Performance of post-processed methods in hydrological predictions evaluated by deterministic and probabilistic criteria

Xiang-quan Li[a] Jie Chen[a,*] Chong-Yu Xu[a,b] Lu Li[c], and Hua Chen[a]

[a] *State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China*

[b] *Department of Geosciences, University of Oslo, Oslo N-0316, Norway*

[c] *Uni Research Climate, Bjerknes Centre for Climate Research, Bergen, Norway*

[*] Correspondence to:
Jie Chen, State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China.
E-mail: jiechen@whu.edu.cn

Abstract

Meteorological Ensemble Streamflow Prediction (ESP), which uses Ensemble Weather forecasts (EWFs) to drive hydrological models, is a useful methodology for extending forecast periods and to provide valuable uncertainty information to improve the operation of future water resources. However, raw EWFs are usually biased and under-dispersive and so cannot be directly used in ESP, leading to the development of several post-processing methods. The performance of these methods needs to be evaluated/compared in building ESP based on deterministic and probabilistic criteria. In addition, likely influencing factors also need to be identified. This study evaluated the performance of four state-of-the-art methods: the Generator-based Post-Processing (GPP) method, Extended Logistic Regression (ExLR), Bayesian Model Averaging (BMA) and Affine Kernel Dressing (AKD), using a simple bias correction (BC) method as a benchmark. The evaluation was carried out over four watersheds with different basin areas in the humid region of central-south China based on the weather reforecasts from the Global Ensemble Forecasting System (GEFS). The results show that the performance of the post-processing methods varies with the forecast variable (precipitation, or air temperature or streamflow), but all of them outperform the BC and GEFS. For the four post-processing methods, the advantage of the generator-based methods (GPP and ExLR) lies in their probabilistic performance, which outperforms the distribution-based methods (BMA and AKD) by about 10% in precipitation forecasts and about 20% in streamflow forecasts, while the distribution-based methods (BMA and AKD) are better at their deterministic performance for precipitation forecasts, with a benefit of about 15%. Meanwhile, the post-processing methods generally perform better for precipitation and streamflow forecasts, but worse for air temperature forecasts for a bigger basin compared to the distribution-based methods. The results of this study emphasize the importance of considering the uncertainty of post-processing methods in ESP.

Key Words

Ensemble Streamflow Prediction (ESP); Ensemble Weather Forecast (EWF); Post-Processing Method; Deterministic Criteria; Probabilistic Criteria

1. Introduction

The daily operations of water resource management rely extensively on hydrological forecasting of the highest possible accuracy and with the longest possible forecast period. Assessing the risk of water resource management also requires knowledge about the possible uncertainties in hydrological forecasting information (Roulin 2007; Xu and Tung, 2008; Shukla et al., 2012). One viable way to meet these requirements is to build meteorological Ensemble Streamflow Prediction systems (ESPs) by driving hydrological models using Ensemble Weather forecasts (EWFs). Many studies have shown that using meteorological ESPs can achieve longer forecast periods and derive more reliable probabilistic forecasts. For example, Cloke and Pappenberger (2009) reviewed the use of ESPs in flood forecasting and found the “added value” of ESPs over deterministic forecasts, including improved flood forecasting accuracy, greater probabilistic skill, reduced forecast uncertainty. Alfieri et al. (2014) evaluated the European operational ESP system (European Flood Awareness System (EFAS)) for flood awareness and found that 10-day ESPs in medium to large rivers are considered as skillful when comparing with the reference simulation using the observed meteorological fields.

However, building a meteorological ESP system is not as easy as driving hydrological models by directly using EWFs. On the one hand, raw EWFs are biased and under-dispersive when compared to the observations (Hagedorn et al. 2008; Hamill et al. 2008; Scheuerer and Hamill 2015). On the other hand, the spatial resolution of EWFs is generally too coarse to drive hydrological models for ESPs (Kavetski et al., 2006a, 2016b; Vetter et al. 2016).

Statistical post-processing methods have been used over the past few years to reduce the bias and to reconstruct the proper ensemble spread for EWFs. These methods can be divided into two categories, distribution-based and generator-based methods.

Distribution-based methods need to calibrate the probability distribution function (PDF) of the weather variable based on raw EWFs, thereby allowing the post-processed forecast ensemble to be generated by randomly sampling the calibrated PDF. These methods include Bayesian Model Averaging (BMA) (Raftery et al., 2005; Sloughter et al., 2007), ensemble dressing (Bröcker and Smith, 2008), and Non-Gaussian Regression (NGR) (Hagedorn et al., 2008; Baran and Nemoda, 2016).

Generator-based methods generate the forecast ensemble by conditionally resampling the historical observations using the forecast information from raw EWFs. These methods include Modified Extended Logistic Regression (ExLR, Roulin and Vannitsem, 2012) and the Generated based Post-Processing method (GPP, Chen et al., 2014).

In meteorological forecasting, the promising post-processing methods were found to be ensemble dressing, logistic regression, and BMA (Wilks, 2006; Wilks and Hamill, 2007; Schmeits and Kok, 2010). However, studies have shown that it is generally not sufficient to only use probabilistic criteria. For example, Vannitsem and Hagedorn (2011) compared Error-in-Variable Model Output Statistics (EVMOS) and the probabilistic-like method NGR in terms of their improvement of ECMWF temperature forecasts. The results showed that EVMOS is comparable to NGR in generating the ensemble consistent with the observations, and is even better than NGR at predicting extreme events. Therefore, to determine the “best” method for weather forecasting, both deterministic and probabilistic criteria must be considered.

When post-processed EWFs are further used to drive a hydrological model to generate ensemble streamflow forecasts, there are many factors that influence the performance of the streamflow forecasts, such as the propagation of bias, and the uncertainty from meteorological input data or from the hydrological model. Verkade et al. (2013) investigated how the biases in mean, spread, and forecast probabilities are propagated to streamflow ensemble forecasts. The temperature and precipitation reforecasts from the European Centers for Medium-Range Weather Forecasts (ECMWF) were post-processed by the quantile-to-quantile transform with linear regression and with logistic regression. The post-processed ensemble forecasts were then used to drive the hydrological model for several basins with different spatial scales. It was found that the significant biases of the raw ensemble forecasts would be largely propagated to the streamflow ensembles, and the improvements to streamflow accrued by post-processing were generally modest. Siddique and Mejia (2017) built the regional hydrological ensemble prediction system (RHEPS) for short- to medium-range (6-168 h) forecasting over the U.S. mid-Atlantic region (MAR), combining the EWF from Global Ensemble Weather Forecasting System (GEFS), the statistics output from the post-processing model, and a distributed hydrological model. Their results found that the hydrological uncertainty was dominant for 1-3 days, while the uncertainty of meteorological input was more notable after 3 days.

Although it is widely accepted that a certain post-processing method is needed to bridge EWFs and ESPs, it is not clear how the various post-processing methods perform, or how to select the most reliable methods to building ESPs. When considering meteorological forecasts, the performances of different post-processing methods vary in terms of deterministic or probabilistic metrics. For streamflow forecasts, the different performance of post-processed EWFs may be further amplified and new influencing factors in hydrological modeling may also be introduced. Therefore, an evaluation of the commonly used post-processing methods can provide meaningful results for choosing appropriate methods in building ESPs. Accordingly, the objectives of this study can be specified by answering the following two questions: (1) how do the commonly used post-processing methods perform in ESPs? And (2) what is the main influencing factor for the performance of post-processing methods?

2. Study area and dataset

2.1 Study area

The ESPs created by four different post-processing methods were evaluated/compared in four basins of different sizes located in central-south China: Daxitan (#1: 3,312 km^2), Xiangxiang (#2: 6,053 km^2), Ganxi (#3: 9,972 km^2) and Hengyang (#4: 52,150 km^2). These basins were selected to evaluate the impact of basin characteristics, especially basin size, on the performance of ESPs. These four basins are from the same watershed, the Xiangjiang watershed (see Supplementary Material 1), which guarantees they share similar meteorological and hydrological conditions.

2.2 Dataset

The dataset consists of both observations and EWFs. The observations cover 36 years (from 1979 to 2014, for basin #1, #2 and #4) or 31 years (from 1979 to 2009, for basin #3) of daily basin-averaged meteorological data and discharge data. The meteorological variables include daily mean air temperature, precipitation, and potential evaporation. All of the meteorological and hydrological data were quality controlled by the China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn) and the Hydrology and Water Resources Bureau of Hunan Province (http://www.hnwr.gov.cn/).

The EWFs for precipitation and mean air temperature used in this study were taken from the second version of the Global Ensemble Forecast System (GEFS) reforecasts (http://portal.nersc.gov/project/refcst/v2/), which provide 11-member ensemble forecasts for up to 16 days from December 1984 to the present.

The common period for observations and EWFs is 1985-2014 for basins #1, #2 and #4, and 1985-2009 for basin #3. According to previous studies, precipitation forecasts lose their skill when the lead time is over 7 days (e.g. Liu and Coulibaly, 2011; Chen et al., 2014), and so only reforecasts for up to one week periods are utilized in this study.

