Factors Influencing the Performance of Regression-Based Statistical Postprocessing Models for Short-Term Precipitation Forecasts^①

WENTAO LI

Beijing Normal University, Beijing, China, and University of Melbourne, Melbourne, Victoria, Australia

QINGYUN DUAN

Hohai University, Nanjing, China

QUAN J. WANG

University of Melbourne, Victoria, Australia

(Manuscript received 18 June 2019, in final form 10 September 2019)

ABSTRACT

Statistical postprocessing models can be used to correct bias and dispersion errors in raw precipitation forecasts from numerical weather prediction models. In this study, we conducted experiments to investigate four factors that influence the performance of regression-based postprocessing models with normalization transformations for short-term precipitation forecasts. The factors are 1) normalization transformations, 2) incorporation of ensemble spread as a predictor in the model, 3) objective function for parameter inference, and 4) two postprocessing schemes, including distributional regression and joint probability models. The experiments on the first three factors are based on variants of a censored regression model with conditional heteroscedasticity (CRCH). For the fourth factor, we compared CRCH as an example of the distributional regression with a joint probability model. The results show that the CRCH with normal quantile transformation (NQT) or power transformation performs better than the CRCH with log-sinh transformation for most of the subbasins in Huai River basin with a subhumid climate. The incorporation of ensemble spread as a predictor in CRCH models can improve forecast skill in our research region at short lead times. The influence of different objective functions (minimum continuous ranked probability score or maximum likelihood) on postprocessed results is limited to a few relatively dry subbasins in the research region. Both the distributional regression and the joint probability models have their advantages, and they are both able to achieve reliable and skillful forecasts.

1. Introduction

Ensemble weather forecasts have been provided by meteorological ensemble prediction systems (EPSs) in weather centers around the world. The ensemble weather forecasts are aimed to quantify the uncertainty of the forecasts for use in various applications. For example, hydrologists use precipitation ensemble forecasts as inputs to hydrological models to generate ensemble

hydrological forecasts (Cloke and Pappenberger 2009; Schaake et al. 2007b).

However, current ensemble prediction systems cannot fully simulate all sources of uncertainty. The NWP forecasts usually suffer from bias and dispersion errors (Buizza et al. 2005; Park et al. 2008). Statistical postprocessing models can be applied to correct the bias and dispersion errors in raw weather forecasts (Cuo et al. 2011; Gneiting and Katzfuss 2014; Schaake et al. 2007b). Early postprocessing models include analog (Hamill and Whitaker 2006), model output statistics (Glahn and Lowry 1972), and perfect prognosis (Klein et al. 1959). A wide variety of postprocessing models have been developed during recent decades. Parametric postprocessing models can generally be divided into kernel density estimation (KDM) models and regression-based

^① Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-19-0121.s1>.

Corresponding author: Qingyun Duan, qyduan@hhu.edu.cn

models (Wilks 2011). Examples of the KDM models include Bayesian model averaging (Raftery et al. 2005; Sloughter et al. 2007) and ensemble dressing (Boucher et al. 2015; Fortin et al. 2006; Roulston and Smith 2003; Wang and Bishop 2005). The regression-based models include logistic regression (Messner et al. 2014a; Scheuerer and Hamill 2015; Wilks 2009), quantile regression (Friederichs and Hense 2007), and ensemble model output statistics (EMOS; Fraley et al. 2010; Gneiting et al. 2005; Scheuerer and Hamill 2015; Sloughter et al. 2007). Other regression-based models include the Bayesian processor of forecast (Krzysztofowicz and Evans 2008), the meta-Gaussian distribution (MGD) models (Schaake et al. 2007a; Wu et al. 2011), and Bayesian joint probability (BJP; Robertson et al. 2013; Shrestha et al. 2015; Wang et al. 2009). More details about the recent development of these postprocessing models for hydro-meteorological forecasting can be found in related books (Duan et al. 2019; Vannitsem et al. 2018) and a recent review (Li et al. 2017).

In this study, we mainly focus on regression-based postprocessing models that include transformations to treat nonnormal precipitation forecasts. There are generally three steps in most of this kind of postprocessing model, including 1) transformations to normalize data, 2) modeling of the conditional distribution of observations given forecasts, and 3) back-transformation. Transformations are applied to the forecasts and observations, then the regression model is fitted to the transformed data. There are generally two schemes to model the conditional distribution of the observations given forecasts. One scheme is EMOS or distributional regression (Klein et al. 2015), which uses statistics of raw forecasts to predict parameters of the conditional distribution of the observations. Another scheme is to fit a joint probability model first, then to derive the conditional distribution from the joint probability model, such as in BJP or MGD. The parameters in the postprocessing models can be estimated by different objective functions including minimum continuous ranked probability score (CRPS) or maximum likelihood. Finally, new predictions from the regression models are transformed back into the original space.

Although there are already some studies that investigated several factors in the above postprocessing schemes, such as the choice of the distribution for predictand (Gebetsberger et al. 2017) or the choice of different parameter estimation methods (Gebetsberger et al. 2018), we found there are still four factors that need to be further investigated for precipitation forecasts in the above postprocessing schemes. 1) Traditionally, different transformations have been used in different postprocessing models, such as the normal

quantile transformation (NQT) in MGD, the log-sinh transformation in BJP, and the power transformation in logistic regression models. The relative performance of these transformations for daily precipitation forecasts is still unknown and will be investigated in this study. 2) The incorporation of an ensemble spread predictor in postprocessing models can help to quantify the flow-dependent uncertainty, but it may also lead to overfitting if the signal-to-noise ratio in raw forecasts is low (Scheuerer and Hamill 2015). We will test whether the incorporation of a spread predictor can improve the postprocessing results in our research region. 3) Two objective functions for parameter estimation (minimum CRPS or maximum likelihood) have been compared for temperature forecasts (Gebetsberger et al. 2018), but these two objective functions haven't been compared specifically for precipitation forecasts, which will be performed in this study. 4) Further comparison between the distributional regression scheme and the joint probability scheme is still needed. Although there are already some comparison works (Zhang et al. 2017), other models belong to the two schemes still need to be compared. In this study, we used the model named "censored regression with conditional heteroscedasticity" (CRCH; Messner et al. 2015) and a joint probability model similar to BJP as examples to compare the two schemes.

Therefore, we proposed four questions in this study as follows: 1) Will different transformations lead to different performance of postprocessed forecasts? 2) Will the ensemble spread predictor provide useful information to quantify forecast uncertainty in the research region? 3) Will different objective functions (minimum CRPS or maximum likelihood) influence the postprocessing performance for precipitation forecasts? 4) What are the strengths and weaknesses of the two postprocessing schemes, namely the distributional regression and the joint probability model?

To answer the above questions, we designed and conducted a series of comparison experiments for daily precipitation forecasts in Huai River basin in China. The structure of the paper is as follows. Section 2 describes the data, the experiment design and the models used in this paper. Section 3 presents the results. Section 4 discusses the four factors in the postprocessing of precipitation forecasts. Finally, section 5 summarizes the main conclusions.

