrologists hesitate to use weather forecasts is that conventional weather forecasts are usually given as single-valued deterministic forecasts and thus lack the uncertainty information needed for risk-based decisions. Finally, weather forecasts often have phase shifts in space and time between what is predicted and what actually happens.

Over recent years, remarkable progress has been made as numerical weather prediction (NWP) models with better physics representations and better forecasting methodologies have been developed. Ensemble weather forecasting methodology has emerged as an effective way to account for uncertainty in weather forecast products (Lewis, 2005). The rationale behind ensemble forecasting was laid by Edward Lorenz many years ago when he discovered the inherent chaotic nature of the weather system (Lorenz, 1963), but it was not until 1990s when the European Center for Medium Range Weather Forecasts and National Center for Environmental Predictions (NCEP) of the US National Weather Service (NWS) started to produce operational ensemble weather forecasts (Molteni *et al*., 1996; Toth and Kalnay, 1997). Today, ensemble weather forecast products have become standard in many NWP centres around the world

*Correspondence to: Qingyun Duan, State Key Laboratory of Earth Surface Processes and Resource Ecology, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China, 100875. E-mail: qyduan@bnu.edu.cn

(Bougeault *et al.*, 2010). In recent years, hydrologists have been taking advantages of the newly available ensemble weather forecasts in producing hydrologic ensemble forecasts (Cloke and Pappenberger, 2009, Thiemig *et al.*, 2010, Webster *et al.*, 2010, Pappenberger *et al.*, 2011). Despite the advantage and usefulness of ensemble weather forecasts (Park *et al.*, 2008; Froude, 2010; He *et al.*, 2010), raw ensemble forecasts still suffer from various limitations for hydrologic applications, including the needs for spatial downscaling and for correction of forecast biases in terms of not only means but also ensemble spread (Hamill *et al.*, 2004).

Statistical post-processing has been a common approach used to enhance the usability of raw weather forecasts. A widely used approach is the Model Output Statistics (MOS) method, which uses multivariate regression techniques to relate NWP model generated forecasts (e.g. forecasts of precipitation, temperature and geopotential height) to relevant observations to obtain improved weather forecasts for given lead times and locations (Glahn and Lowry, 1972). However, MOS outputs are still not well suited for hydrologic applications because they cannot preserve the space–time statistical relationships of the forecasted variables (Clark *et al.*, 2004). To make weather forecasts more useful, techniques specifically designed for hydrologic applications are needed. Krzysztofowicz and Sigrest (1999) presented a Bayesian ensemble method that can generate short-range (with lead times of a few days) probabilistic quantitative precipitation forecasts from raw NWP, MOS and/or human forecasts. Clark *et al.* (2004) improved MOS multivariate regression approach by introducing a procedure known as 'Schaake shuffle' to create ensembles of precipitation forecasts that preserve space–time statistical relationships of observed precipitation.

Schaake *et al.* (2007) developed a method (hereafter referred to as Schaake's method) to construct ensemble precipitation and temperature forecasts from single-valued meteorological forecasts. Schaake's method has been implemented in the Ensemble Pre-Processor (EPP) of the National Weather Service River Forecast System and been run experimentally at several river forecast centres in the NWS for several years (Wu *et al.*, 2011). One requirement for using this method is the availability of a relatively long historical archive of weather forecasts and corresponding observations, so the parameters of the probabilistic model can be calibrated properly (Fundel and Zappa, 2011). NCEP has produced a medium range (1–14 day), 15-member ensemble reforecast product over a period starting from 1 January 1979, to present using a fixed version of the Global Forecast System (GFS) (Hamill *et al.*, 2004). The GFS temperature and precipitation reforecast products have been widely used by river forecast centres to generate basin scale ensemble quantitative temperature forecasts and quantitative precipitation forecasts (QPFs), which in turn have been successfully used to make ensemble hydrological forecasts (Schaake *et al.*, 2007; Wu *et al.*, 2011).

The China Meteorological Administration (CMA) has been working to improve forecasts of high impact weather events for China. Numerous studies were conducted specifically to improve precipitation forecasting (Zhou *et al.*, 2001; Song *et al.*, 2006; Xu *et al.* 2007; Zhu *et al.*, 2007). Zhu *et al.* (2007) described a study to use the GRAPES model, an experimental operational NWP model in CMA, to make 72-h precipitation forecast for the 2005 summer monsoon period in the Huai river basin. Xu *et al.* (2007) presented the results from an ensemble precipitation forecasting experiment that aims to improve the prediction of heavy rainfall events and found that ensemble prediction performs better than single deterministic control run in forecasting precipitation amount of severe storms. One problem for CMA, however, is lack of a long historical archive of China's own NWP reforecasts that are needed to generate basin scale ensemble QPF for hydrologic applications. Because the NCEP GFS reforecast product has global coverage, it would be interesting to see if the GFS products from the US can be used effectively in other countries including China.

The primary purpose of this paper is, therefore, to demonstrate the potential usefulness of the GFS reforecasts in Chinese settings, particularly by analysis of precipitation forecasts for the Huai river basin. The scientific questions we try to address are as follows: (1) Is there verifiable skill in GFS precipitation reforecasts in the Huai River basin? and (2) How does the forecast skill vary with lead time and seasonality? This paper is organized as follows. The second section provides a brief description of the methodology used in the study. The third section describes the basin and experimental design. The fourth section presents and discusses the experimental results. Summary and conclusion are given in the fifth section.

A BRIEF DESCRIPTION OF THE METHODOLOGY

A brief description of Schaake's method, as implemented in the Ensemble Pre-Processing system version 3 (EPP3) is presented here. EPP3 is part of the operational National Weather Service River Forecasting System, and its source code can be obtained from the NWS Hydrology Laboratory. More detailed descriptions of EPP3 with rigorous mathematical derivations can be found in Schaake *et al.* (2007) and Wu *et al.* (2011). Figure 1 is the flowchart summarizing the main steps in Schaake's method. Here, a non-mathematical description is given. Let X and Y denote sets of single-valued QPFs and corresponding observations on a specific forecast day and with a specific lead time over the study domain. To ensure enough sample points in X and Y, a time window of adequate length is selected to include forecast/observation pairs before and after the forecast day. The length of window should meet two conditions: (1) there should be more than a minimum number of non-zero rain events within the window (e.g. say 30) and (2) the stochastic nature of the considered variables within the time window remains approximately the same. In EPP3, an upper limit of 60 days is set for the precipitation window, and a fixed 45-day window is used for temperature. These numbers can be changed depending on the climatic conditions of the study area and the forecast day. In determining whether a

1. Precipitation distribution:

A. *Choosing a time window:*
The width of window should assure a certain number of non-zero forecasts or observations.

B. *Generating marginal distribution of forecast and observation respectively:*
$F_X(x) = 1 - p_X + p_X F_{XC}(x|x > 0)$
where p_X is the probability of occurrence of precipitation; $F_Y(y)$ have similar form.

C. *Applying normal quantile transform (NQT):*
$x = F^{-1}{}_X(Q(u))$, if $u > u_0$; $x = 0$, otherwise.
$y = F^{-1}{}_Y(Q(v))$, if $v > v_0$; $y = 0$, otherwise.
where Q is standard normal cdf; u_0 satisfies $Q(u_0) = 1 - p_X$; v_0 satisfies $Q(v_0) = 1 - p_Y$.

D. *Assuming U and V are bivariate standard normal (BSN):*

$$(U,V) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}, \rho_{UV}\right)$$

where ρ_{UV} is the correlation coefficient of U and V. Thus, given $u = u_1$,

$V|U_1 \sim N(\rho_{UV} u_1, 1 - \rho_{UV}^2)$

E. *Generating conditional distribution of y corresponding to x:*
Given $x = x_1 > 0$:
$F_{Y|X}(y|x = x_1) = B_{V|U}(v|u = u_1)$
where $B_{V|U}(v|u=u_1)$ is the conditional distribution of BSN.
Given $x = 0$:

$$F_{Y|X}(y|x = 0) = B_{V|U}(v|u \le u_0)$$

$$= \frac{\int_{-\infty}^{v}\left(\int_{-\infty}^{u_0} b(u|v)du\right)b(v)dv}{\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{u_0} b(u|v)du\right)b(v)dv} = \frac{\int_{-\infty}^{v} B(u_0|v)b(v)dv}{\int_{-\infty}^{+\infty} B(u_0|v)b(v)dv}$$

where $U_0|V \sim N(\rho_{UV} v, 1 - \rho_{UV}^2)$.

2. Ensemble construction:

A. *Stratified Sampling:*
Produce n values of Y from $F_{Y|X}(y|x = 0)$ using stratified sampling if there are n

Figure 1. The flowchart describing major steps of the Schaake method

precipitation forecast is a rain or a no-rain event, we used a threshold value that is corresponding to 97% exceedance probability value of all non-zero forecasts. The marginal distributions of X and Y are fitted to a parametric probability distribution function (PDF). EPP3 allows Gamma, Weibull or Exponential distribution function to be used. For mathematical convenience, the fitted X and Y are then transformed from original space to Gaussian space (denoted as U and V) by using a procedure known as Normal Quantile Transform (Krzysztofowicz, 1997). Thus, the joint probability distribution between U and V, $F(U,V)$, can be assumed as Bivariate Standard Normal under certain conditions (i.e. both U and V and Gaussian, and there exists meaningful correlation between U and V). Given a forecast date, conditional PDF $F_{Y|X}(Y|X)$, which is equal to $F_{V|U}(V|U)$, is first estimated for a future event (with a given lead time and duration) on the basis of $F(U,V)$. Once $F_{Y|X}(Y|X)$ is determined, a pre-specified number (i.e. ensemble size, which is equal to the number of years of available observed data) of stratified samples are taken from this conditional distribution. After $F_{Y|X}(Y|X)$ for all defined events is determined, the sampled values for different lead times are then connected using the 'Schaake shuffle' procedure to form time series of individual ensemble members (Clark *et al.*, 2004). The space–time Spearman rank correlation structure of the generated ensemble members is similar to that of the observed historical events. Each of the ensemble members can be used directly to drive the hydrologic model to produce ensemble streamflow forecast.