3. Methodology

3.1 Post-processing methods

Let y denote the weather variable of interest, like precipitation or air temperature, and x_1, …, x_K represent the ensemble weather forecasts with K members. The parametric PDF denoted by g for the weather variables then takes the form:

$$y \mid x_1, \ldots, x_M \sim g(y \mid x_1, \ldots, x_K) \qquad (1)$$

The distribution type of g depends on the type of the variable. The air temperature is usually represented by a two-parameter normal distribution, while the precipitation is characterized by a mixed discrete/continuous distribution with a positive probability of being zero and a continuous skewed distribution for positive precipitation amounts. The forecast uncertainty for precipitation has been found to be generally higher for larger precipitation amounts and infrequent high-precipitation events (Scheuerer and Hamill, 2015). Sloughter et al. (2007) proposed a mixed distribution model for precipitation in the following form.

$$g(y \mid f_k) = P(y=0 \mid f_k) \cdot I(y=0) + P(y>0 \mid f_k) \cdot g_k(y \mid f_k) \cdot I(y>0) \qquad (2)$$

where $g(y|f_k)$ is the probability distribution given the member forecast f_k. $I[\ldots]$ is unity if the condition in brackets holds, and zero otherwise; $P(y=0|f_k)$ and $P(y>0|f_k)$ are the probabilities of non-precipitation and precipitation given the forecast f_k, respectively; and $g_k(y|f_k)$ is a two-parameter gamma distribution.

The distribution-based method and the generator-based method differ in how the PDF for the weather variable is calibrated and in how the post-processed ensemble weather forecasts are generated.

For the distribution-based methods, like BMA (Raftery et al., Sloughter et al., 2007) and AKD (Bröcker and Smith, 2008), the PDF of the weather variable at the given day or period is calibrated by fitting the forecast ensemble based on a historical training set containing EWFs and observations. BMA and AKD only differ in the model specification (Equation (3) for BMA, and Equation (4) for AKD). The post-processed ensemble weather forecasts at the given day or period are then generated by random sampling.

$$p(y \mid f_1,\ldots,f_K)=\sum_{k=1}^{K} W_k g_k(y \mid f_k) \qquad (3)$$

$$p(y \mid f_1,\ldots,f_K)=\frac{1}{K\sigma}\sum_{k=1}^{K} N\left(\frac{y-z_k}{\sigma}\right) \qquad (4)$$

where the weight W_k is the posterior probability of ensemble member k being selected, and reflects the model's relative contribution to predictive skill over the training period; N(…) is the kernel distribution (normal distribution used in this study) with the mean Z_k and deviation σ linked to the ensemble forecasts during the training period.

For the generator-based methods, like GPP (Chen et al., 2014) and ExLR (Wilks, 2009 and Roulin and Vannitsem, 2012), the PDF in different seasons or magnitude is separately calibrated by fitting the corresponding observations. The post-processed ensemble weather forecasts are conditionally resampled from the built PDF according to the forecast information in EWFs.

A simple bias correction method proposed by Chen et al. (2014) is used as the benchmark method. Generally, a linear correction equation with form $y = ax$ (where a is the correction parameters) is used for precipitation, and $y = ax + b$ (where a and b are the correction parameters) is used for air temperature.

Given the 36-year (basins #1, #2 and #4) or 31-year (basin #3) available GEFS forecasts and observations, a cross-validation method is used to implement the BC and post-processing methods. Specifically, when making forecasts for a particular year, the remaining years are used as the training period for calibrating the post-processing method. The BC method is applied to each ensemble member and does not alter the ensemble size. However, the ensemble size can be arbitrarily enlarged by sampling or resampling when using post-processing methods. To better represent the calibrated PDF, the ensemble size for the post-processing methods is set as 1000.

3.2 Hydrological model

The Xin'anjiang (XAJ) model (see Supplementary Material 2) is by far the most popular hydrological model for hydrological simulation and forecasting in China, especially in semi-humid and humid regions. It is a lumped model that requires the daily areal precipitation and the measured daily pan evaporation (or the measured daily mean air temperature) as input. The output is the simulated discharge for the basin outlet. The results (see Supplementary Material 2) show that the model's performance for both the calibration and validation periods are satisfactory and so the hydrological model can be used in building ESPs.

3.3 Verification metrics

Both deterministic and probabilistic metrics from the Ensemble Verification System (EVS) by Brown et al. (2010) are used to evaluate the performance of EWFs and ESPs. The Mean Absolute Error (MAE) and Continuous Ranked Probability Skill Score (CRPSS) are used for assessing deterministic performance and probabilistic performance, respectively.

The proposed ESP is verified by two deterministic metrics, the Nash–Sutcliffe coefficient (NSE) and the Relative Error (RE), and two probabilistic metrics, the Brier

Skill Score (BSS) and the CRPSS. The plot of the reliability diagram is used for a probabilistic diagnosis. Detailed descriptions of these metrics are shown in Supplementary Material 3.

4. Results

4.1 Performances of post-processed EWFs

Figure 1 presents the performance of ensemble mean forecasts measured in MAE by the GEFS, BC, GPP, ExLR, BMA, and AKD methods. The MAE of the GEFS ensemble mean precipitation forecasts ranges from around 3.2 to 4.1 mm for 1 lead day to around 4.1 to 5.0 mm for 3 lead days. It shows a clear increase with the increase of lead days, with an overall increase of 22% when moving from a 1day to a 3-day lead time. The MAE for the GEFS air temperature has a range of about 1.4-2.0 ^{o}C for 1 lead day and of about 1.7-2.4 ^{o}C for 3 lead days, which shows an increase of about 17% from 1-day to 3-day lead times on average. The results for the BC method indicate that it works better for air temperature forecasts than for precipitation and that for precipitation it is especially not desirable for the small basin (Daxitan (basin #1)). Four post-processing methods gain in skill compared to the GEFS forecasts for both precipitation and air temperature. Specifically, the BMA and AKD methods stand out in terms of their accuracy (MAE) in forecasting precipitation, indicating a decrease of 0.8 mm and 1.2 mm in the MAE for 1 and 3 lead days, respectively. Their ranges of MAE are also narrower than those of the GEFS, followed by those of the GPP and the ExLR. The GPP's MAE decreased about 0.5 mm and 0.7 mm for 1 and 3 lead days, respectively, and for ExLR it decreased by about 0.4 mm and 0.6 mm for 1 lead and 3 lead days, respectively. Generally, the MAE of the distribution-based methods (BMA and AKD) is smaller than that of the generator-based methods (GPP and ExLR) by about 15%. For air temperature, four post-processing methods have similar performances with their MAE decreasing by about 0.4 ^{o}C for 1 lead day and for 3 lead days. Their MAE ranges are also narrower than that of the GEFS. Furthermore, according to the MAE values, the post-processing methods perform better for precipitation in large river basins than in small basins, while this is the opposite for air temperature.

Figure 2 presents the probabilistic performance (forecast skill) of ensemble forecasts measured in CRPSS. The graphic composition is similar to that of Figure 1. The GEFS ensemble forecasts present an inferior CRPSS performance, with their CRPSS for air temperature (about 0.8 for 1 lead day and 0.77 for 3 lead days on average) substantially higher than that of precipitation (about 0.44 for 1 lead day and 0.35 for 3 lead days on average). The simple BC method shows little improvement compared to the GEFS, with CRPSS of 0.05 for precipitation and 0.07 for air temperature. Using a BC method is not enough to improve the forecast skill and thus certain post-processing methods are needed. Four post-processing methods are capable of improving the forecast skill of the ensemble forecasts, especially for precipitation. Specifically, for precipitation, four post-processing methods increase the CRPSS compared to the simple BC method as well as decrease the performance difference among different regions. The performance ranking for the four post-processing methods is GPP (increased by 0.15 and 0.18 for 1 and 3 days), ExLR (0.14 and 0.17), BMA (0.11 and

0.13), and AKD (0.09 and 0.11). The CRPSS of the generator-based methods is higher than that of the distribution-based method by about 10%. For air temperature, the GPP, ExLR, and BMA present a slight improvement compared to BC, but the AKD shows no competitiveness over BC. Furthermore, for the largest basin used in this study (Hengyang (basin #4)), the CRPSS metric is the highest for precipitation but the lowest for air temperature, for all six methods.

Figure 3(1) depicts the variation of CRPSS against the lead day for ensemble precipitation forecasts. These results were obtained by averaging the CRPSS over all four basins. The results clearly show a marked decreasing trend for the GEFS, BC and the four post-processing methods against the lead day. The CRPSS of the GPP and ExLR methods are consistently larger than those of the BMA and AKD methods for all lead times, with all four methods all outperforming BC and GEFS. The BC method shows an effective lead time where its performance is inferior to that of the GEFS. Figure 3 (Subplots 2-5) displays the significant lead time of BC in more detail. The effective lead days of BC are generally earlier for a smaller basin, with Daxitan (basin #1) indicating 2 lead days, Xiangxiang (basin #2) 5 lead days, Ganxi (basin #3) 4 lead days and Hengyang (basin #4) 7 lead days. In conclusion, the performance of BC is unstable for different basins as well as for different lead days.

4.2 Performance of the proposed ESPs

Figure 4 gives the deterministic performance of ensemble streamflow forecasts using two deterministic metrics, RE and NSE, over 1 and 3 lead days. To highlight the useful information, those basins with NSE values less than 0 and RE out of the range of -20%-20% are not shown in this figure. The GEFS tends to have a large RE (close to 11% for 1 lead day and 17% for 3 lead days on average) and a small NSE (0.18 for 1 lead day and less than 0 for 3 lead days), indicating that it is inadvisable to directly use the GEFS forecast as the hydrological model input and that certain post-processing methods are needed. The performances of the BC method in generating the streamflow forecasts are highly influenced by factors such as basin size and lead days. The positive performance by BC over GEFS is only achieved for bigger basins like the Ganxi (basin #3) and the Hengyang (basin #4). The four post-processing methods, however, are capable of improving the EWF performance consistently compared to both the BC and the GEFS. In general, the post-processing methods slightly decrease the RE to about 1.7% (mm/mm) for 1 lead day (1.3% (mm/mm) for 3 lead days) and increase the NSE by about 0.4 for 1 lead day (0.34 for 3 lead days). Specifically, for the RE, the BMA and AKD methods are slightly better than the GPP and ExLR, while for the NSE, the GPP and ExLR are slightly better for smaller basins (basins #1 and #2), but the BMA and AKD methods are slightly better for bigger basins (basins #3 and #4). In general, the deterministic performances of the streamflow forecasts made by these four post-processing methods are not significantly different and largely resemble the deterministic performance for precipitation in Figure 1.