2. Data and methods

a. Research region and data

The research region is the Huai River basin located in eastern China (30°55'–36°36'N, 111°55'–121°25'E).

FIG. 1. Illustration of the Huai River basin. The JJA-mean observed monthly precipitation is shown at the center of each subbasin.

The drainage area of the basin is about 270 000 km². The mean annual precipitation is appropriately 700–1600 mm. The Huai River basin is under the influence of the Asian monsoon system. The rainy season is during June–August. The Huai River basin (Fig. 1) is divided into 15 subbasins by the China Meteorological Administration. The characteristics of the 15 subbasins are shown in Table 1 (Liu et al. 2013).

The precipitation forecasts are from the 11-member Global Ensemble Forecast System (GEFS) version 2 from GEFS version 10 on a Gaussian grid of about 0.5° resolution (Hamill et al. 2013). The observations are the 0.25° × 0.25° precipitation analysis by optimal interpolation (Shen and Xiong 2016) from the China Meteorological Administration. The mean areal precipitation (MAP) of the forecasts and observations for each subbasin was calculated by averaging the forecasts or the observations of the grids within each subbasin.

b. Experimental design and verification methods

Our experiments included seven postprocessing models as shown in Table 2. First, we compared the influence of three transformations including a power transformation, log-sinh transformation, and normal quantile transformation (NQT) using the first-third CRCH models. As will be shown in the result section, NQT is generally suitable for the research region, so we

fixed NQT as the transformation and compared the other factors. Specifically, the comparison between the third and the fourth model, and the comparison between the fifth and sixth model are designed for the factor of incorporation of an ensemble spread predictor, i.e., the mean absolute difference (MD; see section 2c for details). The comparison between the third and the fifth model, and the comparison between the fourth and sixth model are designed for the comparison of different objective functions including minimum CRPS or maximum likelihood. Finally, we compared the above CRCH models with a joint probability model to compare two postprocessing schemes, namely the distributional regression and the joint probability model. The joint probability model only includes the ensemble mean predictor and is estimated by MLE, so the joint probability model can be compared with the third CRCH model to investigate the influence of the two postprocessing schemes. Details of these models will be described in sections 2c–e.

A 25-fold leave-one-year-out cross-validation was conducted for each of the 15 subbasins of the research region during the rainy season (June–August) using the 25-yr GEFS reforecasts and observations in 1985–2009. In other words, 24 years of data is used to fit postprocessing models and the left-out 1 year of data is used to verify the postprocessed forecasts each time.

TABLE 1. Main characteristics of the 15 subbasins of Huai River basin.

ID	Catchment name	Center	Area (10 ³ km ²)	Annual mean precipitation (mm)
A1	Dapoling upstream of Huaihe to Xixian catchment	32.31°N, 114.01°E	16.5	583
A2	Xixian upstream of Huaihe to Wangjiaba catchment	32.21°N, 115.02°E	8.8	622
A3	Ruhe and upstream of Honghe catchment	33.04°N, 114.12°E	9.5	683
B1	Shihe catchment	32.19°N, 115.73°E	10.6	827
B2	Upstream of Yinghe to Zhoukou catchment	34.07°N, 113.33°E	27.4	755
B3	Midstream of Yinghe and Zhoukou to Fuyang catchment	33.47°N, 114.99°E	14.3	640
B4	Pihe, downstream of Huaihe and Huaigan catchment	32.03°N, 116.29°E	11.4	643
B5	Wohe, midstream of Huaihe and Huaigan catchment	33.48°N, 116.04°E	28.7	827
C0	Bengbu to Hungtse, midstream and downstream of Huaigan and Huihe catchment	33.65°N, 117.40°E	42.3	889
D1	Hungtse to downstream of Huaihe catchment	33.18°N, 119.82°E	30.6	1056
D2	Downstream of Yihe and Shuhe catchment	34.41°N, 118.95°E	26.9	973
D3	Upstream of Shuhe catchment	35.61°N, 118.84°E	4.4	1105
D4	Upstream of Yihe catchment	35.62°N, 118.12°E	10.1	1114
D5	Zaozhuang and Xuzhou catchment	34.63°N, 117.78°E	9.2	1015
E0	Nansihu catchment	35.31°N, 116.32°E	30.8	947

That process is repeated 25 times. Then, the data to be verified are pooled together to calculate the verification metrics. The training dataset for each date was composed of a 45-day window centered on each date during the training years, thus a training dataset of 45 days $\times$ 24 years = 1080 days can be obtained. The postprocessed results were verified using the Ensemble Verification Service (EVS; Brown et al. 2010) software by pooling all the samples of the 15 subbasins together. Several verification metrics were used, including relative mean error, root-mean-square error (RMSE), continuous ranked probability skill score (CRPSS), Brier skill score (BSS), and relative operating characteristic (ROC) score. The sampling uncertainty for the verification metrics was estimated by the stationary block bootstrap technique (Politis and Romano 1994) in EVS.

The stratified probability integral transform (PIT) diagram was used in this study to show the reliability of different precipitation amounts. Specifically, all the samples for verification were stratified into three strata by the 85% and 95% quantiles of the raw forecast mean. The PIT histograms of the three strata of samples

represent the reliability of light, moderate and heavy rain. (The PIT histograms for each stratum are plotted, as will be shown in Fig. 8.) Note that the stratification should be based on the forecasts (e.g., raw forecasts mean as we used here) instead of the observations to ensure the reliability for each stratum (Bellier et al. 2017; Lerch et al. 2017). More details of the verification metrics are described in the appendix A and related references (Wilks 2011).

c. The CRCH model

As the distribution of daily precipitation amounts is skewed, a power transformation is applied to the forecasts and observations in the original CRCH before fitting the regression model. Then, the transformed observations y_t is assumed to follow a left-censored normal distribution in this study as follows:

$$y_t \sim \text{Normal}(\mu_t, \sigma_t), \quad (1)$$

$$\mu_t = \beta_0 + \beta_1 \bar{x}_t, \quad \text{and} \quad (2)$$

$$g(\sigma_t) = \gamma_0 + \gamma_1 g(\text{SD}_{x_t}), \quad (3)$$

TABLE 2. Experiment design. Censored regression with conditional heteroscedasticity (CRCH), normal quantile transform (NQT), maximum likelihood estimation (MLE), continuous ranked probability score (CRPS), and mean absolute difference (MD).