STUDY AREA, DATA AND EXPERIMENTAL DESIGN

Overview of study area

The Huai river basin, one of the seven major river systems in China, is chosen as the study area. Situated between Yangtze and Yellow rivers in Eastern China with an approximate drainage area of 270,000 km^2, the Huai river basin encompasses the whole or part of five provinces with a dense population of 185 million people. With 14.27 million hectares of cultivated lands, agriculture is an important sector that accounts for 17.4% of China's agricultural outputs (Wang *et al.*, 2009). Water resources in the Huai river basin are regulated by four large natural lakes and a network of over 300 man-made reservoirs to control floods, generate electricity, supply water and protect the environment (Huai River Commission, 2009). The Huai river basin is known to be most affected by various water-related hazards because of its transitional climate that shifts from the north sub-tropical zone to the warm temperate zone in the south. The space–time distribution of precipitation varies greatly within the basin, with southern mountainous and coastal regions receiving annual precipitation over 1600 mm and northern inland regions under 700 mm. About 50%–80% of precipitation occurs during the June–September flooding season. Floods and droughts have frequently ravaged the area. Over the last 20 years, three major floods (in 1991, 2003 and 2007) wreaked havoc on the Huai river basin, causing tremendous losses of human lives, properties and economic production. Compounding flooding hazards are the occurrences of abrupt switches from drought to flooding during the spring/summer transition period. This poses special difficulties to agricultural, navigational and fishery sectors. Wang *et al.* (2009) reported that abrupt drought-to-flooding switch occurs about once every 4 years and is becoming more frequent in recent years. Accurate and reliable weather and hydrologic forecasts are critical to the sustainable economic development and the protection of human lives and the environment.

Data used for the study

Two kinds of data are used in the study: the historical GFS precipitation reforecasts and gauge-based mean areal precipitation (MAP) data. The raw GFS data used in this study are a 2.5° × 2.5° gridded ensemble forecast product with a daily time step, a 14-day lead time and 15 ensemble members. Reforecasts cover the data period from 1 January 1979 to 30 April 2004. The particular grid points used for the Huai river basin are shown in Figure 2. The ensemble mean of the raw ensemble forecasts is computed and is used as the single-valued forecast in this study. The reason that ensemble mean instead of the individual ensemble members is used is because the ensemble mean has more significant correlation (0.4 or above) with the observation compared with the individual ensemble members. Hamill *et al.* (2008) also pointed out that 'raw probabilistic forecasts from the ensemble prediction systems' relative frequency possessed little or negative skill' when they were evaluating GFS raw ensemble precipitation forecasts. Daily MAP data for 15 sub-areas of the Huai river basin are computed from historical observed precipitation station data from 167 stations for the data period from 1 January 1981 to 31 December 2009.

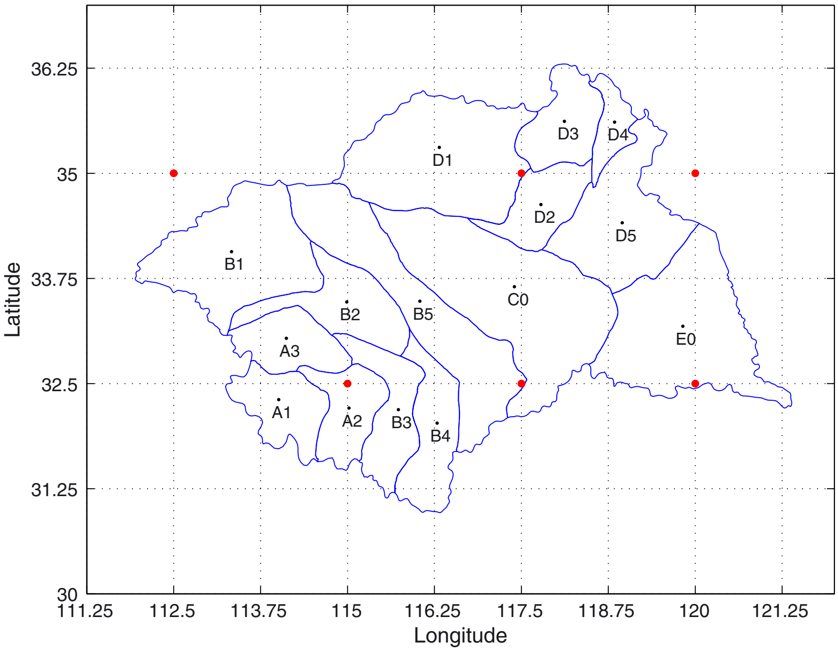

Figure 2. The illustration of the catchments in Huaihe basin (the red dots are the location of the GFS NWP model data, and the asterisks are the centres of each catchments

Table I. The information of Huai river basin's sub-areas

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

The ensemble forecast experiment

The Huai river basin is subdivided into 15 sub-areas. The division is chosen to be consistent with the operational meteorological forecast units for the Huai river basin. Figure 2 shows the boundaries of the 15 sub-areas. Circles represent the centres of the GFS data grids used, whereas solid dots represent the centres of the sub-areas. Table I lists the sub-area geographic information and annual precipitation statistics. Figure 3 displays monthly climatic precipitation statistics that show the intra-annual patterns of precipitation.

The MAP data for the 15 sub-areas are computed using Thiessen Polygon method. These MAP and corresponding GFS ensemble mean forecast data at the nearest grid points are used to estimate the parameters of the statistical model defining the joint forecast and observation PDFs for each of the 15 sub-areas for lead times 1–14 days for every 5 days in the year. The parameters are then linearly interpolated on each day of the year. The calibration data for this study are from 1 January 1981 to 31 December 2003 (note that the partial year GFS data from 2004 are not used in the calibration). Using the calibrated joint PDFs and GFS ensemble mean forecasts, the conditional PDFs on each day of the calibration period and for each lead time (1–14 days) are computed. These conditional PDFs are then used to construct ensemble members by using the Schaake shuffle procedure (Clark *et al*., 2004). The results and verification statistics are presented later in Results and analysis section.

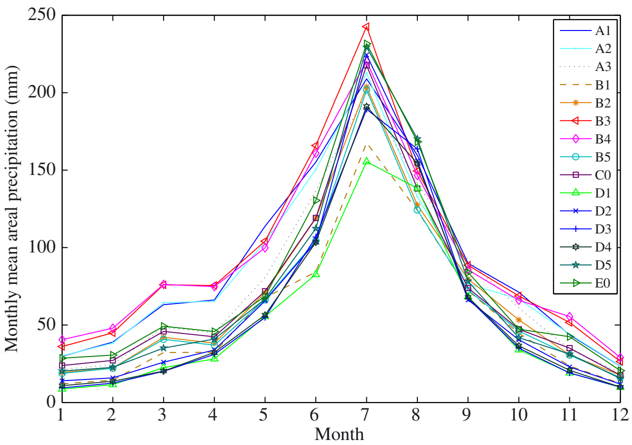

Figure 3. Monthly mean areal precipitation for catchments of Huai river basin (mm)

Forecast verification statistics

To evaluate the forecast skill of raw forecasts and statistically post-processed forecasts, we utilize two types of statistical measures. One type of statistical measure is goodness of fit between ensemble mean forecast and observation. These include Pearson correlation coefficient, relative bias and root mean square error (RMSE) between ensemble means and the observed values. The second type is ensemble verification measures, including Brier skill score (BSS) to measure the improvement of the probabilistic forecast of a binary event (i.e. rain or no rain) relative to a reference forecast, continuous ranked probability skill score (CRPSS) to measure the improvement of the probabilistic forecast of a continuous quantity (i.e. precipitation amount) relative to a reference forecast and reliability diagram (RD) to indicate the closeness of forecast probability and observed frequency. See the Appendix for mathematical definitions of these statistical verification measures. The reference forecast used in the verification of ensemble forecasts is the climatological forecasts. Raw individual ensemble members are not analysed because the raw ensemble spread has little skill.

RESULTS AND ANALYSIS

We obtained the EPP3 post-processed ensemble precipitation forecasts for all days, all lead times (1–14 days) and all sub-areas for the calibration period (1/1/1981–31/12/2003). These forecasts are compared with raw GFS ensemble mean forecasts and climatological forecasts by using the statistical verification measures described in Forecast Verification Statistics section.

Verification of mean ensemble forecasts

The post-processed ensemble forecast means are first compared with raw mean ensemble forecasts of daily precipitation and cumulative precipitation over different lead times. Forecasts of exact daily precipitation traces are important in flood forecasting because timing is critical. However, forecasts of cumulative precipitation are sometimes more useful to water resources managers who care about how much inflow comes into the reservoirs they manage. Table II shows the biases of the raw and post-processed mean ensemble forecasts relative

Table II. Basin averaged absolute relative biases of raw and post-processed mean ensemble forecasts of daily precipitation for different lead times (1–14 days) for the Huai river basin

	Lead time (day)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Raw–Spring	1.29	1.43	1.31	1.00	0.72	0.65	0.66	0.68	0.66	0.66	0.66	0.63	0.64	0.61
Raw–Summer	2.39	2.27	2.36	2.27	2.14	2.11	2.08	2.11	2.10	2.07	2.13	2.19	2.26	2.30
Raw–Fall	1.34	2.41	2.83	2.86	2.90	3.20	3.44	3.43	3.37	3.40	3.47	3.41	3.31	3.30
Raw–Winter	0.54	1.04	0.95	0.63	0.63	0.70	0.85	0.98	0.97	1.00	1.00	1.00	0.99	0.98
Post-Processed–All seasons	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.04

to the MAP daily observations averaged over 15 sub-areas of the Huai river basins. Figure 4 displays biases of the raw and post-processed mean ensemble daily forecasts relative to the MAP daily observations for different lead times and seasons for two typical sub-areas: the relatively wet sub-area B3 and the relatively dry sub-area D1. Table III reveals the biases of raw and post-processed cumulative precipitation mean ensemble forecasts averaged over 15 sub-areas of the Huai river basins. Figure 5 displays the biases of raw and post-processed cumulative precipitation mean ensemble forecasts for different lead times and seasons for sub-areas B3 and D1. For both raw mean ensemble forecasts of daily and cumulative precipitation amounts, relative biases tend to be positive, and the biases can be larger than 100% of the observed values (i.e. with bias value over 1.0). Biases in summer and fall are larger than those in spring and winter with values as high as 3 or above. This is due to the fact that summer and fall precipitation events are more severe and are convective type, which is harder to forecast than the frontal events encountered mostly in winter and spring seasons. The raw daily averaged mean biases do not seem to have a definitive relationship with the length of lead times. On the other hand, post-processed ensemble forecast means can effectively remove all the mean biases in forecasts of daily values as well as in cumulative precipitation forecasts, with the