Figure 5 gives the probabilistic performance of ensemble streamflow forecasts using two probabilistic metrics. The probability exceeding 50% is used to calculate the Brier Skill Score (BSS). The results show that the probabilistic performance of GEFS is inferior in comparison, with CRPSS values of 0.39 and 0.35 and BSS values of 0.33

and 0.33 for 1 and 3 lead days, respectively. For BC, the improvement in the probabilistic performance measured by the CRPSS is not significant (increased by less than 10%). For the post-processing methods, the performance rank of these methods is GPP, ExLR, BMA, and AKD when measured in terms of both CRPSS and BSS values. The CRPSS of the generator-based methods (GPP and ExLR) is higher than that of the distribution-based methods by about 20%. For the BSS, the performances of the deterministic methods (BMA and AKD) are even lower than those of the BC. Furthermore, the ESP probabilistic performance is the best for the largest basin (Hengyang (basin #4)).

Figure 6 depicts the variation of CRPSS against the lead day (Subplot 1) and the effective lead day of the BC method (Subplot 2-5). As expected, the performances of the GEFS, BC and the four post-processing methods become worse with the increase of the lead day. The rates of the decrease in performance for the GPP, ExLR, BMA, and the AKD are-0.57/day, -0.55/day, -0.53/day, and -0.48/day, respectively. As reflected by the precipitation forecasts, there is an effective lead time for the BC method. The effective lead time is generally earlier for a small basin, with that for the Daxitan (basin #1) being less than 1 day, 5 lead days for the Xiangxiang (basin #2), 3 lead days for the Ganxi (basin #3), and 6 lead days for the Hengyang (basin #4). As expected, the effective lead time for BC in streamflow forecasts is shorter than that in weather forecasts, with a lag of about 1 day.

The ESP performance is further evaluated at the median runoff event (with the probability of 50%), as shown in Figure 7. The raw GEFS tends to be under-forecasting for low forecast probability (<50%) but over-forecasting for high forecast probability (>50%), indicated by the blue curve lying above the 1:1 diagonal line for forecast probability less than 50 %, and otherwise for forecast probability higher than 50%. A simple BC method brings almost no improvement compared to the GEFS because it does not explicitly reconstruct the ensemble spread. The ESPs produced by the four post-processing methods are considerably better than the GEFS-ESP, especially for forecast probability of less than 50%. The GPP-ESP gives the best reliability diagram, followed by the ExLR-ESP, BMA-ESP, and the AKD-ESP.

5. Discussion and conclusion

The use of a post-processing method when building a meteorological ESP is necessary, as it can remove the bias and reconstruct the proper ensemble spread for the raw ensemble forecasts. However, the post-processing methods most widely used in post-processing EWFs need to be further evaluated/compared when they are to be used in building the meteorological ESP. In this study, four state-of-the-art post-processing methods (GPP, BMA, AKD, and ExLR) are evaluated to compare their performance in EWFs and ESPs and to identify the possible influencing factors that would determine when and where they should best be used.

5.1 The performances of the post-processing methods

Four widely used post-processing methods were chosen for this study: two distribution-based methods (BMA and AKD) and two generator-based methods (GPP and ExLR). The post-processed EWFs and the corresponding ESPs were evaluated by both deterministic and probabilistic metrics. The BC method introduced by Chen et al.

(2014) is also included in the evaluation comparison as a benchmark method. The BC method adopts a simple linear correction to remove the bias and does not explicitly reconstruct the ensemble spread. For shorter lead days, the BC performance is comparable to that of the post-processing methods in terms of deterministic metrics for temperature forecasts, but its advantages in precipitation and streamflow forecasts are not significant. The performances of the post-processing methods vary with the forecast variables, e.g. precipitation or air temperature, and in different forecast categories, e.g. weather forecasts or streamflow forecasts. Post-processing precipitation forecasts are much more difficult than air temperature forecasts because the error in precipitation forecasts is highly non-normally distributed, and precipitation involves other difficulties including its discrete-continuous nature, the forecast uncertainty that varies with the precipitation magnitude, all of which are compounded by the insufficient records for extreme precipitation (Scheuerer et al., 2015). For this study, the generator-based methods, i.e. GPP and ExLR, show their advantages in the probabilistic performance for precipitation forecasts (about 10% better than the distribution-based methods), and the benefit can be propagated to the probabilistic performance for streamflow forecasts (for about a 20% better than the distribution-based methods). In contrast, the distribution-based methods, i.e., BMA and AKD, are more competitive in their deterministic performance than in their probabilistic performance for precipitation forecasts (by about 15% better than the generator-based methods), but the benefit in the corresponding streamflow forecasts is generally not significant.

5.2 Performance uncertainty

The performance of post-processing methods can vary depending upon the basin characteristics, including basin size, climate type, topography and hydrological conditions (Hagedorn et al., 2008; Scheuerer and Hamill, 2015; Siddique and Mejia, 2017). In this study, four sub-basins of different sizes were obtained from the same basin to ensure they share a similar climate and similar rainfall-runoff characteristics for hydrological forecasts. This study found that when using the BC method, the effective lead time for precipitation forecasts increased from 2 to 7 days from the smallest basin (Basin #1) to the largest basin (#4), with about a 1-day lag for streamflow forecasts. The post-processing methods performed better in the larger sub-basins than in the smaller ones for precipitation and streamflow, but that was not the case for air temperature.

5.3 Further Discussion

One concern with this study is that the auto-correlation of precipitation or of temperature stations was not considered in the post-processing methods. However, the correlation structure of the raw ensemble forecasts may be implicitly conveyed to the post-processing method. For example, if the forecasts show rainy events for several consecutive days, the distribution model built by the post-processing method will likely predict a bigger precipitation occurrence for those days. Also, if there is one day with heavy precipitation, the predictive distribution for that day may be skewed to the right (larger precipitation), and the generated ensemble may have a larger proportion of big rainfall events.

The other concern is that the dependence between precipitation and air temperature is violated in the post-processing ensemble. For hydrological modeling, precipitation and air temperature together control the water balance in a basin. The inconsistency between the two variables from the post-processed results may result in a deviation of the forecasted discharge hydrograph compared with observations. Taking the Hengyang Basin (#4) as an example, the precipitation-air temperature (P-T) correlation of the 1-lead-day ensemble weather forecasts for the GEFS, BC, GPP, ExLR, BMA, and AKD methods is about 0.35, 0.22, 0.17, 0.19, 0.20, and 0.20, respectively, compared to 0.08 from the observation data. This indicates that the real P-T correlation is over-estimated by the raw GEFS forecasts, and while all of the post-processing methods tend to decrease the P-T coefficient compared to the GEFS, they are still over-estimated. The drop of the P-T coefficient from the GEFS to the post-processing methods is as expected since the precipitation and air temperature are post-processed independently. For future studies, it will be interesting to investigate whether the specific procedure of constructing P-T correlation in post-processing will improve the performance of ESP.

6. Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (Grant No. 51779176, 51539009), the Overseas Expertise Introduction Project for Discipline Innovation (111 Project) funded by the Ministry of Education and State Administration of Foreign Experts Affairs P.R. China (Grant No. B18037), the Thousand Youth Talents Plan from the Organization Department of the CCP Central Committee (Wuhan University, China), and the Research Council of Norway

(FRINATEK Project 274310). The authors would like to acknowledge the National Oceanic and Atmospheric Administration (NOAA), Boulder, Colorado, USA for providing GEFS ensemble precipitation and air temperature reforecasts. The authors wish to thank the China Meteorological Data Sharing Service System and the Hydrology and Water Resources Bureau of Hunan Province, China for providing the daily meteorological and hydrological data in the Xiangjiang basin.

7. Conflict of Interest

The authors declare that they have no conflict of interest with the information presented in this study.

8. Reference

Alfieri L, Pappenberger F, Wetterhall F, Haiden T, Richardson D, Salamon P (2014) Evaluation of ensemble streamflow predictions in Europe. Journal of Hydrology 517:913–922. http://doi.org/10.1016/j.jhydrol.2014.06.035

Bröcker J, Smith LA (2008) From ensemble forecasts to predictive distribution functions. Tellus A 60:663–678. http://doi.org/10.1111/j.1600-0870.2008.00333.x

Baran S, Nemoda D (2016) Censored and shifted gamma distribution based EMOS m odel for probabilistic quantitative precipitation forecasting. Environmetrics 27:28 0–292. http://doi.org/10.1002/env.2391

Brown JD, Demargne J, Seo D-J, Liu Y (2010) The Ensemble Verification System

(EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environmental Modelling and Software 25:854–872. http://doi.org/10.1016/j.envsoft.2010.01.009

Chen J, Brissette FP, Li Z (2014) Postprocessing of ensemble weather forecasts using a stochastic weather generator. Monthly Weather Review 142:1106–1124. http://doi.org/10.1175/MWR-D-13-00180.1

Cloke HL, Pappenberger F (2009) Ensemble flood forecasting: a review. Journal of Hydrology 375:613–626. http://doi.org/10.1016/j.jhydrol.2009.06.005

Hagedorn R, Hamill TM, Whitaker JS (2008) Probabilistic forecast calibration using ecmwf and GFS ensemble reforecasts Part I: two-meter temperatures. Monthly Weather Review 136:2608–2619. http://doi.org/10.1175/2007MWR2410.1