	Name of the model	Transformation	Objective function	Predictor of ensemble spread
1	CRCH log-sinh MLE	Log-sinh	Likelihood	Not incorporated
2	CRCH power MLE	Power	Likelihood	Not incorporated
3	CRCH NQT MLE	NQT	Likelihood	Not incorporated
4	CRCH NQT MLE MD	NQT	Likelihood	Incorporated
5	CRCH NQT CRPS	NQT	CRPS	Not incorporated
6	CRCH NQT CRPS MD	NQT	CRPS	Incorporated
7	Joint probability model	NQT	Likelihood	Not incorporated

where x_t is the transformed ensembles at time t , $\bar{x}_t$ is the mean of the transformed ensembles x_t , SD_{x_t} is the standard deviation of the transformed ensemble forecasts, and g is the link function for the scale parameter. The transformed observations less than a threshold are considered as left censored to deal with the intermittency nature of precipitation (Messner et al. 2014a). A threshold of 0.1 mm day^{-1} is used because 0.1 mm is the minimum measurable rainfall amount for most of the rain gauges in China. Equation (3) is generally similar to the model described in Gebetsberger et al. (2017), except that we used normal distribution instead of logistic distribution, because we used normalization transformations with parameters that estimated by MLE to ensure the transformed variates to follow normal distribution in this study.

In this study, we used the mean absolute difference (MD) of the ensemble members as a measure of the ensemble dispersion, which is able to robustly quantify the forecast dispersion (Scheuerer and Hamill 2015). The MD is defined as follows:

$$MD_{x_t} = \frac{1}{m^2} \sum_{j,j'=1}^m |x_t^j - x_t^{j'}|, \quad (4)$$

where m is the ensemble size, x_t^j and $x_t^{j'}$ are two ensemble members at time t .

We chose a quadratic function as the link function g in Eq. (3) according to our previous experiments. Then the scale submodel in Eq. (3) becomes

$$\sigma_t^2 = \gamma_0 + \gamma_1 \cdot (MD_{x_t})^2. \quad (5)$$

As we are interested in whether the incorporation of the ensemble spread predictor leads to improvements, we compared the scale submodel above with a constant-scale submodel, namely

$$\sigma_t^2 = \gamma_0. \quad (6)$$

In this case, the variance of the regression model in transformed space becomes constant. As shown in Table 2, we used the constant-scale submodel for the first three models and the fifth model. For the fourth and sixth models, we used the MD as the predictor in the scale submodel as shown in Eq. (5).

The parameters in the CRCH models can be obtained by minimum CRPS or maximum likelihood by the R package “crch” (Messner et al. 2015). Given a new forecast, 100 evenly spaced quantiles of the conditional distribution of the observations are generated from the CRCH models in the transformed space. Then, the 100 quantiles are transformed back to the original

space and are used as calibrated ensemble forecasts. The forecasts that are less than the threshold of 0.1 mm day^{-1} are transformed to a value that equals to the transformed value of the threshold during the fitting and prediction procedure for the CRCH models. The Schaake shuffle (Clark et al. 2004) can be applied to the ensembles to preserve the spatiotemporal correlation for further application such as hydrological forecasting.

d. The CRCH model with different transformations

As we aimed to compare the influence of different transformations, we implemented three transformations including power transformation, NQT, and log-sinh transformations in CRCH models.

A parametric NQT for forecasts z is defined as follows:

$$x = \Phi_N^{-1}[F_z(z)], \quad (7)$$

where F_z is the cumulative distribution function (CDF) of the marginal distribution of the forecasts z in the untransformed space; Φ_N^{-1} is the inverse of the CDF of the standard Normal distribution. In this study, we used a mixed-type distribution for the marginal distribution of forecasts as follows:

$$F_z(z) = p_0 + (1 - p_0)F_{z|z>0}(z), \quad (8)$$

where p_0 is the probability of the precipitation amount that is less than or equal to the threshold of 0.1 mm day^{-1} ; $F_{z|z>0}$ is the CDF of the precipitation amount given the occurrence of precipitation. The distribution for the nonzero precipitation amount is chosen as Weibull distribution according to our previous experiments in this research region (Li et al. 2019). The NQT for the observations is similar and is omitted here.

The log-sinh transformation for forecasts z is defined as follows (Wang et al. 2012):

$$x = \frac{1}{\lambda} \log[\sinh(\varepsilon + \lambda z)]. \quad (9)$$

The two parameters ε and λ are estimated by maximum likelihood estimation (MLE). The data less than or equal to the threshold of 0.1 mm day^{-1} is considered as censored data during MLE (Wang and Robertson 2011; Wang et al. 2012). The log-sinh transformation for the observations is similar and is omitted here.

The power transformation for forecasts z is as follows:

$$x = z^p. \quad (10)$$

The power parameter is fitted by MLE, similar to the method used for log-sinh transformation. The data less

than or equal to the threshold of 0.1 mm day^{-1} is considered as censored data during MLE, similar to the estimation method used for log-sinh transformation. Detail of the likelihood function is shown in [appendix B](#). The power transformation for the observations is similar and is omitted here.

The parameters of the NQT, log-sinh, and power transformations are fitted for observations and forecasts separately. For the CRCH models, we fitted the transformations for forecasts by pooling all the ensemble members together, and then applied the transformations to each of the ensemble members. For the joint probability model, we fitted the transformation for the ensemble mean directly, because only the ensemble mean is used in the joint probability model.

e. The joint probability model

The joint probability model used in this study is similar to the BJP ([Robertson et al. 2013](#); [Shrestha et al. 2015](#); [Wang and Robertson 2011](#); [Wang et al. 2009](#)). The main difference between the joint probability model in this study and the original BJP is we used maximum likelihood estimation (MLE) here to obtain the parameters of the model, instead of the Bayesian inference in the original BJP. As NQT has been found to perform generally well in our research region from previous experiments, we also applied the NQT to the forecast mean and observations in the joint probability model. Then, the joint distribution of the transformed forecast mean x and observations y is assumed to be a bivariate normal distribution:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \right). \quad (11)$$

The forecasts or observations less than the censoring threshold are treated as censored data during the parameter estimation ([Shrestha et al. 2015](#); [Wang and Robertson 2011](#); [Wang et al. 2009](#)), similar to the censoring technique for the predictand in CRCH models. After obtaining the parameters of the joint probability distribution, the conditional distribution of the observations given a new forecast can be obtained:

$$p(y|x) \sim N \left[\mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x), (1 - \rho^2) \sigma_y^2 \right]. \quad (12)$$

A large number (here we use 1000) of random samples can be generated from the conditional distribution and then transformed back to the original space. Then we calculated 100 evenly spaced quantiles from the 1000 random samples to generate a 100-member

ensemble forecasts. If the new forecast is less than the threshold, we used the “data augmentation” method, that is, to draw a random sample that is equal to or less than the transformed threshold from the marginal distribution of the ensemble mean forecasts, and then use the random sample to substitute x in Eq. (12) and get one sample of y from the conditional distribution ([Robertson et al. 2013](#); [Shrestha et al. 2015](#); [Wang et al. 2009](#)). This process is repeated to get all the postprocessed ensemble members. For more details of the joint probability model, please refer to related references ([Robertson et al. 2013](#); [Shrestha et al. 2015](#); [Wang et al. 2009](#)).