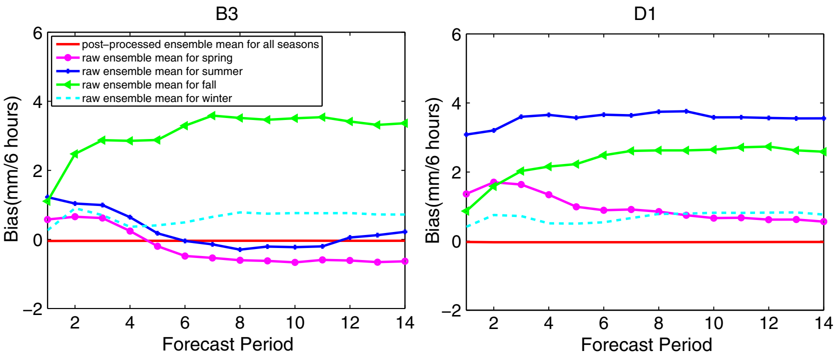

Figure 4. Relative biases of raw and post-processed mean ensemble forecasts of daily precipitation for different lead times (1–14 days) and sub-areas B3 and D1. The first line indicates biases of the post-processed ensemble forecast means. The other four lines indicate biases of raw ensemble forecast means for four different seasons

Table III. Basin averaged absolute relative biases of raw and post-processed mean ensemble forecasts of cumulative precipitation for different lead times (1–14 days) for the Huai river basin

	Lead time (day)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Raw–Spring	1.28	1.35	1.34	1.25	1.11	0.98	0.90	0.85	0.81	0.77	0.75	0.72	0.71	0.70
Raw–Summer	2.38	2.33	2.33	2.28	2.19	2.15	2.12	2.11	2.10	2.09	2.09	2.10	2.11	2.12
Raw–Fall	1.33	1.88	2.19	2.36	2.47	2.59	2.71	2.80	2.86	2.92	2.97	3.01	3.03	3.05
Raw–Winter	0.54	0.79	0.84	0.78	0.76	0.75	0.76	0.79	0.80	0.83	0.84	0.86	0.87	0.88
Post-Processed–All seasons	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

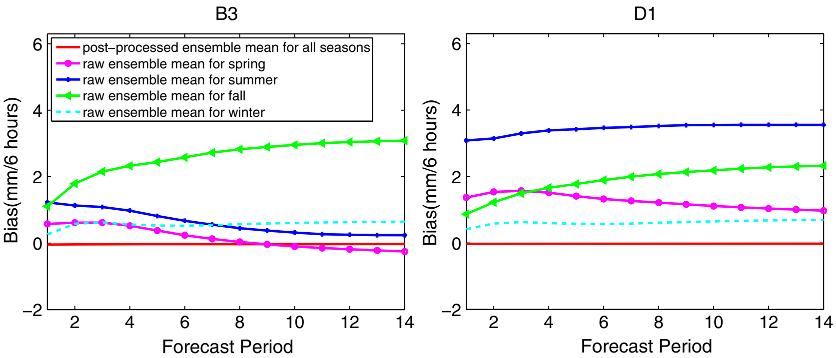

Figure 5. Relative biases of raw and post-processed mean ensemble forecasts of cumulative precipitation for different lead times (1–14 days) and sub-areas B3 and D1. The first line indicates biases of the post-processed ensemble forecast means. The other four lines indicate biases of raw ensemble forecast means for four different seasons

absolute relative bias values of less 0.04 (i.e. 4%) (see Tables II and III).

Tables IV and V present the root mean square error skill scores (RMSESs) of the post-processed ensemble forecast means. RMSES measures the improvement in RMSE of a forecast relative to that of a reference forecast, which is the raw mean ensemble forecast in this case. A RMSES value greater than 0 indicates the forecast is better than the reference forecast. Figures 6 and 7 exhibit the RMSES values of post-processed mean ensemble forecasts for sub-areas B3 and D1. Figure 6 shows the RMSES values for daily mean ensemble forecasts at different lead times and

Table IV. Basin averaged RMSES of post-processed ensemble forecast means of daily precipitation for different lead times (1–14 days) for the Huai river basin

	Lead time (day)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Spring	0.081	0.130	0.125	0.108	0.078	0.070	0.064	0.061	0.053	0.048	0.048	0.044	0.045	0.041
Summer	0.186	0.270	0.288	0.264	0.217	0.203	0.178	0.168	0.153	0.138	0.131	0.126	0.124	0.119
Fall	0.320	0.474	0.491	0.468	0.422	0.419	0.423	0.393	0.359	0.349	0.342	0.318	0.295	0.281
Winter	0.297	0.497	0.456	0.347	0.318	0.324	0.332	0.341	0.318	0.300	0.288	0.258	0.248	0.224

Table V. Basin averaged RMSES values of post-processed ensemble forecast means of cumulative precipitation for different lead times (1–14 days) for the Huai river basin

	Lead time (day)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Spring	0.081	0.125	0.145	0.151	0.148	0.146	0.143	0.140	0.138	0.136	0.135	0.135	0.136	0.136
Summer	0.186	0.270	0.317	0.338	0.343	0.347	0.351	0.355	0.358	0.361	0.365	0.370	0.376	0.381
Fall	0.320	0.467	0.544	0.586	0.607	0.628	0.651	0.668	0.680	0.691	0.704	0.715	0.723	0.730
Winter	0.297	0.463	0.523	0.527	0.522	0.520	0.524	0.532	0.542	0.550	0.559	0.567	0.574	0.579

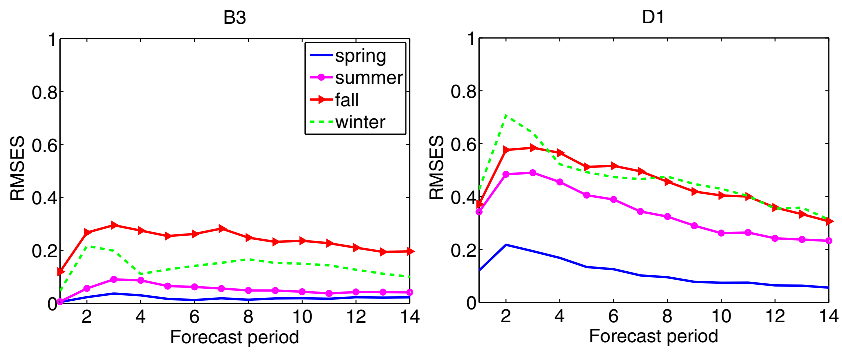

Figure 6. Root mean square error skill score (RMSES) of post-processed mean ensemble forecasts of daily precipitation for different lead times (1–14 days) and sub-areas B3 and D1. The four lines denote values for four different seasons

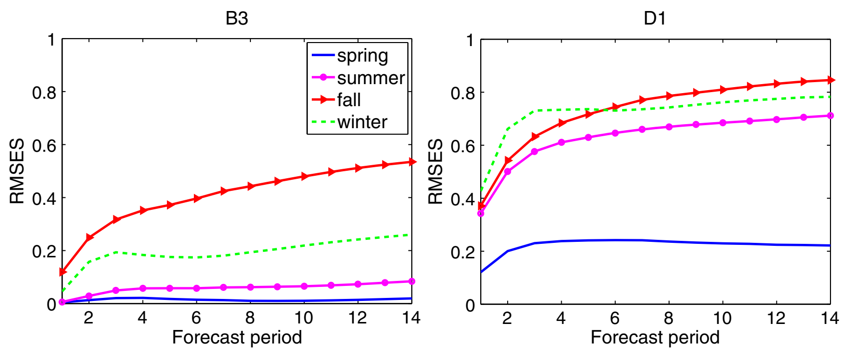

Figure 7. Root mean square error skill score (RMSES) of post-processed mean ensemble forecasts of cumulative values for different lead times (1–14 days) and sub-areas B3 and D1. The four lines denote values for four different seasons