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. http://doi.org/10.1175/2007MWR2411.1

Kavetski D, Kuczera G, Franks SW (2006a) Bayesian analysis of input uncertainty in hydrological modeling: 1. theory. Water Resources Research 42:W03407. http://doi.org/10.1029/2005WR004368

Kavetski D, Kuczera G, Franks SW (2006b) Bayesian analysis of input uncertainty in hydrological modeling: 2. application. Water Resources Research 42:W03408. http://doi.org/10.1029/2005WR004376

Liu X, Coulibaly P (2011) Downscaling ensemble weather predictions for improved week-2 hydrologic forecasting. Journal of Hydrometeorology 12:1564–1580. http://doi.org/10.1175/2011JHM1366.1

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review 133:1155–1174.

http://doi.org/10.1175/MWR2906.1

Roulin E (2007) Skill and relative economic value of medium-range hydrological ensemble predictions. Hydrology and Earth System Sciences 11:725-737.

http://doi.org/10.5194/hess-11-725-2007

Roulin E, Vannitsem S (2012) Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. Monthly Weather Review 140:874–888.

http://doi.org/10.1175/MWR-D-11-00062.1

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

http://doi.org/10.1175/MWR-D-15-0061.1

Schmeits MJ, Kok KJ (2010) A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. Monthly Weather Review 138:4199–4211. http://doi.org/10.1175/2010MWR3285.1

Shukla S, Voisin N, Lettenmaier DP (2012) Value of medium range weather forecasts in the improvement of seasonal hydrologic prediction skill. Hydrology and Earth System Sciences 16:2825–2838. http://doi.org/10.5194/hess-16-2825-2012

Siddique R, Mejia A (2017) Ensemble streamflow forecasting across the U.S. Mid-Atlantic region with a distributed hydrological model forced by GEFS reforecasts. Journal of Hydrometeorology 18:1905–1928. http://doi.org/10.1175/JHM-D-16-0243.1

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

Vannitsem SP, Hagedorn R (2011) Ensemble forecast post-processing over Belgium: comparison of deterministic-like and ensemble regression methods. Meteorological Applications 18:94–104. http://doi.org/10.1002/met.217

Vetter T, Reinhardt J, Flörke M, van Griensven A, Hattermann F, Huang S, et al. (2016) Evaluation of sources of uncertainty in projected hydrological changes under climate change in 12 large-scale river basins. Climatic Change 141:419–433. http://doi.org/10.1007/s10584-016-1794-y

Verkade JS, Brown JD, Reggiani P, Weerts AH (2013) Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. Journal of Hydrology 501:73–91. http://doi.org/10.1016/j.jhydrol.2013.07.039

Wilks DS (2006) Comparison of ensemble-MOS methods in the Lorenz '96 setting. Meteorological Applications 13:243–14. http://doi.org/10.1017/S1350482706002192

Wilks DS, Hamill TM (2007) Comparison of ensemble-MOS methods using GFS reforecasts. Monthly Weather Review 135:2379–2390. http://doi.org/10.1175/MWR3402.1

Wilks DS (2009) Extending logistic regression to provide full-probability-distribution MOS forecasts. Meteorological Applications 16:361–368. http://doi.org/10.1002/met.134

Xu Y-P, Tung Y-K (2007) Decision making in water management under uncertainty. Water Resources Management 22:535–550. http://doi.org/10.1007/s11269-007-9176-x

9. Figure Captions

Figure 1. Bubble plot of MAEs for the ensemble mean forecasts by the GEFS, BC, GPP, ExLR, BMA, and AKD methods over four basins. The different basins for each subplot are represented by circles of different sizes and colors. A larger basin is associated with a larger circle. The left column MAEs are for 1 lead day and the right is for 3 lead days; the top row is for precipitation and the bottom row is for temperature.

Figure 2. Bubble plot of CRPSS for the ensemble forecasts by the GEFS, BC, GPP, ExLR, BMA, and AKD methods over four basins. The different basins for each subplot are represented by circles with different sizes and colors, and a larger basin is associated with a larger circle. The left column columns are for 1 lead day and the right is for 3 lead days; the top row is for precipitation and the bottom row is for air temperature.

Figure 3. (1) 3-D bar plot of the variation of CRPSS against lead time for precipitation; results obtained by averaging the CRPSS over 4 basins. (2-5) Line chart of the CRPSS for ensemble precipitation forecasts by the GEFS and BC method

Figure 4. Bubble plot of the deterministic metrics for the ensemble mean forecasts by the GEFS, BC, GPP, ExLR, BMA and AKD methods over four basins. The different basins for each subplot are represented by circles with different sizes and colors, and a larger basin is associated with a larger circle. The top row shows the RE while the bottom row shows the NSE. The left column gives the 1-lead day results while the right column offers the 3 lead day results.

Figure 5. Bubble plot of the probabilistic metrics for the ensemble mean forecasts by GEFS, BC, GPP, ExLR, BMA, and AKD over four basins. Different basins for each subplot are represented by different circles with different size and color. And a larger basin is associated with a larger circle. And the top row is for BSS while the bottom row is for CRPSS. Results for 1 lead day are in the left column and results for 3 lead days are in the right column.

Figure 6. (1) 3-D bar plot of CRPSS values for ensemble streamflow forecasts for different lead days and post-processing methods. (2-5) Line chart of CRPSS results for ensemble precipitation forecasts by GEFS and BC.

Figure 7. Reliability diagram of the ESPs by the GEFS, BC, GPP, ExLR, BMA, and AKD methods for 1 lead day over four basins, for: (1) the Xiangxiang station, (2) the Daxitan station, (3) the Hengyang station, and (4) the Ganxi station. The forecast event is the median runoff with the probability of 50%.

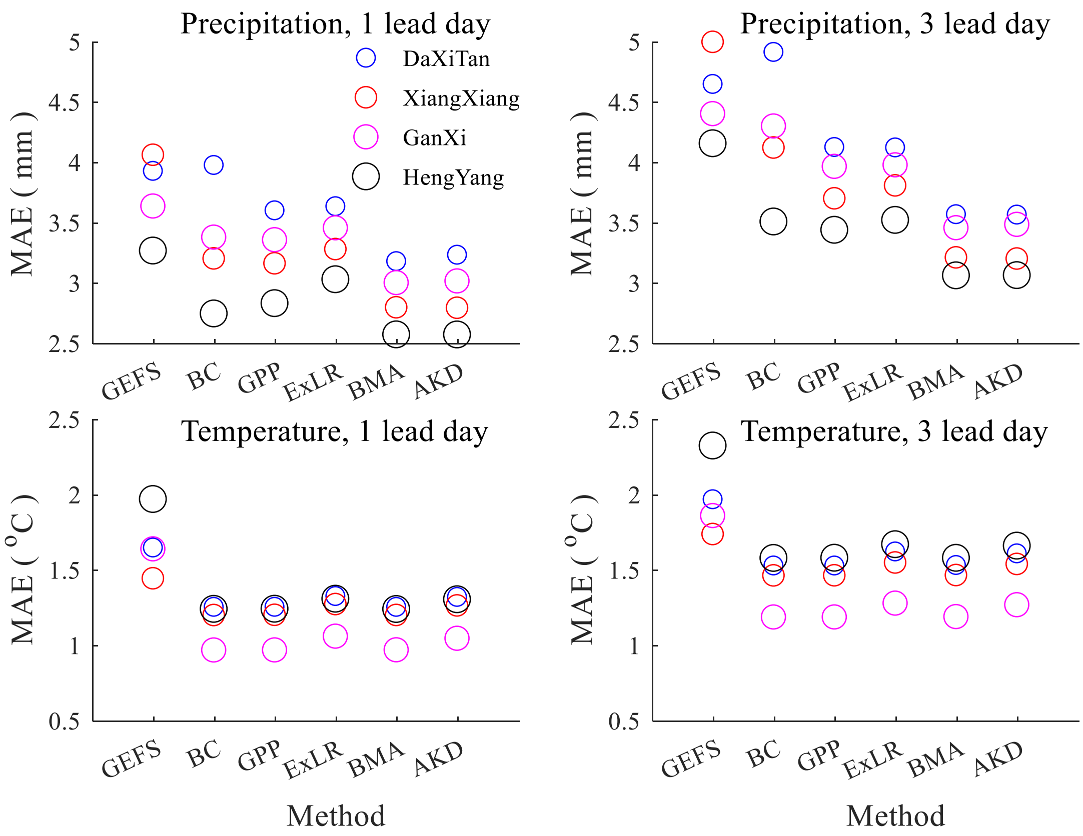


Figure 1. Bubble plot of MAEs for the ensemble mean forecasts by the GEFS, BC, GPP, ExLR, BMA, and AKD methods over four basins. The different basins for each subplot are represented by circles of different sizes and colors. A larger basin is associated with a larger circle. The left column MAEs are for 1 lead day and the right is for 3 lead days; the top row is for precipitation and the bottom row is for temperature.