3. Results

In this section, the verification results of 15 subbasins together are briefly described in [section 3a](#). Then, the detailed results of each of the 15 subbasins are described in [sections 3b–3e](#), corresponding to the four scientific questions mentioned in the introduction. The postprocessed forecast intervals for one subbasin (B1) in Huai River basin are shown as an example in [section 3f](#).

a. Overall verification results for all 15 subbasins together

Verification results by pooling all the samples from 15 subbasins together are shown in [Figs. 2–5](#). In this subsection, only a general description of the results is given. More details will be further described in the following four subsections focusing on each of the four methodological factors.

The relative mean error results in [Fig. 2a](#) show that raw GEFS forecasts (black bars) suffer from obvious overestimation. All seven postprocessing methods achieve unbiased results in general and reduce the RMSE relative to that of raw forecasts.

[Figure 3](#) shows the CRPSS of the results of the seven postprocessing methods. There are obvious differences between the CRPSS of some of these postprocessing methods. For example, the fourth and sixth CRCH models with the MD predictor (dark red or dark green bars) outperform over the third or fifth models without the MD predictor (light red or light green bars) at lead times of 1–5 days. The best models among the seven postprocessing methods are the sixth CRCH model with the MD predictor and estimated by CRPS minimization (dark green bars), followed by the joint probability model (purple bars).

[Figure 4](#) shows the verification results of ROC scores at four thresholds, namely 75%, 85%, 95%, and 97.5% quantiles of observations. The relative performance of

FIG. 2. Deterministic verification for the raw GEFS reforecasts and the postprocessed results by the seven models. (a) Relative mean error; (b) root-mean-square error. The 90% confidence intervals by bootstrapping are shown for each metric.

seven methods generally follows a similar pattern with that of Fig. 3. The difference between the ROC scores of seven models is more remarkable for heavy rain events (Figs. 4c,d).

Figure 5 shows the stratified PIT histograms for the seven postprocessing models at the lead time of three days. The three rows in Fig. 5 are the stratified PIT histograms for light rain, moderate rain, and heavy rain, respectively. Figure 5 shows that the reliability for different precipitation amounts is different. The PIT histograms of the light rain (the first row) for the seven postprocessed models are generally flat, which means all the postprocessing models are able to achieve reliable forecasts for light rain. For the moderate rain (the second row), the first CRCH model with log-sinh transformation suffers from overdispersion (“∩” shape). For the heavy rain, the results of the CRCH models that only include ensemble mean predictors (Figs. 5o–q,s) suffer from slight overestimation (downward “∪” shape), while the CRCH models include ensemble spread as a predictor are still reliable. The stratified PIT histograms for other lead times are generally similar to Fig. 5 and are presented in the online supplemental material.

b. The results for different normalization transformations

Figure 6 shows the quantile–quantile plots of the forecasts or observations using three normalization transformations at a relatively dry subbasin B1 (Figs. 6a,b) and a relatively wet subbasin D1 (Figs. 6c,d) in Huai River basin at lead time of one day. As shown in Fig. 6, the results of NQT (blue dots) and power transformation (red crosses) are quite similar and almost overlap

with each other. The quantiles of log-sinh transformation (green circles) are obviously higher than those of the other two transformations at the upper tail. As shown in Figs. 6a and 6b, the transformed variates by NQT and power transformation are both closer to the diagonal line than those by log-sinh transformation in the relatively dry subbasin B1. The transformed variates by log-sinh transformation suffer from overestimation at the upper tail in subbasin B1. On the contrary, the transformed variates by log-sinh transformation are closer to the diagonal line than those by the other two transformations at the upper tail in the relatively wet subbasin D1, as shown in Figs. 6c and 6d. The transformed variates by NQT or power transformation suffer from slight underestimation at the upper tail in subbasin D1.

FIG. 3. The continuous ranked probability skill score (CRPSS) of the postprocessed results by the seven models. The 90% confidence intervals by bootstrapping are shown for the metric.

FIG. 4. The relative operating characteristic (ROC) score for the postprocessed results by the seven models at thresholds of (a) 75%, (b) 85%, (c) 95%, and (d) 97.5% quantiles of observations. The 90% confidence intervals by bootstrapping are shown for each metric.

As shown in section 3a, the log-sinh transformation performs slightly worse than the other two transformations in terms of CRPSS (Fig. 3), but performs better than the other two transformations in terms of ROC score (Fig. 4) in the study region. We mainly focused on the relative performance of NQT and log-sinh transformation here, and showed the Brier skill score for the results by the CRCH with NQT relative to the results of the CRCH with log-sinh transformation (the latter is used as reference forecasts) at each of the 15 subbasins at three thresholds (85%, 95%, and 97.5% quantiles of the observations) in Fig. 7. As shown in Fig. 7, the results of the CRCH model with NQT are better than those of the CRCH model with log-sinh transformation for most of the subbasins and lead times, especially at relatively dry subbasins (e.g., subbasin A1–B4 in blue color). The CRCH model with

log-sinh transformation is better than the CRCH model with NQT at a few relatively wet subbasins (e.g., C0 and D1 in red color) at the lead time of one day, which is in accordance with the quantile–quantile plots in Fig. 6. For longer lead times the differences between the two methods gradually diminish. The results for the CRCH model with power transformation relative to log-sinh transformation are similar and not shown here.

The results indicate that NQT or power transformation performs better than log-sinh transformation in most of the subbasins in Huai River basin. However, there are also a few relatively wet subbasins where log-sinh transformation is more suitable than the other two transformations. The results show that the performance of these three transformations depends on the data of specific locations. The suitable transformations

FIG. 5. The stratified PIT histograms of the forecasts by seven postprocessing models at the lead time of three days. (a)–(g) PIT histograms corresponding to the samples with raw forecast mean within the 0%–85% quantiles (light rain), (h)–(n) 85%–95% quantiles (moderate rain), and (o)–(u) 95%–100% quantiles (heavy rain).

for other basins under different climate still need to be further tested.

c. The results for the incorporation of ensemble spread as a predictor

As shown in the results in section 3a, the fourth and sixth CRCH models with MD predictor perform better than the third or fifth models without MD predictor in terms of CRPS (Fig. 3), ROC score (Fig. 4) and reliability (Fig. 5) mainly at lead times of 2–4 days. Figure 8 shows the Brier skill score for the CRCH with ensemble spread relative to the CRCH without ensemble spread (the latter is used as reference forecasts) at each of the 15 subbasins. The results in Figs. 8a–c are for the third–fourth CRCH models by MLE. The results in Figs. 8d–f are for the fifth–sixth CRCH models by CRPS minimization. Figure 8 shows that the improvements (blue color) by incorporation of ensemble spread in CRCH exist in most of the 15 subbasins, especially at lead times of 1–3 days for heavy rain. In summary, the incorporation of ensemble spread as predictors in CRCH models will improve the forecast quality in terms of accuracy, discrimination and reliability in most of the subbasins in the research region at short lead times.

d. The results for the objective functions for parameter inference

As mentioned previously, the comparison between the results of the third and fifth CRCH models, as well as the comparison between the fourth and sixth CRCH models exhibit whether CRCH models by CRPS minimization outperform the models by MLE. From the results shown in section 3a, the CRCH models by CRPS minimization improve the results over the CRCH models by MLE mainly for lead times of 2–5 days in terms of CRPS (Fig. 3) and ROC score (Fig. 4), but the improvement is very marginal. The difference between those two groups of CRCH models in terms of reliability is also not significant (Fig. 5).