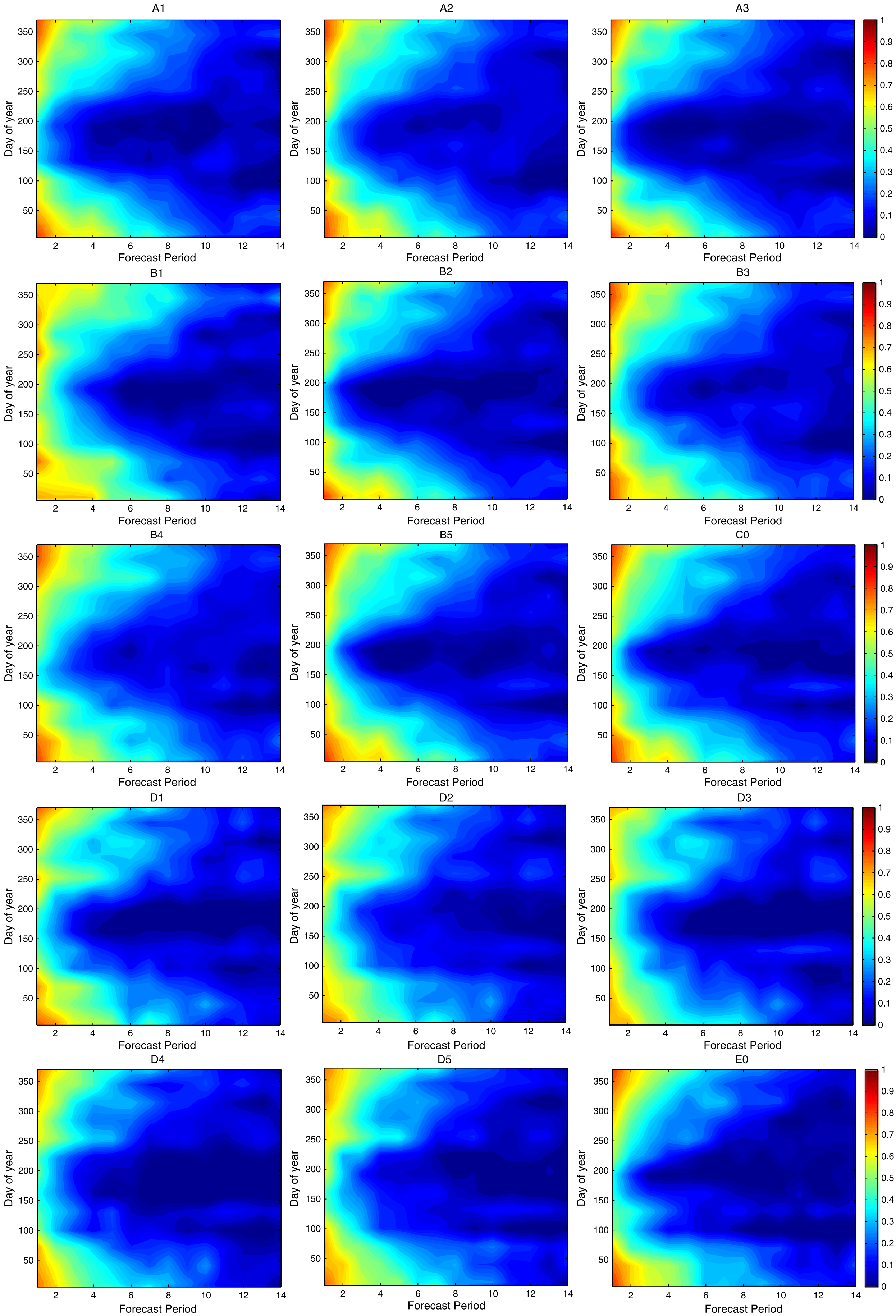

Figure 8. Correlation coefficients between post-processed mean ensemble forecasts of daily values and corresponding observations for different lead times (1–14 days) and the 15 sub-areas

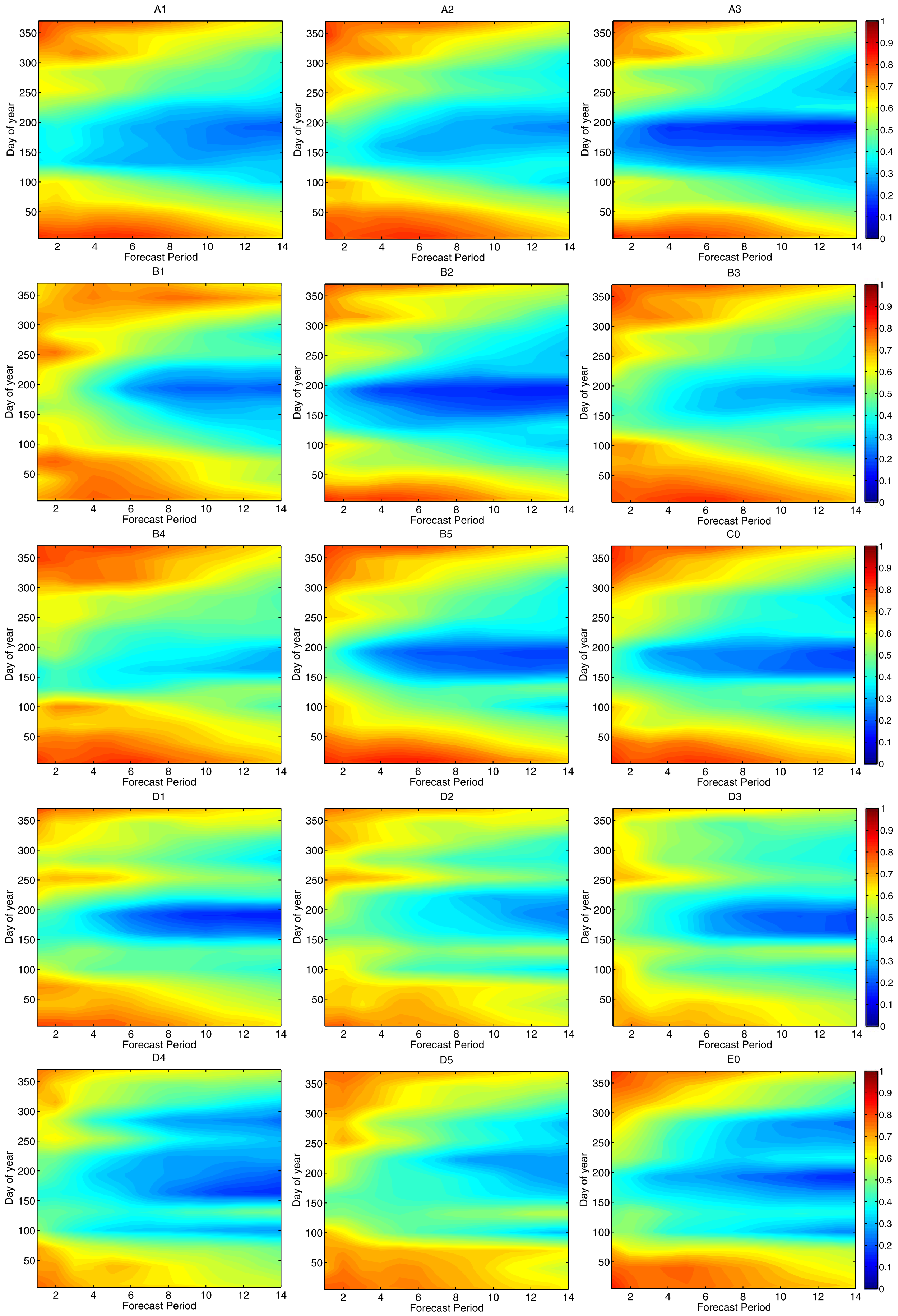

Figure 9. Correlation coefficients between post-processed mean ensemble forecasts of cumulative precipitation and corresponding observation for different lead times (1-14 days) and the 15 sub-areas

Table VI. Basin averaged BSS values of post-processed probability of precipitation (PoP) forecasts of daily precipitation for different lead times (1–14 days) for the Huai river basin

	Lead time (day)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Spring	0.338	0.212	0.133	0.094	0.076	0.052	0.045	0.031	0.021	0.017	0.012	0.010	0.008	0.007
Summer	0.275	0.125	0.060	0.029	0.019	0.009	0.004	0.003	0.003	0.002	0.002	0.002	0.001	0.001
Fall	0.433	0.283	0.200	0.146	0.117	0.101	0.076	0.047	0.032	0.015	0.010	0.010	0.008	0.004
Winter	0.513	0.376	0.272	0.226	0.174	0.125	0.113	0.082	0.058	0.040	0.027	0.022	0.016	0.018
All seasons	0.393	0.252	0.169	0.127	0.098	0.073	0.061	0.042	0.029	0.019	0.013	0.011	0.008	0.008

Table VII. Basin averaged BSS values of post-processed probability of precipitation (PoP) forecasts of cumulative precipitation over different lengths of lead times (1–14 days) for the Huai river basin

	Lead time (day)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Spring	0.338	0.335	0.313	0.291	0.276	0.261	0.247	0.231	0.217	0.202	0.190	0.181	0.168	0.158
Summer	0.275	0.250	0.202	0.160	0.129	0.106	0.090	0.079	0.071	0.065	0.058	0.052	0.048	0.044
Fall	0.433	0.408	0.364	0.325	0.290	0.258	0.234	0.205	0.186	0.167	0.151	0.136	0.119	0.103
Winter	0.513	0.497	0.470	0.458	0.440	0.412	0.384	0.353	0.328	0.310	0.289	0.271	0.247	0.223
All seasons	0.393	0.376	0.341	0.313	0.288	0.264	0.243	0.221	0.204	0.190	0.175	0.163	0.148	0.134

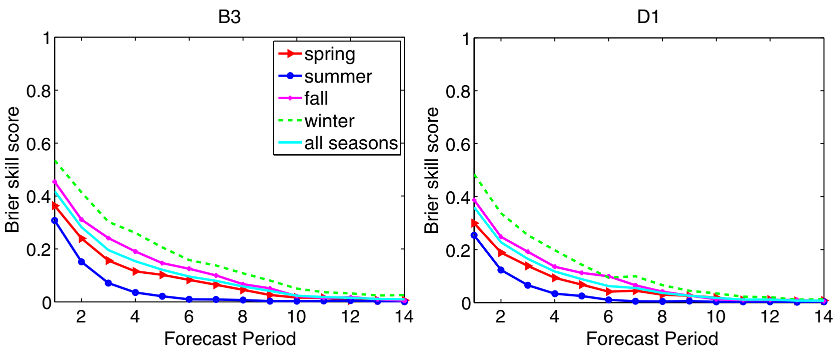

Figure 10. Brier skill scores (BSS) of post-processed ensemble forecasts of daily precipitation over corresponding climatological ensemble forecasts for different lead times (1–14 days) and sub-areas B3 and D1

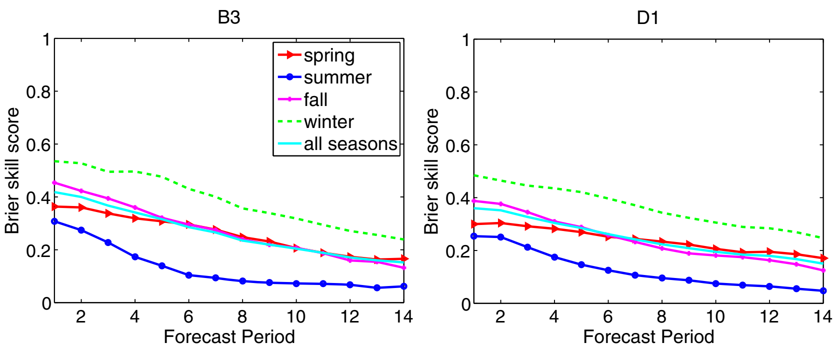

Figure 11. Brier skill scores (BSS) of post-processed ensemble forecasts of cumulative precipitation over corresponding climatological ensemble forecasts for different lead times (1–14 days) and sub-areas B3 and D1