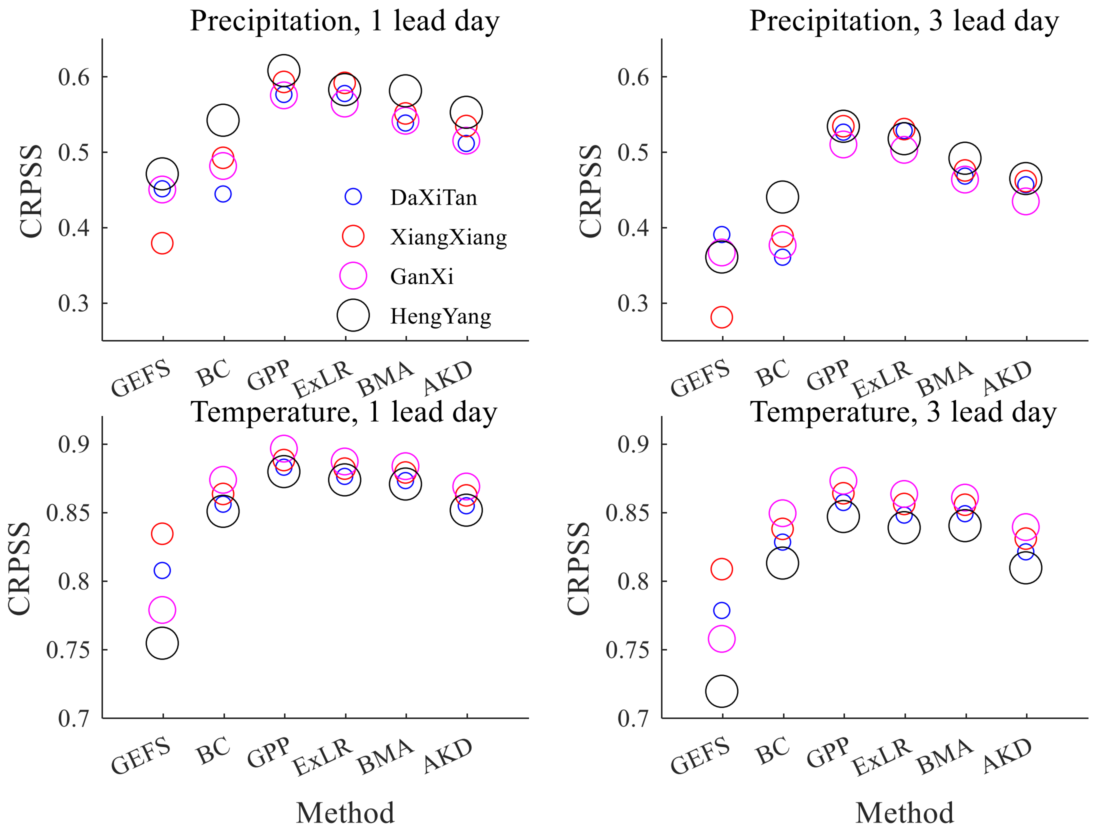


Figure 2. Bubble plot of CRPSS for the ensemble forecasts by the GEFS, BC, GPP, ExLR, BMA, and AKD methods over four basins. The different basins for each subplot are represented by circles with different sizes and colors, and a larger basin is associated with a larger circle. The left column columns are for 1 lead day and the right is for 3 lead days; the top row is for precipitation and the bottom row is for air temperature.

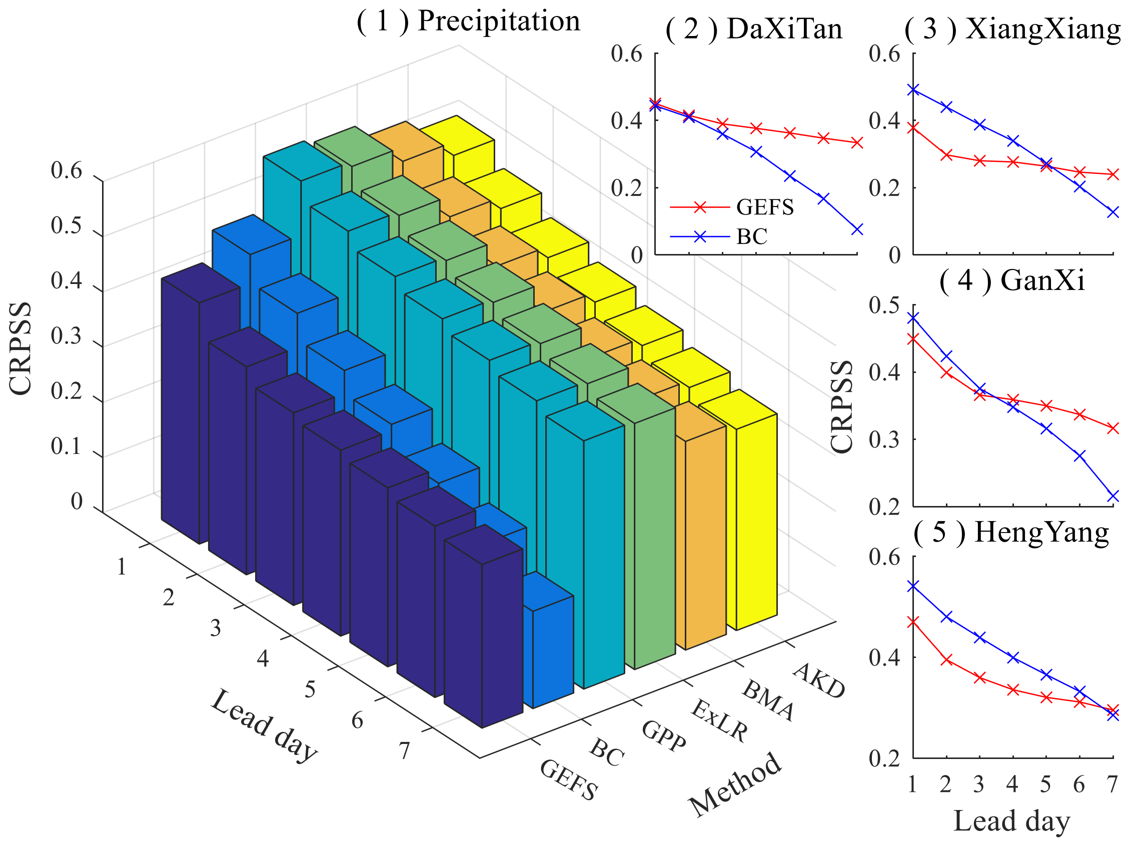


Figure 3. (1) 3-D bar plot of the variation of CRPSS against lead time for precipitation; results obtained by averaging the CRPSS over 4 basins. (2-5) Line chart of the CRPSS for ensemble precipitation forecasts by the GEFS and BC method

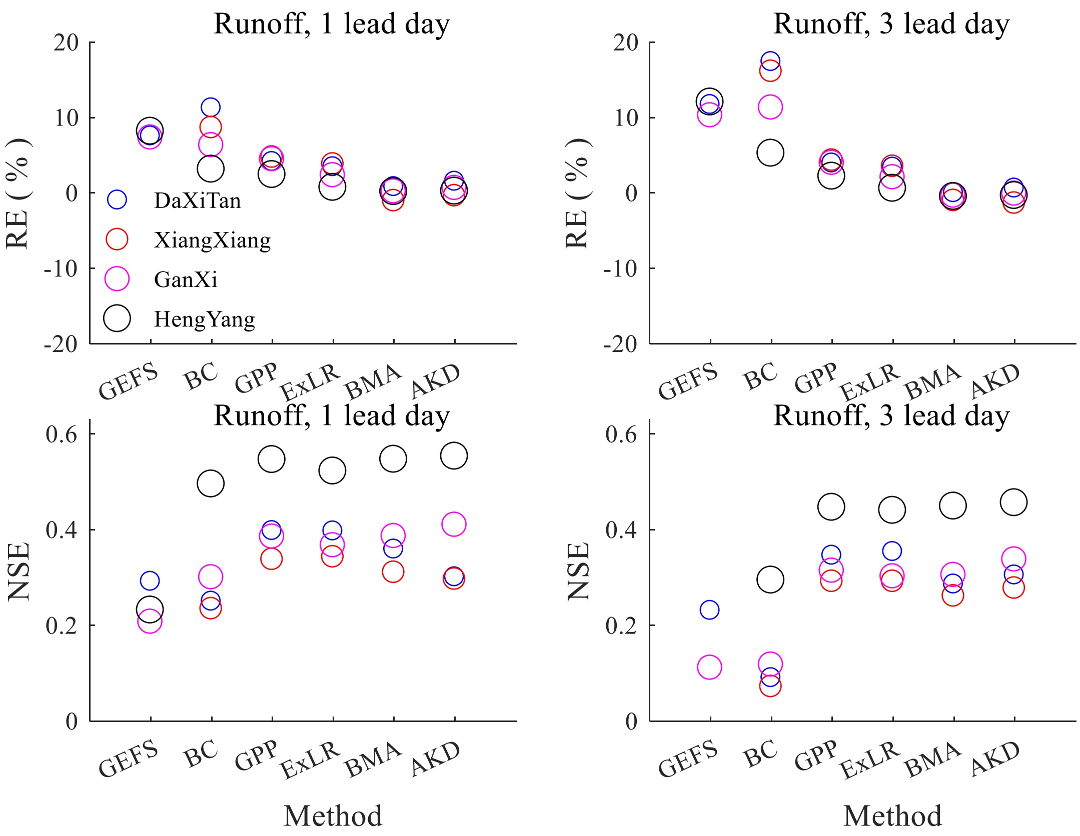


Figure 4. Bubble plot of the deterministic metrics for the ensemble mean forecasts by the GEFS, BC, GPP, ExLR, BMA and AKD methods over four basins. The different basins for each subplot are represented by circles with different sizes and colors, and a larger basin is associated with a larger circle. The top row shows the RE while the bottom row shows the NSE. The left column gives the 1-lead day results while the right column offers the 3 lead day results.