Figure 9 shows the Brier skill score for the post-processed results of the CRCH models by CRPS minimization relative to the CRCH models by MLE (the latter is used as reference forecasts) at each of the 15 subbasins. The results in Figs. 9a–c are for the third and fifth CRCH models with ensemble mean as the only predictor. The results in Figs. 9d–f are for the fourth and sixth CRCH models with ensemble spread as predictors. Figure 9 shows that the improvements by CRPS minimization in CRCH mainly exist in a few

FIG. 6. The quantile–quantile plots for the transformed forecasts and observations by three normalization transformations at a (a),(b) relatively dry subbasin B1 and (c),(d) relatively wet subbasin D1 in the Huai River basin at a lead time of one day. The black solid lines are the diagonal lines.

relatively dry subbasins (subbasins A1–B2 in blue color) for lead times of 1–5 days, while no remarkable improvements are shown in other relatively wet subbasins. In summary, the advantage of CRPS minimization over MLE mainly exists in some of the subbasins with a relatively dry climate, but not in all of the subbasins in Huai River basin. That can explain why the improvements in the verification results by pooling all the samples of the 15 subbasins together are not significant in [section 3a](#).

e. The results for two postprocessing schemes

The results of two postprocessing schemes, namely the joint probability model and CRCH model are compared in this subsection. The joint probability model is comparable with the third CRCH model, because both models do not include ensemble spread predictors and are both estimated by MLE. As shown in [section 3a](#), the joint probability model outperforms the third CRCH

model (light red color) for most of the verification metrics such as RMSE ([Fig. 2b](#)), CRPS ([Fig. 3](#)), and ROC score ([Fig. 4](#)). On the other hand, the joint probability model is still worse than the best CRCH model, namely the sixth CRCH model for short lead times, but the difference gradually vanishes with the increase of lead times.

[Figures 10a–c](#) shows the Brier skill score for the postprocessed results of the third CRCH model relative to the joint probability model (the latter is used as reference forecasts) at each of the 15 subbasins. The third CRCH model is obviously worse than the joint probability model for most of the subbasins and lead times, except for a few subbasins (e.g., subbasins B5 and D5) at a lead time of one day. [Figures 10d–f](#) shows the Brier skill score for the postprocessed results of the sixth CRCH model (with ensemble spread as predictors, estimated by CRPS minimization) relative to the joint probability model (the latter is used as

FIG. 7. The Brier skill score for the postprocessed results by the CRCH with NQT relative to the CRCH with log-sinh transformation at each of the 15 subbasins for three thresholds (85%, 95%, and 97.5% quantiles of the observations). The results of the CRCH with log-sinh transformation are used as the reference forecasts. A blue color indicates the CRCH with NQT is better than the CRCH with log-sinh transformation.

reference forecasts). The sixth CRCH model is better than the joint probability model for several subbasins (e.g., subbasins B3, B5, and D5) mainly at short lead times.

The results show that the joint probability model outperforms the CRCH model if both use ensemble mean as the only predictor and are estimated by MLE. However, the advantages of the CRCH scheme are that it is able to include ensemble spread predictors and the parameters can be estimated by CRPS minimization,

which leads to further improvements over the joint probability model for short lead times.

f. Postprocessed forecast intervals for one subbasin

Figure 11 shows the forecast intervals of the seven postprocessing models against the raw forecast median at the lead time of one day in subbasin B1. The forecast intervals of raw forecasts are too narrow, with many observations (red dots) falling out of the 90% forecast intervals (blue bars). The forecast intervals of the

FIG. 8. The Brier skill score for the postprocessed results by the CRCH with ensemble spread relative to the CRCH without ensemble spread at each of the 15 subbasins for three thresholds (85%, 95%, and 97.5% quantiles of the observations). The results of the CRCH models without ensemble spread are used as the reference forecasts. (a)–(c) The third or fourth CRCH models estimated by MLE; (d)–(f) The fifth or sixth CRCH models estimated by minimization of CRPS. A blue color means the CRCH models with ensemble spread are better than those without ensemble spread.

FIG. 9. The Brier skill score for the postprocessed results of the CRCH model by CRPS minimization relative to the CRCH model by MLE at each of the 15 subbasins for three thresholds (85%, 95%, and 97.5% quantiles of the observations). The results of the CRCH models by MLE are used as the reference forecasts. (a)–(c) The third or fifth CRCH model that only includes ensemble mean as predictors; (d)–(f) The fourth or sixth CRCH model that includes ensemble mean and spread as predictors. A blue color means the CRCH models using minimization of CRPS are better than those using MLE.

postprocessed results are all wider than those of raw forecasts to include most of the observations within the 90% forecast intervals. The forecast intervals of the CRCH model with log-sinh (Fig. 11b) are narrower than those of the CRCH model with power transformation or NQT (Figs. 11c,d) for large events, but more observations fall lower than the 90% forecast intervals when the raw forecasts (x axis) are extremely high, which indicates the CRCH model with log-sinh suffers from overestimation problem for these large events.

The widths of all the postprocessed forecast intervals generally increase with the raw forecast median, which exhibits an increasing forecast uncertainty with the increase of the forecast amounts. Moreover, the interval widths of the CRCH models with spread predictors of MD (Figs. 11e,g) exhibit similar fluctuations to the interval widths of the raw forecasts (Fig. 11a). In other words, the incorporation of ensemble spread as a predictor makes the CRCH models be able to characterize the flow-dependent uncertainty represented by the ensemble spread of the raw forecasts. The difference between the forecast intervals of CRCH models by MLE (Figs. 11d,e) and those by minimum CRPS (Figs. 11f,g) is not obvious. The forecast intervals of the joint probability model (Fig. 11h) are generally similar to those of the

CRCH model without an ensemble spread predictor (Fig. 11d).