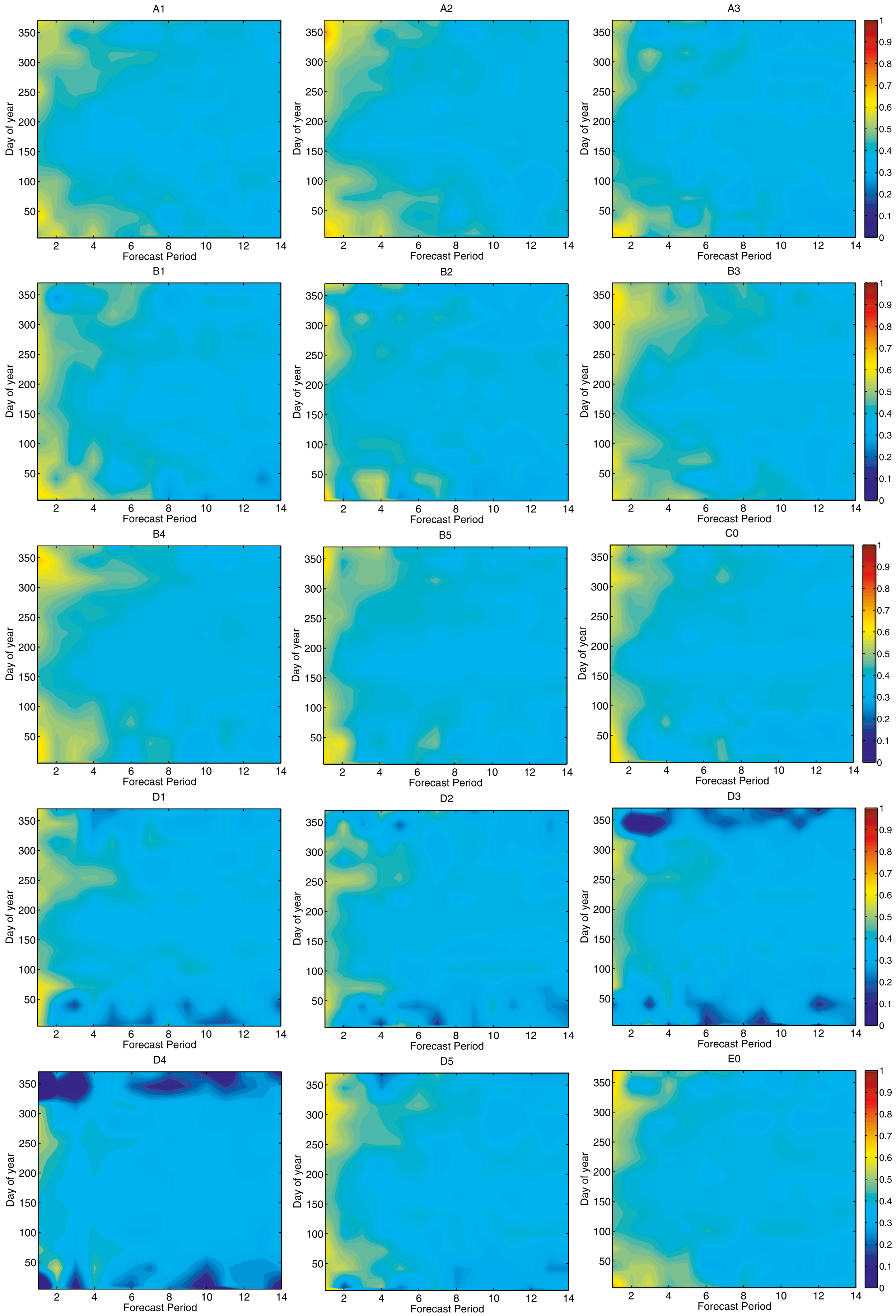

Figure 12. Continuous ranked probability skill score (CRPSS) of post-processed ensemble forecasts of daily values over corresponding climatological ensemble forecasts for different lead times (1–14 days) and the 15 sub-areas

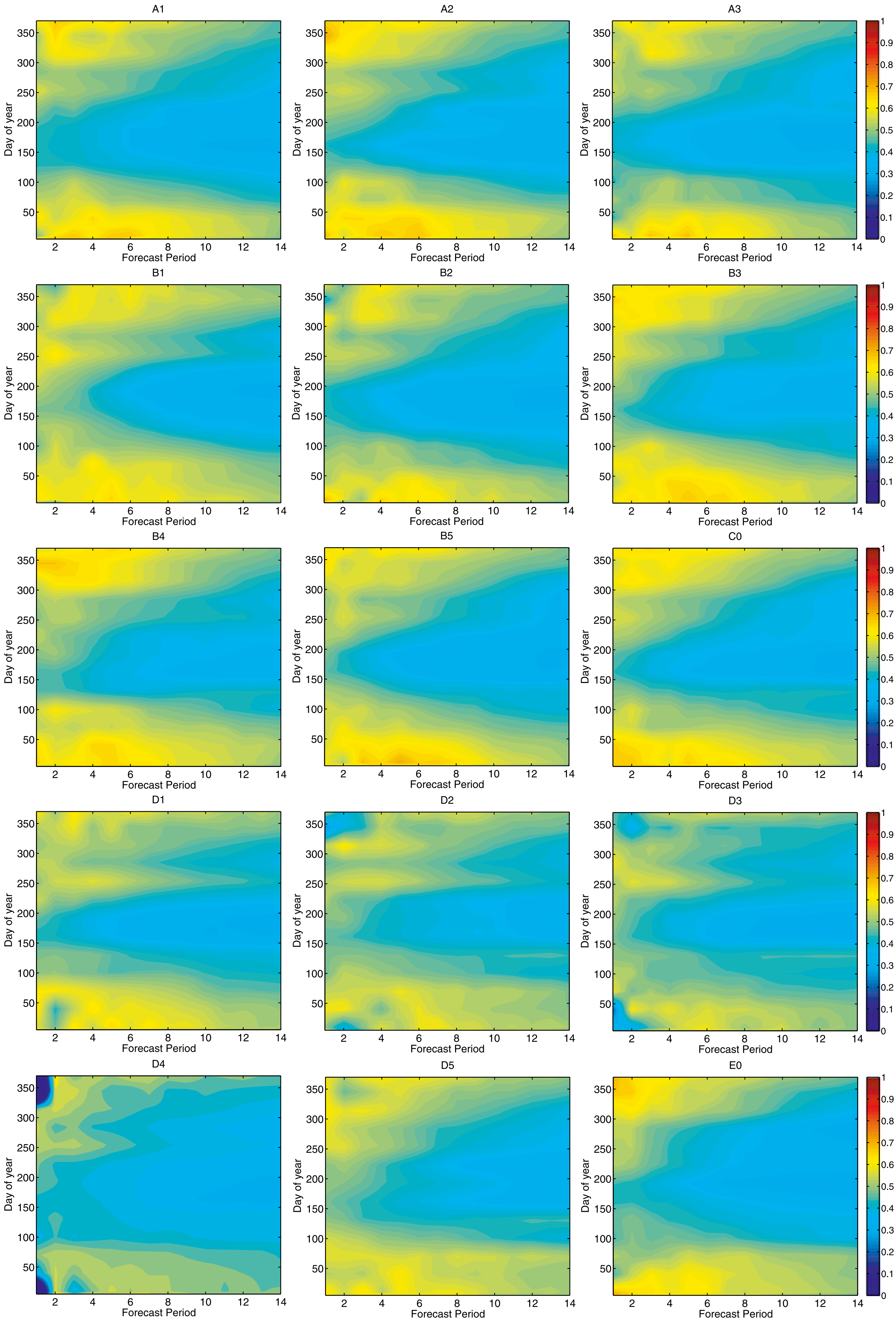

Figure 13. Continuous ranked probability skill score (CRPSS) of post-processed ensemble forecasts of cumulative values over corresponding climatological ensemble forecasts for different lead times (1–14 days) and the 15 sub-areas

seasons, whereas Figure 7 is for cumulative mean ensemble forecasts at different lead times and seasons. In daily mean ensemble forecasts, the RMSES values tend to peak between days 2 and 7 (i.e. the biggest improvement over the reference forecast occurs during this period). On the other hand, the RMSES values increase with lead times for cumulative ensemble forecasts. The RMSES values of the post-processed daily mean ensemble forecasts for spring appear to be the relatively small (between 0.04 and 0.13, or 4% and 13%), whereas the RMSES values for fall and winter are relatively large (up to 0.49 or 49%). For the RMSES values of the mean cumulative ensemble forecasts, the RMSES values are higher than the daily mean ensemble forecasts (up to 0.723 or 72.3%). These observations suggest that post-processed GFS precipitation forecasts do a better job of forecasting the total amount of precipitation than forecasting the exact timing of the precipitation events.