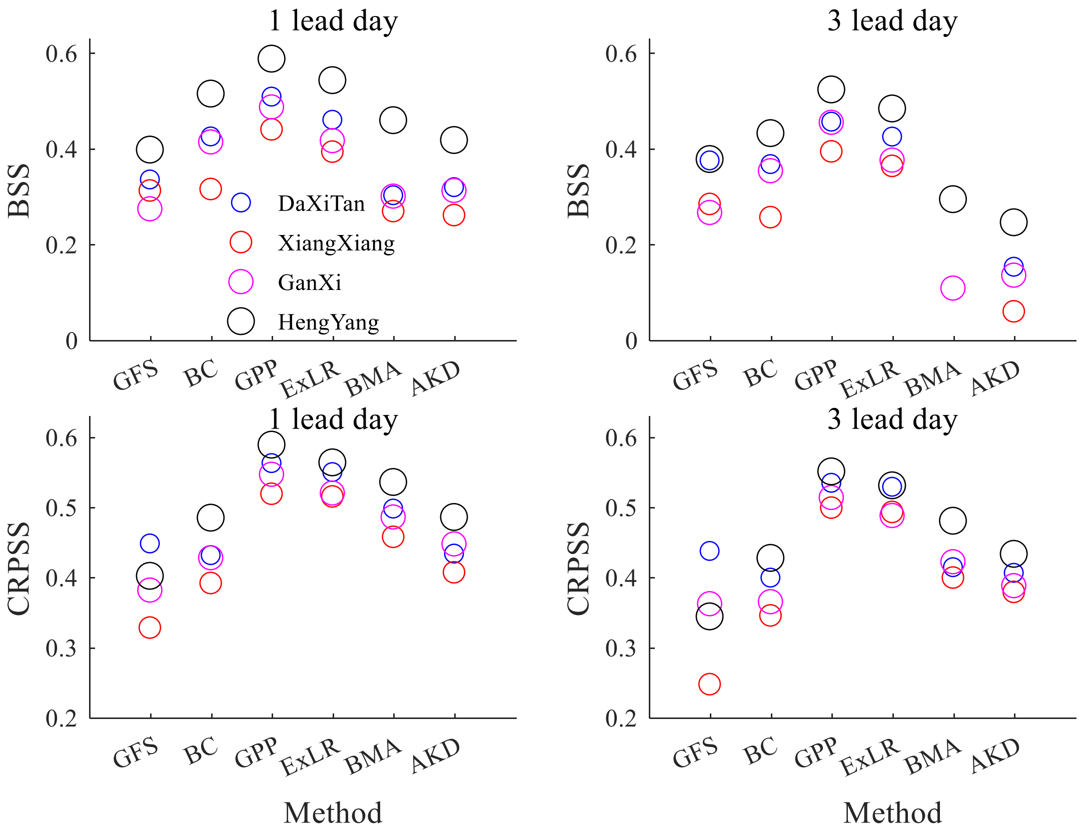


Figure 5. Bubble plot of the probabilistic metrics for the ensemble mean forecasts by GEFS, BC, GPP, ExLR, BMA, and AKD over four basins. Different basins for each subplot are represented by different circles with different size and color. And a larger basin is associated with a larger circle. And the top row is for BSS while the bottom row is for CRPSS. Results for 1 lead day are in the left column and results for 3 lead days are in the right column.

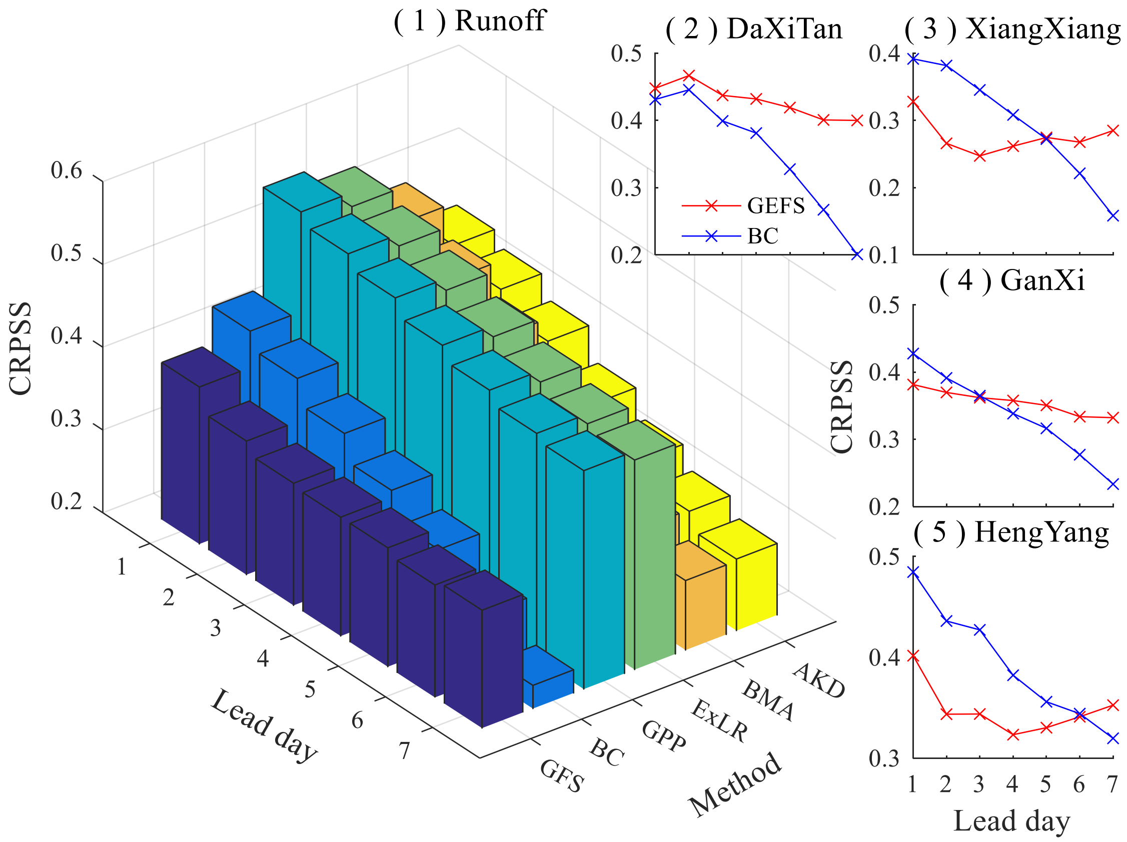


Figure 6. (1) 3-D bar plot of CRPSS values for ensemble streamflow forecasts for different lead days and post-processing methods. (2-5) Line chart of CRPSS results for ensemble precipitation forecasts by GEFS and BC.

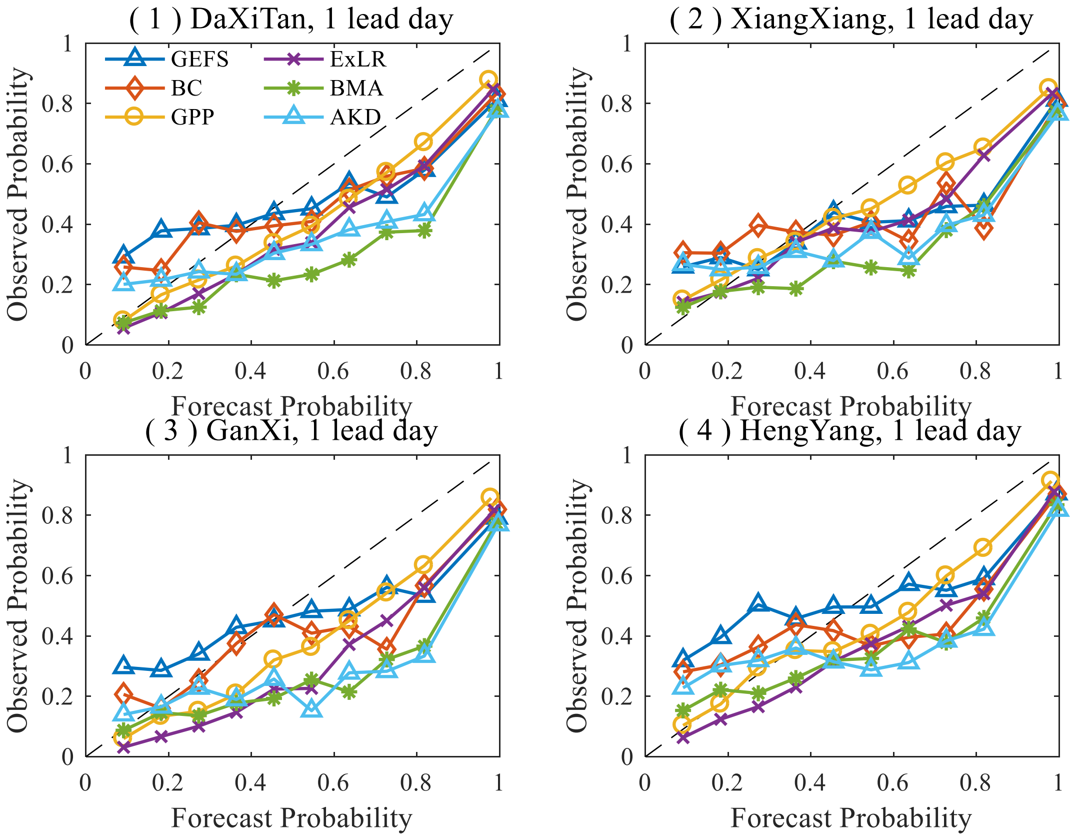


Figure 7. Reliability diagram of the ESPs by the GEFS, BC, GPP, ExLR, BMA, and AKD methods for 1 lead day over four basins, for: (1) the Xiangxiang station, (2) the Daxitan station, (3) the Hengyang station, and (4) the Ganxi station. The forecast event is the median runoff with the probability of 50%.