4. Discussion

According to the results in the previous section, all seven postprocessing models improve the forecast skill over the raw GEFS reforecasts and generally achieve reliable forecasts. Detailed discussions about the four factors in regression-based postprocessing models are as follows.

a. Normalization transformations

We investigated three normalization transformations in this study, including NQT, power transformation and log-sinh transformation. It is found that the results of CRCH models with NQT and power transformation are quite similar, while the log-sinh transformation generates transformed values with heavier upper tail than the other transformations, which leads to slightly worse performance than the CRCH models with other two transformations for most of the 15 subbasins in Huai River basin. However, there are still several subbasins in the Huai River basin where log-sinh performs better than the other two transformations. Log-sinh transformation is originally designed to stabilize the variance of streamflow variables. In such

FIG. 10. (a)–(c) The Brier skill score for the postprocessed results by the third CRCH (using MLE and only ensemble mean) relative to the joint probability model at each of the 15 subbasins for three thresholds (85%, 95%, and 97.5% quantiles of the observations). (d)–(f) The Brier skill score for the postprocessed results by the sixth CRCH (using CRPS minimization and ensemble spread) relative to the joint probability model at each of the 15 subbasins. The results of the joint probability model are used as the reference forecasts for both cases.

cases, the variance of the original streamflow variables often first increases rapidly, then slowly approaches a constant for large events (Wang et al. 2012). However, the variance of precipitation forecasts may not satisfy

that assumption, which might lead to the slightly inferior performance of log-sinh transformation in this study. Further comparison of log-sinh transformation with other transformations for precipitation forecasts

FIG. 11. The 50% and 90% forecast intervals for the seven models for subbasin B1 at the lead time of one day.

still needs to be conducted for regions under different climates.

b. The incorporation of ensemble spread as a predictor

The results in our experiments show that the incorporation of an ensemble spread as a predictor for the scale parameter of the predictive distribution improves the forecast performance for most of the subbasins in the Huai River basin at short lead times when forecast skill is relatively high. Ensemble spread has been used as predictors of predictive variance in postprocessing models since the EMOS model proposed by [Gneiting et al. \(2005\)](#). However, the influence of the incorporation of ensemble spread as a predictor might depend on specific regions and seasons. [Messner et al. \(2014b\)](#) found that the incorporation of ensemble spread as a predictor for the scale parameter of the predictive distribution in an extended logistic regression model led to improvements in most of the European meteorological stations, but there were also stations where improvement was not obvious (e.g., Berlin station in Fig. 5 of [Messner et al. 2014b](#)). [Scheuerer and Hamill \(2015\)](#) found that the incorporation of ensemble spread predictor improved the results during more predictable cool seasons but didn't improve the results during the less predictable warm seasons, and even deteriorated the results at long lead times in the contiguous United States. They speculated that the reason might be overfitting when the raw forecast skill is limited ([Scheuerer and Hamill 2015](#)). Further experiments on this issue are still needed in other research regions.

c. The objective functions for parameter inference

We compared the results by CRCH models using different objective functions including CRPS minimization and MLE. The results show that both objective functions can improve the forecast skill and generate reliable forecasts in terms of PIT histograms. The CRCH models by CRPS minimization perform slightly better than those by MLE in CRPS evaluation and some of the other skill scores at a few relatively dry subbasins, but the overall improvement in the Huai River basin is not remarkable. [Gebetsberger et al. \(2018\)](#) investigated the influence of objective functions in nonhomogeneous regression models for postprocessing of temperature forecasts by both theoretical derivation and simulation experiments. Their results showed that these two objective functions should lead to similar results if the distribution of the predictand is appropriately specified. More detailed discussion about that issue can be found in their paper.

d. Two postprocessing schemes

Two postprocessing schemes, namely the distributional regression and the joint probability model, are compared in this study. The CRCH models are used as examples of the distributional regression scheme and are compared with a joint probability model similar to BJP. One difference between the two schemes is that the intermittent nature of precipitation is treated differently. As mentioned in the method section, both the forecasts and the observations below a threshold are treated as censored data in the joint probability model. However, only the observations (i.e., the predictand) are treated as censored in distributional regression models and the near-zero forecasts are usually not specially treated. [Gebetsberger et al. \(2017\)](#) found that the regression relationship might collapse in traditional distributional regression models when the raw precipitation forecasts are near to zero. Therefore, the distributional regression models used in this study may not perform as well as the joint probability model when the raw forecasts are near to zero. That can be a possible reason why the performance of the joint probability model is better than the third CRCH model that only uses the ensemble mean as a predictor and estimated by MLE. On the other hand, an advantage of the distributional regression scheme relative to the joint probability model is that the former is more convenient to incorporate ensemble spread predictors to quantify the flow-dependent forecast uncertainty.

Besides the models included in this experiment, there are still other models for postprocessing of short-term precipitation forecasts, such as CSGD EMOS or generalized extreme value distribution-based EMOS ([Scheuerer 2014](#)), and kernel dressing models ([Hamill and Scheuerer 2018](#)). Although there are already some comparisons ([Wilks 2006](#); [Wilks and Hamill 2007](#); [Williams et al. 2014](#)), further experiments are still needed to compare these newly developed postprocessing models.

The conclusions in this study are mainly limited to basins with a semihumid climate similar to the Huai River basin. Further comparison of these methods in other basins with different climate is still needed. Moreover, it should be noted that the above comparison results are obtained by fitting the model using a sufficient long archive of reforecasts (24 years of training data in each cross-validation process). We also did comparison experiments using only five years of forecasts to fit the postprocessing models. We found the difference between the results of different models will become smaller in that case, probably because there are

much less extreme events to fit the model by only five years of dataset, which diminishes the differences between the performances of postprocessing methods for extreme events.

5. Conclusions

In this study, we considered four factors in regression-based postprocessing models for short-term precipitation forecasts including: 1) normalization transformations, 2) the incorporation of an ensemble spread predictor in postprocessing models, 3) the objective function for parameter inference, and 4) different postprocessing schemes. Variants of the CRCH models are used to investigate the first three factors. For the fourth factor, we chose the CRCH model as an example of the distributional regression scheme and compared it with a joint probability model similar to BJP. The main conclusions are as follows:

- 1) The performance of different normalization transformations depends on the data in specific regions. The normal quantile transformation and the power transformation are generally more suitable than log-sinh transformation for most of the subbasins in Huai River basin with a semihumid climate. However, log-sinh transformation may perform better than the other two transformations in some relatively wet subbasins. The suitable transformations for other regions or seasons under different climate still need to be further tested.
- 2) The CRCH models that incorporate ensemble spread predictors are able to better quantify the flow-dependent forecast uncertainty and outperform the CRCH models without spread predictors in most of the subbasins in the Huai River basin, especially for short lead times.
- 3) The CRCH models by CRPS minimization achieve slightly better accuracy than those by MLE, but the improvements mainly exist in a few relatively dry subbasins in our research region. Both objective functions can lead to generally reliable and skillful forecasts in our research region.
- 4) The two postprocessing schemes both have their advantages. The joint probability model is better than the distributional regression scheme if both use the ensemble mean as the only predictor and use MLE for parameter estimation for most of the subbasins in the Huai River basin. The distributional regression scheme is more convenient to incorporate the ensemble spread predictors to quantify the flow-dependent forecast uncertainty and improves the performance at short lead times.

Acknowledgments. We are grateful to the valuable comments from the editor and anonymous reviewers. The study is supported by the National Basic Research Program of China (2015CB953703), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA2006040104), and the National Key Research and Development Program of China (2018YFE0196000). The first author is supported by China Scholarship Council.