Figures 8 and 9 display the values of correlation coefficients between post-processed ensemble forecast means and observations for daily precipitation and cumulative precipitation for each day of the year (vertical axis), for each lead time (horizontal axis) and for all 15 sub-areas. Orange to red colour in the plots indicates high correlation, whereas blue colour means low correlation. Comparing individual sub-areas, we note that the correlation coefficients illustrate that the skill of ensemble mean forecast are decreasing with lead time. Comparing the two sets of correlation coefficients, the correlation coefficients between post-processed ensemble forecast means of cumulative precipitation and corresponding observations are better than for daily precipitation forecasts. Both sets have higher correlation in the cool seasons than in warm seasons for all sub-areas. It is noteworthy that forecasts of cumulative precipitation have significant skills (i.e. the correlation coefficient values >0.4 or above) for even day 14 for almost all of the sub-areas and almost all seasons, except summer. For forecasts of daily precipitation, however, significant skills are shown mostly for the first few days. In winter, skills of daily forecasts can last up to a week or so. The correlation of post-processed mean ensemble forecasts

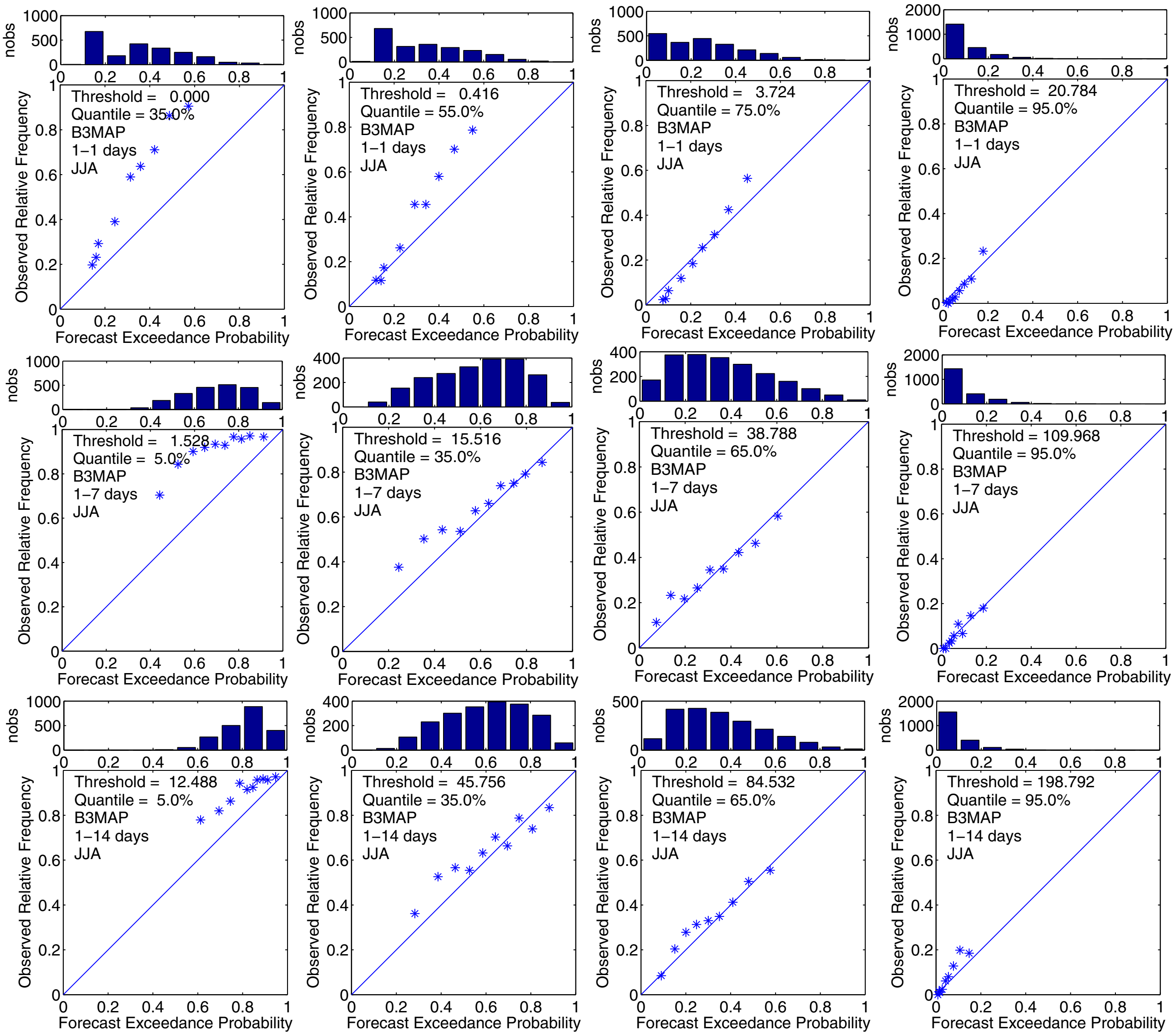

Figure 14. Reliability diagram (RD) of post-processed ensemble forecasts of cumulative precipitation for different lead times (1, 7 and 14 days) for sub-area B3 in summer

with observation does not differ too much from that of the raw mean ensemble forecasts because Schaake's method is designed to make use of the correlation between raw mean ensemble forecast and observation. If the correlation between raw mean ensemble forecast and observation is non-existent, the post-processed mean ensemble forecast will not lead to higher correlation.

Verification of ensemble forecasts

Brier skill score is used to evaluate improvement of the forecasts of probability of precipitation (PoP) relative to a reference forecast for all 15 sub-areas. The reference forecast for the verification of ensemble forecast is the climatological ensemble forecast, which is formed by historical precipitation time series during a given period of the year over a number of years. Tables VI and VII exhibit the BSS values of post-processed PoP forecasts of daily values and cumulative values averaged over the 15 sub-areas of the Huai river basin. The BSS values of post-processed ensemble forecasts are obviously improved over that of the climatological PoP forecasts, i.e. the BSS values are larger than zero for all seasons and lead times. The reason for this improvement is due to the fact that GFS raw mean ensemble PoP forecast has higher correlation with observed PoP than the PoP estimate based on climatology. For the forecast of daily values, the improvement of the post-processed PoP forecasts over the climatological forecasts is obvious only in the first few days, but the improvement diminishes as lead time increases. For forecasts of cumulative precipitation, the improvement in PoP forecasts is significant even at lead time day 14, no matter what season it is. Figures 10 and 11 show the BSS results for the daily ensemble forecasts and ensemble forecasts of cumulative precipitation for sub-area B3 and D1. From the tables as well as the figures, we note that improvement in winter is more apparent than in other seasons. This is probably due to the fact that winter precipitation events tend to be large-scale frontal events that are easier to forecast.

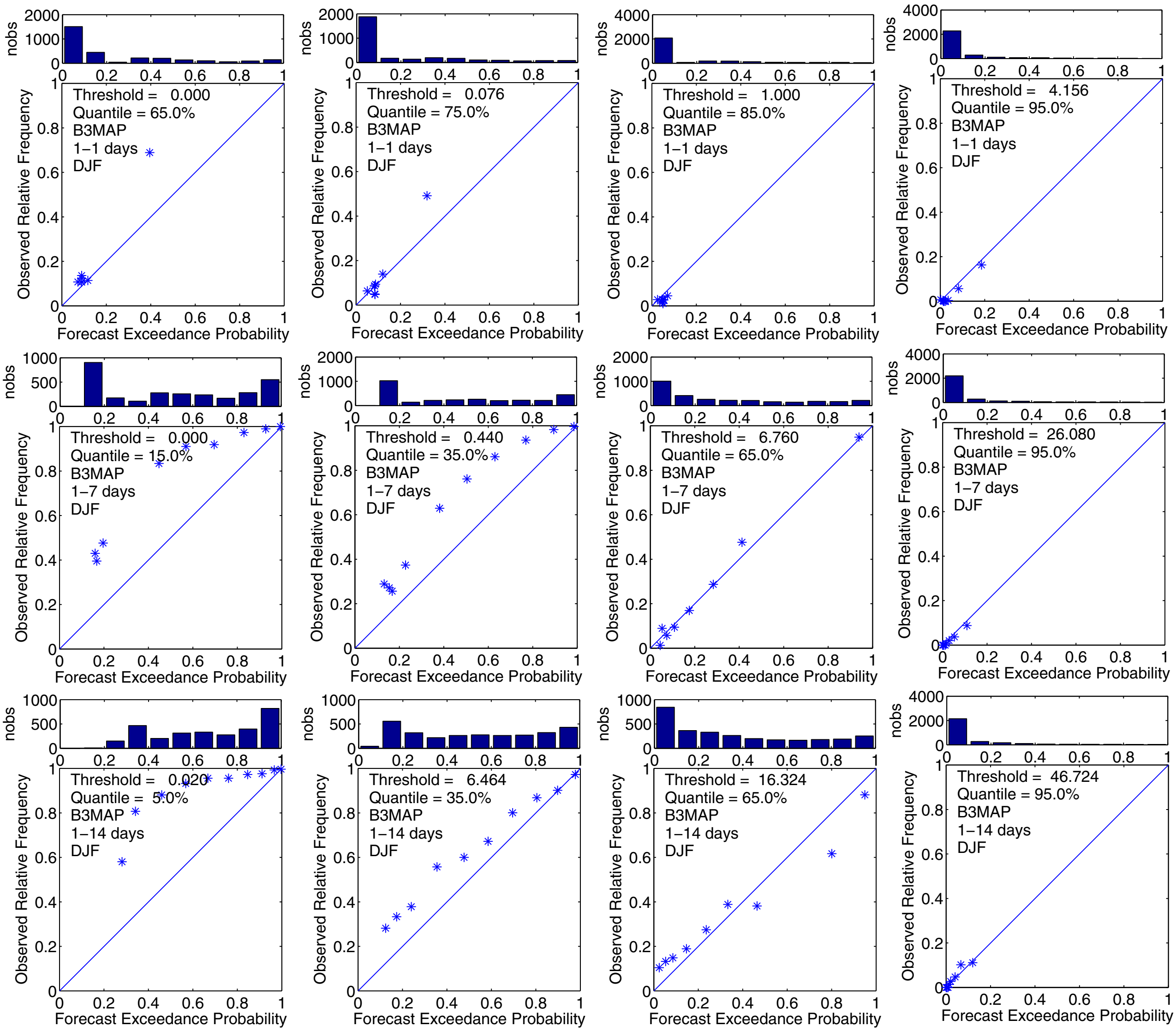

Figure 15. Reliability diagram (RD) of post-processed ensemble forecasts of cumulative precipitation for different lead times (1, 7 and 14 days) for sub-area B3 in winter

Figures 12 and 13 display the CRPSS values for the post-processed ensemble daily precipitation forecasts and ensemble forecasts of cumulative precipitation for different lead times and all sub-areas. All CRPSS values in the figures are shown to be above zero, indicating that the post-processed ensemble precipitation forecasts are better than the climatological forecasts. The CRPSS values for daily precipitation forecasts are smaller than that for the cumulative precipitation forecasts. For the drier sub-areas (D1–D5), the improvement in the winter season is less than that in other seasons, which contrasts to the previously described results, where the improvement in verification statistics in winter is more obvious than in other seasons. For ensemble forecasts of cumulative precipitation, the improvement in CRPSS values is consistent with other verification statistics. For sub-area D4, the CRPSS values for day 1 in winter do not seem to improve. This may indicate that the sample size of precipitation events for the period needs to be larger for the results to be reliable.