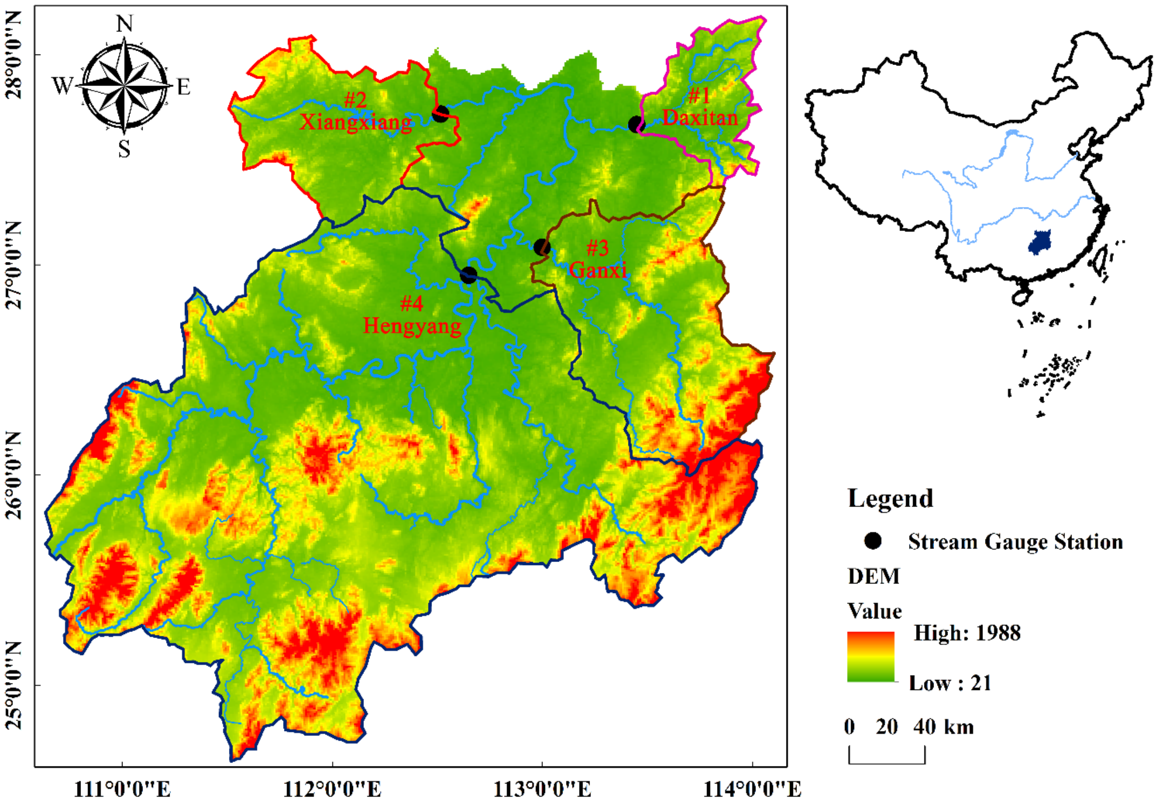
Supplementary Material 1

The Description to the Xiangjiang Basin

Xiangjiang basin owns a total drainage area of 94,660 km^2 and a registered river length of 856 km. As a subtropical monsoon area, the mean annual precipitation of this basin from 1979 to 2014 is about 1,510 mm, of which 65% occurs in the rainy season from April to September. The average annual air temperature is about 18.5 ^{o}C and the mean air temperature for the coldest month (January) is about 3.5 ^{o}C. The upstream region of Xiangjiang basin consists of mountains (mean elevation >200 m) and hills (mean elevation between 100~200m), while the downstream part is mainly plains (mean elevation <100m). The rapid decline in elevation from upstream to downstream accelerate the rainfall convergence and the rapid variation of water level, with about 68% of the total runoff yielding during April to September and 50% of the flooding events occurring in June and July (Xu et al., 2013).

Abstracts:

Xu H, Xu C-Y, Chen H, Zhang Z, Li L (2013) Assessing the influence of rain gauge density and distribution on hydrological model performance in a humid region of China. Journal of Hydrology 505:1–12. http://doi.org/10.1016/j.jhydrol.2013.09.004



Map of Xiangjiang basin as well as four selected sub-basins (Daxitan (#1), Xiangxiang (#2), Ganxi (#3) and Hengyang (#4)

Supplementary Material 2

The Description to the Xin’anjiang Model

Xin’anjiang (XAJ) hydrological model is by far the most widely used lumped model for hydrological simulation and forecasting in China, and its applicability in semi-humid and humid regions has been extensively demonstrated (Jayawardena and Zhou, 2000; Xu et al., 2013; Zeng et al., 2016) since it was developed in 1973 (Zhao, 1992).

For this model, the basin is generalized into a soil box with three stacked layers, that is, the upper soil layer, the lower soil layer, and the deep soil layer. The areal mean tension water storage capacity for three layers are represented as *UM*, *LM*, and *DM*, respectively. The total areal mean tension water storage capacity for this basin is *WM*. The potential evaporation rate (*EP*) is transformed from the pan evaporation measurements using a conversion coefficient (*KE*). The actual evaporation (*E*) of the basin is a sum of evaporation rate happening in three soil layers, as represented by *EU*, *EL*, and *ED*, using a three-layer evaporation conceptual model. The parameter *C* controls the deep layer evaporation. Runoff (*R*) is generated from the rainfall excess and soil storage deficit based on the concept of “runoff formation on repletion of storage” which means that *R* can only be yielded from rainfall when the soil moisture (*W(0)*) of the aeration zone reaches *WM*. The parameter *B* is used to describe the non-uniformity of the surface condition when calculating the generated *R*. The generated *R* is then divided into three parts, surface runoff (*RS*), interflow (*RI*), and groundwater runoff (*RG*), using a free water reservoir whose storage capacity (*S(0)*) is unevenly distributed over the basin associated with the parameter-free water distribution index *EX*, areal mean free water storage capacity *SM,* the coefficient of free water storage to interflow *KI* and groundwater flow *KG*. The *RS* is routed to the basin outlet using a unit hydrograph associated with the parameter *n* and *NK*, while *RI* and *RG* are routed through a linear reservoir with the recession coefficients *CI* and *CG,* respectively. After routing, the sum of surface discharge, interflow discharge, and groundwater discharge is the simulated runoff for this model. Besides, the parameter *IMP* defines the proportion of impermeable area to the total catchment area. Table 1 (below) lists the 15 parameters of the XAJ model. The inputs for this model are daily areal precipitation, the measured daily pan evaporation, while the outputs are the outlet discharge, the actual evaporation rate.

Since the measured daily pan evaporation may be hard to obtain for meteorological forecasting, a simple and efficient potential evaporation (PE) model proposed by Oudin et al. (2005) can be used to transform daily mean air temperature into pan evaporation rate. Besides, the observed meteorological data of stations and gridded EWFs will be averaged for each river basin using the Thiessen polygon method.

When the XAJ model is built for the four basins, a split-sample calibration methods is used. Specifically, the first year is used as warm-up period and the remaining data, of which about 2/3 (Basins #1, #2, and #3: 1980-2004; Basin #4: 1980-1999) is used for calibrating and 1/3 (Basins #1, #2, and #3: 2005-2014; Basin #4: 2000-2009) is used for validation. Two commonly used metrics are chosen for the evaluating the performance of hydrological modeling, that is, Nash–Sutcliffe coefficient (NSE) and relative error (RE). The results displayed in Table 2 show that the NSE value is larger than 0.8 and the RE is falling in the range of 10% for the validation period for 4 river basins, indicating a well-calibrated of the hydrological model.

Abstracts

Jayawardena AW, Zhou, MC (2000). A modified spatial soil moisture storage capacity distribution curve for the Xinanjiang model. Journal of Hydrology 227:93–113. http://doi.org/10.1016/S0022-1694(99)00173-0

Oudin L, Hervieu F, Michel C, Perrin C, Andréassian V, Anctil F, Loumagne C (2005) Which potential evapotranspiration input for a lumped rainfall–runoff model? Journal of Hydrology 303:290–306. http://doi.org/10.1016/j.jhydrol.2004.08.026

Xu H, Xu C-Y, Chen H, Zhang Z, Li L (2013) Assessing the influence of rain gauge density and distribution on hydrological model performance in a humid region of China. Journal of Hydrology 505:1–12. http://doi.org/10.1016/j.jhydrol.2013.09.004

Zeng Q, Chen H, Xu C-Y, Jie M-X, Hou Y-K (2015) Feasibility and uncertainty of using conceptual rainfall-runoff models in design flood estimation. Hydrology Research 47:701–717. http://doi.org/10.2166/nh.2015.069

Ren-Jun Z (1992) The Xinanjiang model applied in China. Journal of Hydrology 135:371–381.

http://doi.org/10.1016/0022-1694(92)90096-E

Table 1 Parameters of the Xin'anjiang model

Number	Parameter	Explanation	Range	unit
1	KE	Ratio of potential evapotranspiration to pan evaporation	0.5-1.5	-
2	WM	Areal mean tension water storage capacity	80-150	mm
3	UM	Upper layer tension water storage capacity	-	mm
4	LM	Lower layer tension water storage capacity	-	mm
5	B	Tension water distribution index	0-0.5	-
6	IMP	Impermeable coefficient	0-0.2	-
7	SM	Areal mean free water storage capacity	0-100	mm
8	EX	Free water distribution index	1-2	-
9	KI	Outflow coefficient of free water storage to interflow	0-1.0	-
10	KG	Outflow coefficient of free water storage to groundwater flow	0-1.0	-
11	C	Deep layer evapotranspiration coefficient	0-0.2	-
12	CI	Interflow recession coefficient	0.5-1.0	-
13	CG	Groundwater recession coefficient	0.5-1.0	-
14	n	The parameter of Nash unit hydrograph	-	-
15	NK	The parameter of Nash unit hydrograph	-	-