APPENDIX A

The Verification Metrics

a. Relative mean error

The relative mean error measures the average difference between the ensemble mean forecast $\bar{f}_i$ and corresponding observation o_i as a fraction of the average observation:

$$\text{Relative mean error} = \frac{\sum_{i=1}^n (\bar{f}_i - o_i)}{\sum_{i=1}^n o_i}, \quad (\text{A1})$$

where n is the total number of forecast–observation pairs.

b. Root-mean-square error

The root-mean-square error (RMSE) is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (f_i - o_i)^2}. \quad (\text{A2})$$

c. Continuous ranked probability skill score

The continuous ranked probability score (CRPS) measures the integral square difference between the cumulative distribution functions (CDF) of a forecast and corresponding CDF of observation as follows:

$$\text{CRPS} = \int_{-\infty}^{\infty} [F_f(q) - F_o(q)]^2 dq. \quad (\text{A3})$$

Then, the mean CRPS of all samples can be computed from the average of the CRPS of all the forecast and observation pairs. The continuous rank probability skill score (CRPSS) can be computed as follows:

$$\text{CRPSS} = 1 - \frac{\overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{ref}}}, \quad (\text{A4})$$

where $\overline{\text{CRPS}}$ is the mean CRPS of the forecasts; $\overline{\text{CRPS}}_{\text{ref}}$ is the mean CRPS for the reference. The climatology is chosen as the reference in this study.

d. Brier skill score

The Brier score measures the mean square error of the forecast probabilities of the meteorological variables exceeding a threshold

$$BS = \frac{1}{n} \sum_{i=1}^n [F_{f_i}(q) - F_{o_i}(q)]^2, \quad (A5)$$

where $F_{f_i}(q)$ is the probabilities of the i th forecast exceeding the threshold q , and $F_{o_i}(q)$ is the corresponding binary observation depending on whether the observation o_i exceeds the threshold as follows:

$$F_{o_i}(q) = \begin{cases} 1, & o_i > q \\ 0, & \text{otherwise} \end{cases}. \quad (A6)$$

The Brier skill score (BSS) measures the relative improvement of the Brier score of the main forecast system over that of a reference system (e.g., the climatology)

$$BSS = 1 - \frac{BS}{BS_{\text{ref}}}. \quad (A7)$$

The BSS is positively oriented. The BSS for the perfect forecast is one, while the BSS is less than zero for forecasts with no skill relative to the reference.

e. Relative operating characteristic score

Forecast discrimination can be by the ROC diagram, which plots the hit rate (HR) versus the false alarm rate (FAR) for a range of probability thresholds. The ROC diagram can be summarized by the area under the ROC curve (AUC). Then, the ROC score can be computed as follows:

$$\text{ROC score} = \frac{\text{AUC} - \text{AUC}_{\text{ref}}}{1 - \text{AUC}_{\text{ref}}}, \quad (A8)$$

where AUC_{ref} is the AUC of the reference forecasts. The ROC score is positively oriented.

f. The PIT diagram

The probability integral transform (PIT) is the value of the predictive cumulative distribution function (CDF) F_f at the corresponding observation o_i as follows:

$$\text{PIT}_i = F_f(o_i). \quad (A9)$$

The PIT for reliable forecasts follows a uniform distribution. The reliability can be checked by plotting the empirical CDF of the PIT values against the CDF of the uniform distribution. For reliable forecasts, the PIT values should align along the diagonal line. PIT diagram can be used to diagnose the over/underestimation

over/underdispersion problems (Laio and Tamea 2007; Thyer et al. 2009). When the observations are equal or below the censoring threshold (0.1 mm day^{-1} in this study), a pseudo PIT value is generated from a uniform distribution with the range of $[0, F_f(y_c)]$, where y_c is the censoring threshold for observation (Robertson et al. 2013).

APPENDIX B

Likelihood Function for Power Transformation

The likelihood function for power transformation for forecasts can be divided into two cases depending on whether the forecasts are larger than the censoring threshold in the original space z_c as follows given a data series with n forecast–observation pairs:

$$L(z; p, \mu, \sigma) = \prod_{i=1}^n l(z_i), \quad (B1)$$

$$l(z_i) = \begin{cases} \phi_N(z_i^p; \mu, \sigma)(z_i^p)', & \text{if } z_i > z_c \\ \Phi_N(z_i^p; \mu, \sigma), & \text{if } z_i \leq z_c \end{cases}, \quad (B2)$$

where ϕ_N and Φ_N are the density and CDF of the distribution of transformed forecasts, respectively; μ and σ are the mean and standard deviation parameter of the transformed variable. The likelihood function of power transformation for observations is similar and is omitted here.

REFERENCES

- Bellier, J., I. Zin, and G. Bontron, 2017: Sample stratification in verification of ensemble forecasts of continuous scalar variables: Potential benefits and pitfalls. *Mon. Wea. Rev.*, **145**, 3529–3544, <https://doi.org/10.1175/MWR-D-16-0487.1>.
- Boucher, M. A., L. Perreault, F. O. Ancil, and A. C. Favre, 2015: Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts. *Hydrol. Processes*, **29**, 1141–1155, <https://doi.org/10.1002/hyp.10234>.
- Brown, J. D., J. Demargne, D.-J. Seo, and Y. Liu, 2010: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Modell. Software*, **25**, 854–872, <https://doi.org/10.1016/j.envsoft.2010.01.009>.
- Buizza, R., P. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, <https://doi.org/10.1175/MWR2905.1>.
- Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake Shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.*, **5**, 243–262, [https://doi.org/10.1175/1525-7541\(2004\)005<0243:TSSAMF>2.0.CO;2](https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2).
- Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *J. Hydrol.*, **375**, 613–626, <https://doi.org/10.1016/j.jhydrol.2009.06.005>.