Figures 14–17 show the RDs of two typical sub-areas, B3 and D1, for ensemble forecasts of cumulative precipitation in both summer (i.e. June–July–August) and winter (i.e. December–January–February) at three different lead times (1, 7 and 14 days). The RD plots are made with bins of uneven probability intervals on the *x*-axis (i.e. forecast probabilities). This is carried out to ensure that there are adequate samples in each bin. This contrasts the traditional RD plot that uses fixed-length bins (say 0.1), which may not contain enough sample points in some of the bins for meaningful statistical calculation. The reliability of large precipitation events (i.e. with high threshold values) is mostly good (i.e. close to the 45° diagonal line) in all figures, indicating that the post-processed ensemble precipitation forecasts do a reasonable job of forecasting larger precipitation events. For low threshold values, especially in day 1 forecasts, the RD plots tend to show points off the diagonal line, probably because there are too many zero rain or very small rain events on

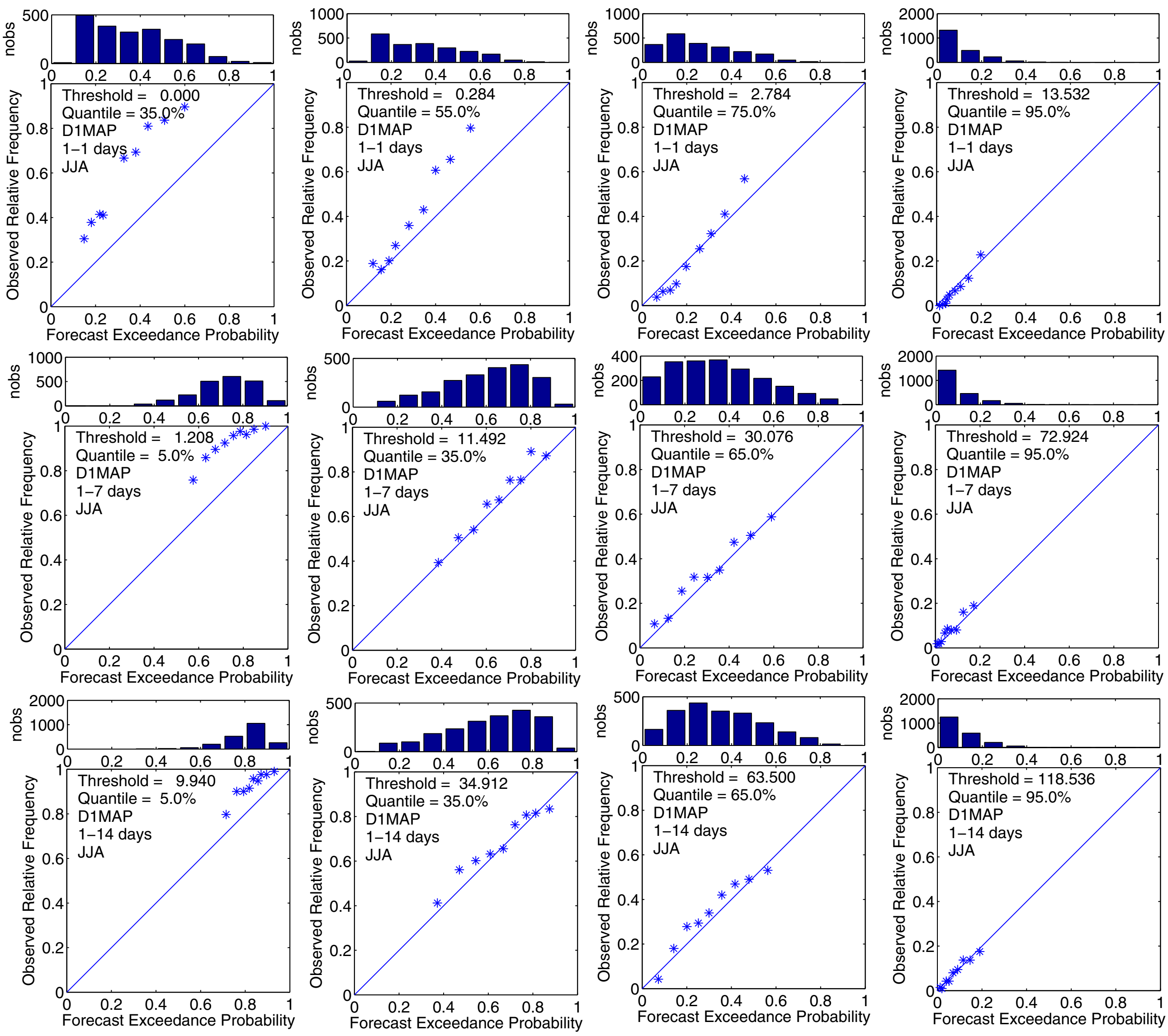

Figure 16. Reliability diagram (RD) of post-processed ensemble forecasts of cumulative precipitation for different lead times (1, 7 and 14 days) for sub-area D1 in summer

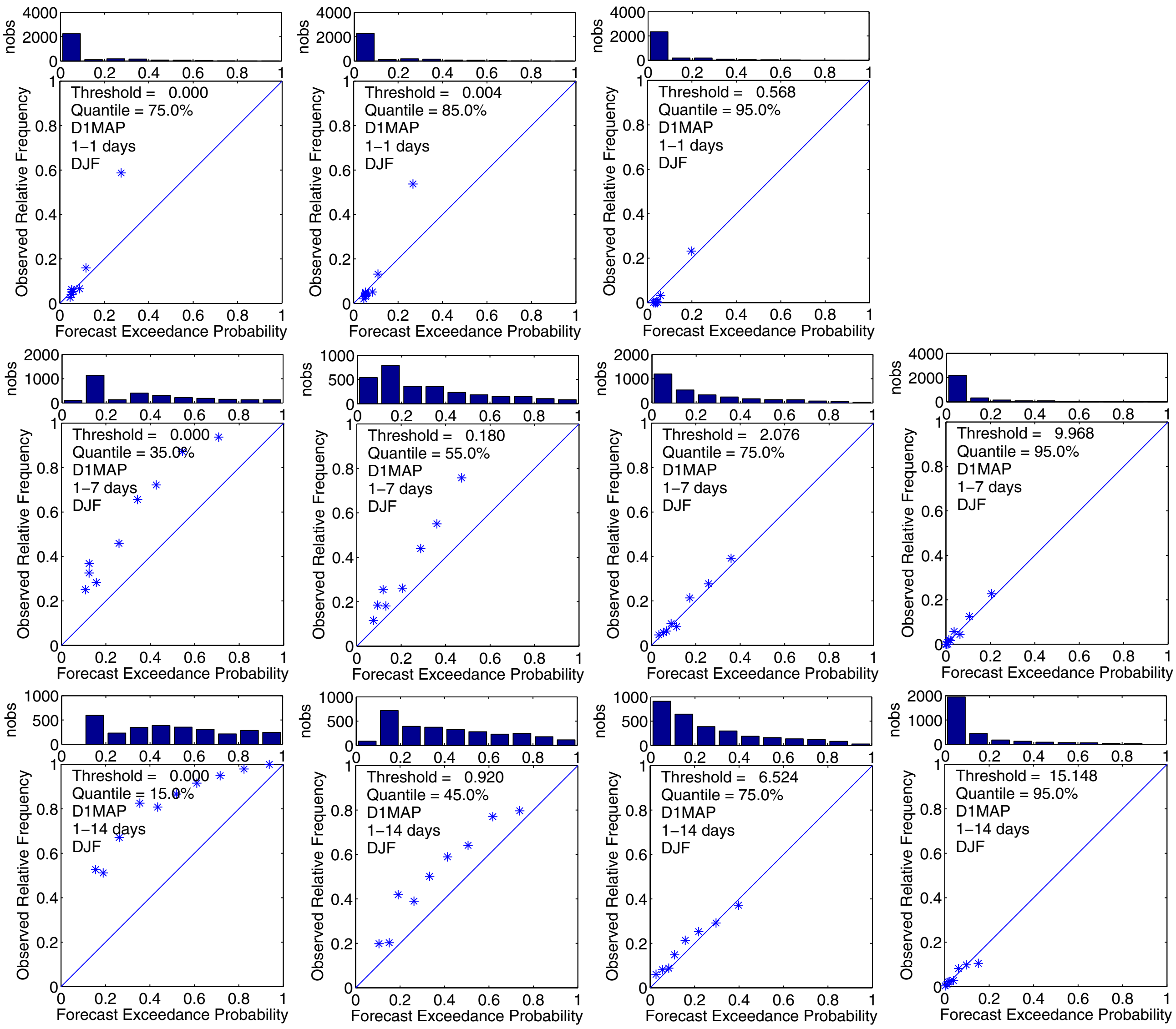

Figure 17. Reliability diagram (RD) of post-processed ensemble forecasts of cumulative precipitation for different lead times (1, 7 and 14 days) for sub-area D1 in winter

any given day and the forecasted probability for small events is not accurate. There are no discernable differences in RD plots between the warm and cool seasons or between the wet and dry sub-areas. The reliability of daily precipitation forecasts (which is not shown here) is somewhat worse than that of the cumulative precipitation forecasts.

SUMMARY AND CONCLUSIONS

This paper investigated the suitability of NCEP's GFS precipitation reforecast product for generating basin scale ensemble precipitation forecasts in the Huai river basin in China. Statistical models of the raw ensemble mean precipitation forecasts and observations are established for 15 sub-areas in Huai river basin by using historical forecast/observation data pairs from the 1981–2003 period. Statistical verification measures are used to evaluate the predictive skills of the post-processed ensemble daily precipitation forecasts and ensemble cumulative precipitation forecasts. The following conclusions are drawn on the basis of the experimental results from this study:

1. The raw mean ensemble precipitation forecasts show significant biases in all sub-areas of the Huai river basin. The biases tend to be larger in summer and fall than in spring and winter. The biases do not seem to be totally dependent on lead times. The biases are removed totally from the post-processed ensemble precipitation forecasts.
2. The post-processed ensemble daily precipitation forecasts possess meaningful skill in the first days, with the value of correlation coefficient greater than 0.4. In winter and spring, the skill can last to a week or so. In terms of post-processed ensemble cumulative precipitation forecasts, the predictive skill is very significant, even for lead time at day 14, except for summer.
3. The ensemble spread of the post-processed ensemble precipitation forecasts is shown to be significantly

better than that of the climatological ensemble precipitation forecasts in terms of both the PoP forecasts and the forecasted precipitation amounts. The RD plots indicate that the post-processed ensemble forecasts can reliably predict large precipitation events (i.e. events above a high threshold values). The reliability of the small event forecasts is not as good, probably because the raw precipitation forecasts are made at $2.5^\circ \times 2.5^\circ$ spatial resolution, making the prediction of the smaller events very hard.