Table 2 Results for model calibration and validation

Station		**Warm-Up**	**Calibration**			**Validation**		
ID	Name	Period	Period	NSE	RE%	Period	NSE	RE%
#1	Xiangxiang	1979-1980	1980-2004	0.87	1.1	2005-2014	0.84	-7.8
#2	Daxitan	1979-1980	1980-2004	0.92	-0.2	2005-2014	0.90	7.1
#3	Hengyang	1979-1980	1980-2004	0.93	-0.7	2005-2014	0.87	-4.5
#4	Ganxi	1979-1980	1980-1999	0.94	-0.9	2000-2009	0.82	2.5

Supplementary Material 3

The Description to the Verification Metrics

Metric	Description
MAE	MAE measure the mean absolute difference between the ensemble mean forecasts and the observations, and a smaller MAE is preferred.
NSE	NSE determines the relative magnitude of the residual variance compared to the measure data variance, and NSE is positive oriented and being 1 means perfect.
RE	RE measures the mean difference between ensemble mean forecasts and the observations, reflecting the model's ability to maintain the water balance.
CRPSS	CRPSS verifies the skill of the ensemble spread of EWF over climatology, where the skill is defined as the mean squared difference between the distribution of ensemble forecasts and corresponding distributions of observations.
BSS	BSS measures the performance of one forecasting system relative to another in terms of the Brier Score (BS), where BS measures the average square error of a probability forecast of a dichotomous event.
Reliability diagram	Reliability diagram plots the observed frequency against the forecast probability for a discrete event, any deviation from the 1:1 diagonal line denotes the conditional bias of the probabilistic forecasts. Over-forecasting occurs when the plotted curve lies below the 1:1 line, while under-forecasting happens when the plotted curve lies on the other side.

Supplementary Material 4

The Additional Results

Figure 1 presents the variation of CRPSS averaged over 4 basins for air temperature forecasts (Subplot 1), and the line plot of CRPS for BC and GEFS against lead time (subplot 2-5). The CRPSS for GEFS, BC and 4 post-processing methods become as expected worse with the increasing lead time. GPP tend to be the best one, followed by ExLR and BMA, and then BC, finally being GEFS. Unlike the precipitation forecasts, there is no valid lead day for air temperature when using BC.

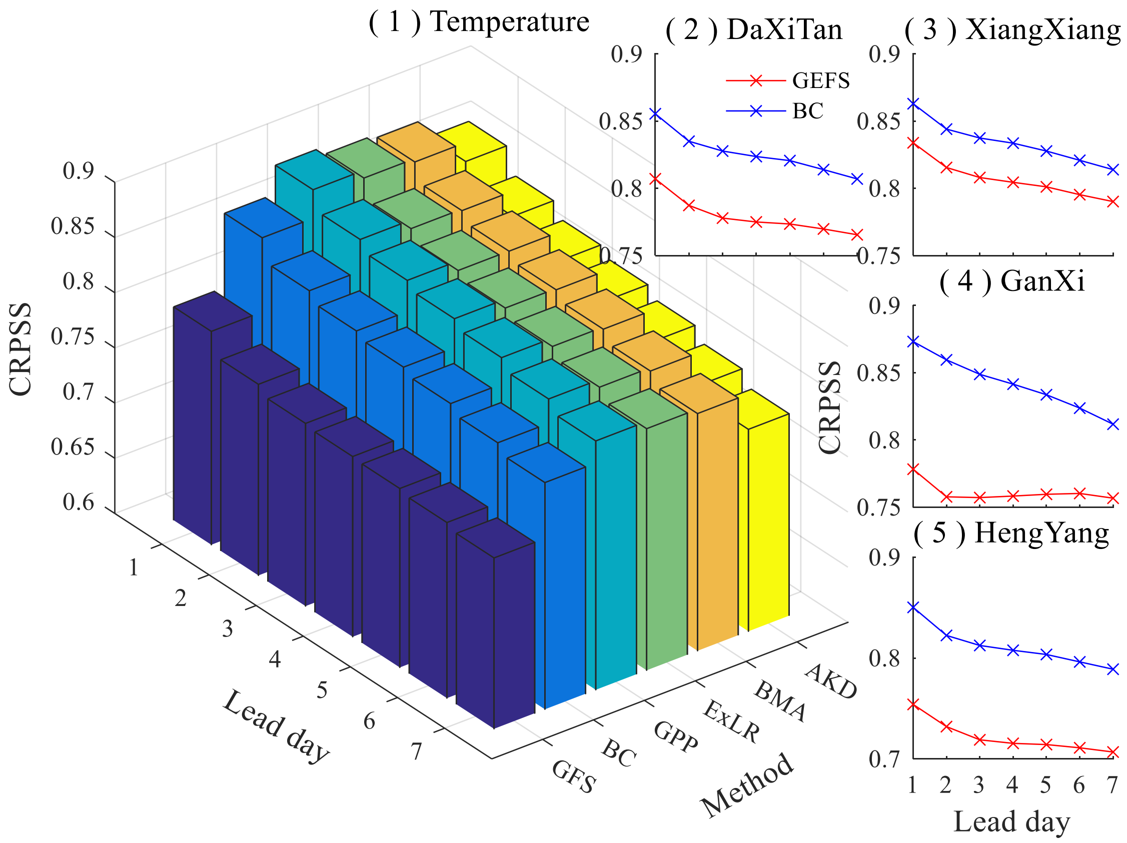


Fig 1. (1) 3-D bar plot of the variation of CRPSS against lead for temperature, the results are averaged for all basins. (2-5) Line chart of CRPSS for ensemble temperature forecasts by GEFS and BC.

Figure 2 presents the variation of CRPSS averaged over 4 basins for streamflow forecasts (Subplot 1), and the line plot of CRPS for BC and GEFS against lead time (subplot 2-5). NSE that is less than zero will be excluded from this figure. The performance of ESP measured in NSE become worse with the lead day, and the decreasing sloop for the 4 post-processing methods is similar (about -0.43/day). The valid lead day for BC method is generally earlier for a small basin, with Daxitan (basin #1) being less than 1 day, Xiangxiang (basin #2) being about 5 lead days, Ganxi (basin #3) being 3 lead days and Hengyang (basin #4) being about 6 lead days.

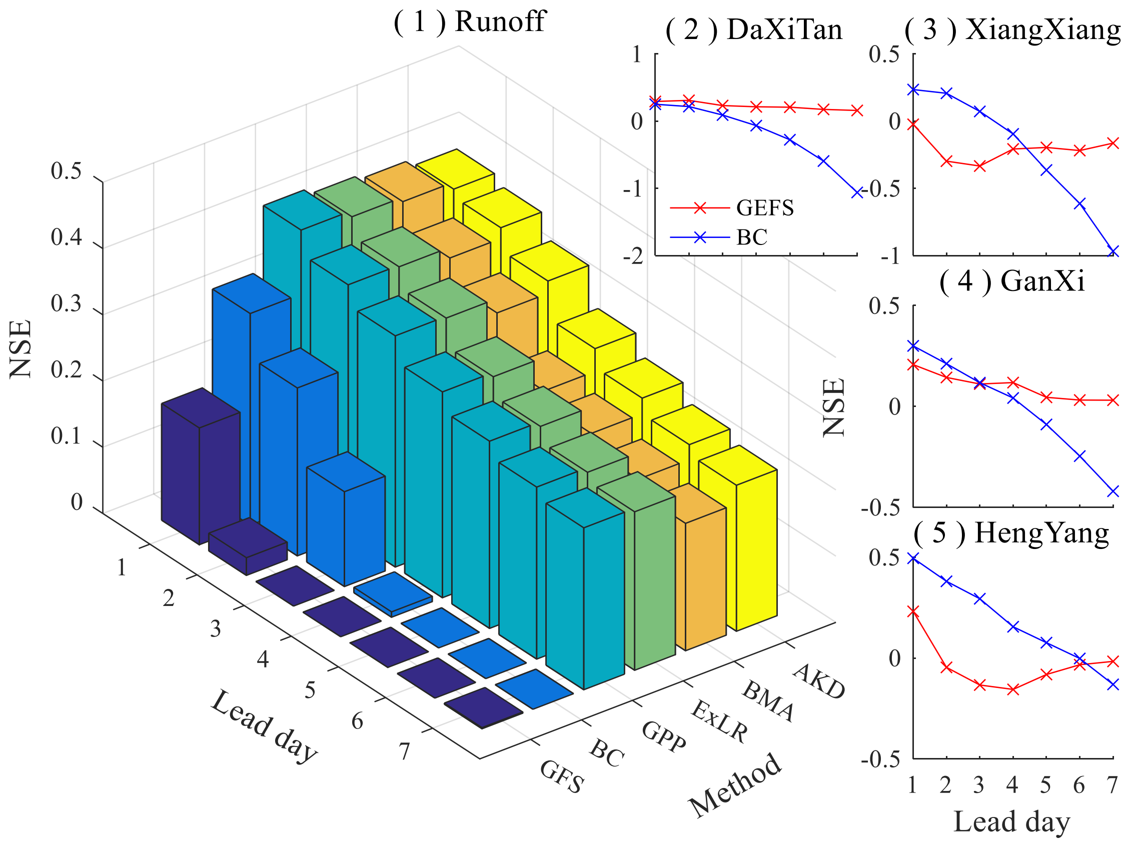


Fig 2. (1) 3-D bar plot of NSE for ensemble streamflow forecasts against different lead day and the post-processing methods. (2-5) Line chart of CRPSS for ensemble precipitation forecasts by GEFS-ESP and BC-ESP.