- Cuo, L., T. C. Pagano, and Q. J. Wang, 2011: A review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting. *J. Hydrometeorol.*, **12**, 713–728, <https://doi.org/10.1175/2011JHM1347.1>.
- Duan, Q., F. Pappenberger, A. Wood, H. L. Cloke, and J. Schaake, 2019: *Handbook of Hydrometeorological Ensemble Forecasting*. Springer, 1528 pp.
- Fortin, V., A.-C. Favre, and M. Said, 2006: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quart. J. Roy. Meteor. Soc.*, **132**, 1349–1369, <https://doi.org/10.1256/qj.05.167>.
- Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190–202, <https://doi.org/10.1175/2009MWR3046.1>.
- Friederichs, P., and A. Hense, 2007: Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Wea. Rev.*, **135**, 2365–2378, <https://doi.org/10.1175/MWR3403.1>.
- Gebetsberger, M., J. W. Messner, G. J. Mayr, and A. Zeileis, 2017: Fine-tuning nonhomogeneous regression for probabilistic precipitation forecasts: Unanimous predictions, heavy tails, and link functions. *Mon. Wea. Rev.*, **145**, 4693–4708, <https://doi.org/10.1175/MWR-D-16-0388.1>.
- , —, —, and —, 2018: Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Mon. Wea. Rev.*, **146**, 4323–4338, <https://doi.org/10.1175/MWR-D-17-0364.1>.
- Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, [https://doi.org/10.1175/1520-0450\(1972\)011<1203:TUOMOS>2.0.CO;2](https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2).
- Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, <https://doi.org/10.1146/annurev-statistics-062713-085831>.
- , A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.
- Hamill, T. M., and Coauthors, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, <https://doi.org/10.1175/BAMS-D-12-00014.1>.
- , and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079–4098, <https://doi.org/10.1175/MWR-D-18-0147.1>.
- , and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, <https://doi.org/10.1175/MWR3237.1>.
- Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.*, **16**, 672–682, [https://doi.org/10.1175/1520-0469\(1959\)016<0672:OPOFDM>2.0.CO;2](https://doi.org/10.1175/1520-0469(1959)016<0672:OPOFDM>2.0.CO;2).
- Klein, N., T. Kneib, S. Lang, and A. Sohn, 2015: Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Stat.*, **9**, 1024–1052, <https://doi.org/10.1214/15-AOAS823>.
- Krzysztofowicz, R., and W. B. Evans, 2008: Probabilistic forecasts from the National Digital Forecast Database. *Wea. Forecasting*, **23**, 270–289, <https://doi.org/10.1175/2007WAF2007029.1>.
- Laio, F., and S. Tamea, 2007: Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrol. Earth Syst. Sci.*, **11**, 1267–1277, <https://doi.org/10.5194/hess-11-1267-2007>.
- Leirich, S., T. L. Thorarindottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: Extreme events and forecast evaluation. *Stat. Sci.*, **32**, 106–127, <https://doi.org/10.1214/16-STS588>.
- Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di, 2017: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdiscip. Rev.: Water*, **4**, e1246, <https://doi.org/10.1002/wat2.1246>.
- , —, A. Ye, and C. Miao, 2019: An improved meta-Gaussian distribution model for post-processing of precipitation forecasts by censored maximum likelihood estimation. *J. Hydrol.*, **574**, 801–810, <https://doi.org/10.1016/j.jhydrol.2019.04.073>.
- Liu, Y., and Coauthors, 2013: Evaluating the predictive skill of post-processed NCEP GFS ensemble precipitation forecasts in China's Huai River basin. *Hydrol. Processes*, **27**, 57–74, <https://doi.org/10.1002/hyp.9496>.
- Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014a: Extending extended logistic regression: Extended vs. separate vs. ordered vs. censored. *Mon. Wea. Rev.*, **142**, 3003–3013, <https://doi.org/10.1175/MWR-D-13-00355.1>.
- , —, A. Zeileis, and D. S. Wilks, 2014b: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, <https://doi.org/10.1175/MWR-D-13-00271.1>.
- , —, and —, 2015: Heteroscedastic censored and truncated regression with crch. *R J.*, **8**, 1–12.
- Park, Y. Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.*, **134**, 2029–2050, <https://doi.org/10.1002/qj.334>.
- Politis, D. N., and J. P. Romano, 1994: The stationary bootstrap. *J. Amer. Stat. Assoc.*, **89**, 1303–1313, <https://doi.org/10.1080/01621459.1994.10476870>.
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, <https://doi.org/10.1175/MWR2906.1>.
- Robertson, D. E., D. L. Shrestha, and Q. J. Wang, 2013: Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrol. Earth Syst. Sci.*, **17**, 3587–3603, <https://doi.org/10.5194/hess-17-3587-2013>.
- Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, <https://doi.org/10.3402/tellusa.v55i1.12082>.
- Schaake, J., and Coauthors, 2007a: Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrol. Earth Syst. Sci. Discuss.*, **4**, 655–717, <https://doi.org/10.5194/hessd-4-655-2007>.
- , T. M. Hamill, R. Buizza, and M. Clark, 2007b: HEPEX: The Hydrological Ensemble Prediction Experiment. *Bull. Amer. Meteor. Soc.*, **88**, 1541–1547, <https://doi.org/10.1175/BAMS-88-10-1541>.
- Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using Ensemble Model Output Statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, <https://doi.org/10.1002/qj.2183>.

- , and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.
- Shen, Y., and A. Xiong, 2016: Validation and comparison of a new gauge-based precipitation analysis over mainland China. *Int. J. Climatol.*, **36**, 252–265, <https://doi.org/10.1002/joc.4341>.
- Shrestha, D. L., D. E. Robertson, J. C. Bennett, and Q. J. Wang, 2015: Improving precipitation forecasts by generating ensembles through postprocessing. *Mon. Wea. Rev.*, **143**, 3642–3663, <https://doi.org/10.1175/MWR-D-14-00329.1>.
- Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, <https://doi.org/10.1175/MWR3441.1>.
- Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan, 2009: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resour. Res.*, **45**, 1–22, <https://doi.org/10.1029/2008WR006825>.
- Vannitsem, S., D. S. Wilks, and J. Messner, 2018: *Statistical Post-processing of Ensemble Forecasts*. Elsevier, 362 pp.
- Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.*, **131**, 965–986, <https://doi.org/10.1256/qj.04.120>.
- Wang, Q. J., and D. E. Robertson, 2011: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resour. Res.*, **47**, W02546, <https://doi.org/10.1029/2010WR009333>.
- , —, and F. H. S. Chiew, 2009: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resour. Res.*, **45**, 1–18, <https://doi.org/10.1029/2008WR007355>.
- , D. L. Shrestha, D. E. Robertson, and P. Pokhrel, 2012: A log-sinh transformation for data normalization and variance stabilization. *Water Resour. Res.*, **48**, W05514, <https://doi.org/10.1029/2011WR010973>.
- Wilks, D. S., 2006: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Appl.*, **13**, 243–256, <https://doi.org/10.1017/S1350482706002192>.
- , 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, <https://doi.org/10.1002/met.134>.
- , 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- , and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, <https://doi.org/10.1175/MWR3402.1>.
- Williams, R. M., C. A. T. Ferro, and F. Kwasniok, 2014: A comparison of ensemble post-processing methods for extreme events. *Quart. J. Roy. Meteor. Soc.*, **140**, 1112–1120, <https://doi.org/10.1002/qj.2198>.
- Wu, L., D. J. Seo, J. Demargne, J. D. Brown, S. Cong, and J. Schaake, 2011: Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *J. Hydrol.*, **399**, 281–298, <https://doi.org/10.1016/j.jhydrol.2011.01.013>.
- Zhang, Y., L. Wu, M. Scheuerer, J. Schaake, and C. Kongoli, 2017: Comparison of probabilistic quantitative precipitation forecasts from two postprocessing mechanisms. *J. Hydrometeorol.*, **18**, 2873–2891, <https://doi.org/10.1175/JHM-D-16-0293.1>.