The fact that the raw GFS precipitation reforecast product generated in the US can be used successfully in China to generate basin scale ensemble precipitation forecast points to the potential of the GFS reforecast products being applicable elsewhere in the world. The findings that there are meaningful prediction skills in ensemble daily precipitation forecasts and ensemble cumulative precipitation forecasts also have significant implications in flood forecasting and reservoir operation applications. In follow-up research, we will try to find out how the ensemble precipitation forecasts from this study can be translated into ensemble hydrological forecasts of floods and in reservoir operation optimization.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support provided by the 2010–2013 Public Service Scientific Research Program of the Chinese Meteorological Administration (Project # GYHY201006037) and by the Chinese Ministry of Science and Technology 973 Basic Science Research Program (Project #2010CB428402). We would also like to extend our appreciation of Dr Florian Pappenberger and an anonymous reviewer whose insightful comments and suggestions lead to a much more readable paper.

REFERENCES

Bougeault P, Toth Z, Bishop C, Brown B, Burridge D, Chen D, Ebert B, Fuentes M, Hamill TM, Mylne K, Nicolau J, Paccagnella T, Park Y-Y, Parsons D, Raoult B, Schuster D, Dias PS, Swinbank R, Takeuchi Y, Tennant W, Wilson L, Worley S. 2010. The THORPEX Interactive Grand Global Ensemble (TIGGE). *Bulletin of the American Meteorological Society* **91**: 1059–1072.

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. *Journal of Hydrometeorology* **5**(1): 243–262.

Cloke HL, Pappenberger F. 2009. Ensemble flood forecasting: a review. *Journal of Hydrology* **375**(3–4): 613–626.

Froude LSR. 2010. TIGGE: comparison of the prediction of northern hemisphere extratropical cyclones by different ensemble prediction systems. *Weather and Forecasting* **25**: 819–836.

Fundel F, Zappa M. 2011. Hydrological ensemble forecasting in mesoscale catchments: sensitivity to initial conditions and value of reforecasts. *Water Resources Research* **47**: W09520. DOI: 10.1029/2010WR009996

Glahn HR, Lowry DA. 1972. The use of Model Output Statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology* **11**(8): 1203–1211.

Hamill T, Whitaker JS, Wei X. 2004. Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review* **132**: 1434–1447.

Hamill TM, Hagedorn R, Whitaker JS. 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Monthly Weather Review* **136**: 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. *Atmospheric Science Letters* **11**: 132–138.

Huai River Commission. 2009. Huai river water resources bulletin for year 2009. The Chinese Ministry of Water Resources, 29p. (in Chinese).

Krzysztofowicz R. 1997. Transformation and normalization of variates with specified distributions. *Journal of Hydrology* **197**(1–4): 286–292.

Krzysztofowicz R, Sigrest AA. 1999. Calibration of probabilistic quantitative precipitation forecasts. *Weather and Forecasting* **14**(3): 427–442.

Lewis JM. 2005. Roots of ensemble forecasting. *Monthly Weather Review* **133**(7): 1865–1885.

Lorenz EN. 1963. Deterministic non-periodic flow. *Journal of Atmospheric Science* **20**: 130–141.

Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of the Royal Meteorological Society* **122**: 73–119.

Pappenberger F, Thielen J, del Medico M. 2011. The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European flood alert system. *Hydrological Processes* **25**(7). DOI: 10.1002/hyp.7772, 2010

Park Y-Y, Buizza R, Leutbecher M. 2008. TIGGE: preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society* **134**: 2029–2050.

Rayner S, Lach D, Ingram H. 2005: Weather forecasts are for wimps: why water resource managers do not use climate forecasts. *Climatic Change* **69**: 197–227.

Schaake J, Demargne J, Hartman R, Mullusky M, Welles E, Wu L, Herr H, Fan X, Seo DJ. 2007. Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrology and Earth System Sciences Discussions* **4**: 655–717.

Schaake J, Pailleux J, Thielen J, Arritt R, Hamill T, Luo L, Martin E, McCollor D, Pappenberger F. 2010. Summary of recommendations of the first workshop on Postprocessing and Downscaling Atmospheric Forecasts for Hydrologic Applications held at Météo-France, Toulouse, France, 15–18 June 2009. *Atmospheric Science Letters* **11**: 59–63.

Song Q, Wei F, Xu C. 2006. Numerical simulation and diagnostic analysis of a heavy rainfall process over the Huaihe river valley. *Journal of Nanjing Institute of Meteorology* **29**: 342–347 (in Chinese).

Thiemig V, Pappenberger F, Thielen J, Gadain H, de Roo A, del Medico M, Muthusi F. 2010. Ensemble flood forecasting in Africa: a feasibility study in the Juba-Shabelle river basin. *Atmospheric Science Letters* **11**(2): 123–131.

Toth Z, Kalnay E. 1997. Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review* **125**: 3297–3319.

Wang S, Tian H, Ding X, Xie W, Tao Y. 2009. Climate characteristics of precipitation and phenomenon of drought-flood abrupt alternation during main flood season in Huaihe river basin. *Chinese Journal of Agrometeorology* **30**: 31–34 (in Chinese).

Webster PJ, Jian J, Hopson TM, Hoyos CD, Agudelo PA, Chang HR, Curry JA, Grossman RL, Palmer TN, Subbiah AR. 2010. Extended-range probabilistic forecasts of Ganges and Brahmaputra floods in Bangladesh. *Bulletin of the American Meterological Society* **91**(11): 1493–U121.

Wu L, Seo DJ, Demargne J, Brown JD, Cong S, Schaake J. 2011. Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *Journal of Hydrology* **399**: 281–298.

Xu G, Zhao S, Wang Y, Yang Y. 2007. Experiment of ensemble forecast of heavy rainfall in the Huaihe river during rainy season of 2003. *Climatic and Environmental Research* **12**: 481–488 (in Chinese).

Zhou Y, Wang L, Xu Y, Zhang J. 2001. An approach to dynamics of the flood series in the Huaihe river basin. *Scientia Geographica Sinica* **20**: 41–45 (in Chinese).

Zhu H, Wang D, Zhu P, Zhou K. 2007. Application of GRAPES model to area-rainfall forecast in the Huaihe river basin. *Meteorological Monthly* **33**: 76–82 (in Chinese).

APPENDIX: STATISTICAL VERIFICATION MEASURE OF ENSEMBLE FORECASTS

For verification measures presented in the following text, we denote x_i, y_i, p_i, o_i and N as single-valued forecast, corresponding observation, forecast probability, observed frequency and number of forecast/observation pairs, respectively, at time i.

A.1 Pearson correlation coefficient

Pearson correlation coefficient, r, measures the degree of association between x_i and y_i. It is computed as

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

where $\bar{x}$ is forecast average and $\bar{y}$ is the observation average. For a perfect forecast, $r = 1$.

A.2 Bias

Bias measures the relative difference between the average forecast and observation over a long time period. It is computed as

$$\text{Bias} = \bar{x} - \bar{y}$$

where $\bar{x}$ is the forecast average and $\bar{y}$ is the observation average.

A.3 Root mean square error skill score

Root mean square error (RMSE) measures the closeness of forecast and observation over a long period. RMSE is computed as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)^2}$$

The perfect RMSE score is 0. Because the squared difference is used in the calculation, RMSE gives more consideration to high values than low ones.

Because RMSE is not meaningful when viewed in absolute terms, RMSE skill score (RMSES) is used in the analysis. It is defined as the percentage improvement of RMSE over a reference (denoted as RMSE*), i.e.

$$\text{RMSES} = \left(1 - \frac{\text{RMSE}}{\text{RMSE}^*}\right)$$

The reference used in this analysis is the RMSE value of the raw mean ensemble forecast.

A.4 Brier skill score

The Brier skill score (BSS) measures the improvement of the probabilistic forecast relative to a reference forecast:

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{ref}}}$$

and

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N} (p_i - o_i)^2$$

Usually, the reference forecast is the climatological probability forecast, $\text{BS}_{\text{ref}} = s(1 - s)$, where s is the climatological probability forecast. The perfect BSS is 1, and 0 indicates no skill in the forecast.

A.5 Continuous ranked probability skill score

Continuous ranked probability score (CRPS) is a measure of the integrated squared difference between the cumulative distribution function of the forecasts and the corresponding cumulative distribution function of the observations. The mathematical expression of CRPS is

$$\text{CRPS} = \int_{-\infty}^{\infty} (p_i(x) - o_i(x))^2 \mathrm{d}x$$

Again, CRPS of a perfect forecast is equal to 0. Continuous ranked probability skill score (CRPSS) measures the relative improvement of a forecast over a reference forecast:

$$\text{CRPSS} = 1 - \frac{\text{CRPS}}{\text{CRPS}*}$$

CRPS* referred here corresponds to the CRPS value of the climatological ensemble forecast.

A.6 Reliability diagram

Reliability diagram is another measure of the closeness between forecast probability and observed frequency. It plots the average forecast probability within each bin on the x-axis.

$$f_i = \frac{1}{N_k}\sum_{i \in I_k} F_{y_i}(t)$$

Y-axis shows the corresponding fraction of observations that fall in the corresponding bin.

$$p_{f_i}(x = 1|f_i) = \frac{1}{N_k}\sum_{i \in I_k} p(t \geq x_i)$$

where the t is a real-valued threshold, $F_{y_i}(t)$ is the probability of forecast value exceeding threshold t for sample i, N_k is the number of sample fall in the kth forecast bin and the I_k denotes index for the sample fall in the kth forecast bin.