A Variable-Correlation Model to Characterize Asymmetric Dependence for Postprocessing Short-Term Precipitation Forecasts

WENTAO LI

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

QUAN J. WANG

University of Melbourne, Melbourne, Victoria, Australia

QINGYUN DUAN

Hohai University, Nanjing, China

(Manuscript received 6 August 2019, in final form 1 October 2019)

ABSTRACT

Statistical postprocessing methods can be used to correct bias and dispersion error in raw ensemble forecasts from numerical weather prediction models. Existing postprocessing models generally perform well when they are assessed on all events, but their performance for extreme events still needs to be investigated. Commonly used joint probability postprocessing models are based on the correlation between forecasts and observations. Because the correlation may be lower for extreme events as a result of larger forecast uncertainty, the dependence between forecasts and observations can be asymmetric with respect to the magnitude of the precipitation. However, the constant correlation coefficient in the traditional joint probability model lacks the flexibility to model asymmetric dependence. In this study, we formulated a new postprocessing model with a decreasing correlation coefficient to characterize asymmetric dependence. We carried out experiments using Global Ensemble Forecast System reforecasts for daily precipitation in the Huai River basin in China. The results show that, although it performs well in terms of continuous ranked probability score or reliability for all events, the traditional joint probability model suffers from overestimation for extreme events defined by the largest 2.5% or 5% of raw forecasts. On the contrary, the proposed variable-correlation model is able to alleviate the overestimation and achieves better reliability for extreme events than the traditional model. The proposed variable-correlation model can be seen as a flexible extension of the traditional joint probability model to improve the performance for extreme events.

1. Introduction

Reliable and unbiased precipitation forecasts are important for hydrological forecasting, water resource management and other applications. Forecasting for extreme events is of special interest because extreme events may bring severe disasters and lead to significant life and property loss to human society. However, raw forecasts from numerical weather prediction (NWP)

models generally contain bias due to different errors in model inputs, initial conditions, model structure and parameters (Schaake et al. 2007b). Moreover, the raw forecast ensembles often suffer from underdispersion problems, namely the ensemble spread is too narrow to represent the real forecast uncertainty (Buizza et al. 2005).

Statistical postprocessing methods have been applied to correct the bias and the dispersion errors in raw forecasts from NWP models and enhance forecast skill (Cuo et al. 2011; Gneiting and Katzfuss 2014; Schaake et al. 2007b). During recent decades, various postprocessing methods have been developed. These models mainly follow the scheme of model output statistic (MOS), namely to establish statistical models of raw

---

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/MWR-D-19-0258.s1>.

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

forecast and observation using historical forecasts and observations, and then apply the fitted model to correct new forecasts in the future (Wilks 2011).

Statistical postprocessing models include both non-parametric models such as analog (Hamill and Whitaker 2006) and parametric models. The latter can be further classified into regression models and kernel density models (Wilks 2011). Ensemble MOS (EMOS) and logistic regression are examples of regression models, which use raw forecasts as predictors to predict the observations (Messner et al. 2014; Scheuerer and Hamill 2015; Wilks 2009). Other examples of regression models are the joint probability models, such as metaGaussian models (Krzysztofowicz and Evans 2008; Schaake et al. 2007a; Wu et al. 2011) and Bayesian joint probability (Robertson et al. 2013; Shrestha et al. 2015; Wang et al. 2009). The kernel density type of models includes ensemble dressing (Boucher et al. 2015; Fortin et al. 2006; Roulston and Smith 2003; Wang and Bishop 2005) and Bayesian model averaging (Raftery et al. 2005; Slougher et al. 2007), which can generate multimodal predictive distributions. Moreover, it is important to preserve the spatiotemporal and intervariable correlation for applications such as hydrological forecasting. The methods to maintain these dependencies including Schaake shuffle, ensemble copula coupling (ECC) and other variants of the two methods (Clark et al. 2004; Schefzik 2016; Schefzik et al. 2013; Wu et al. 2018). For more details of available postprocessing models, readers are referred to related books (Duan et al. 2019; Vannitsem et al. 2018) and a recent review of postprocessing methods for hydrometeorological forecasting (Li et al. 2017).

Existing postprocessing models are generally designed for the postprocessing of common events. They perform well in terms of overall metrics such as the continuous ranked probability score (CRPS) and reliability. There are several comparative studies on postprocessing for extreme events, such as postprocessing of wind speed forecasts (Lerch and Thorarinsdottir 2013), precipitation forecasts (Taillardat et al. 2019) or synthetic data from Lorenz 1996 model (Williams et al. 2014). However, their performance for extreme events still needs to be further investigated. Although Brier score and relative operating characteristic (ROC) can be used for evaluating the forecast performance for extreme events, these metrics are limited for evaluating only binary events (i.e., whether the precipitation amount exceeds a certain threshold). An alternative approach is to apply metrics such as CRPS and reliability that evaluate the full distribution of forecasts to stratified samples corresponding to extreme events. Such a stratified evaluation should be based on forecasts rather than

observations, because the latter will lead to biased results (Bellier et al. 2017; Lerch et al. 2017). Although the approach is straightforward, evaluation of the performance of postprocessing models based on stratified forecasts is rare (Bellier et al. 2017).

In this study, we aimed to evaluate and improve the performance of a joint probability model for extreme events defined by forecast-based stratification. Joint probability models are based on the bivariate normal assumption for transformed forecasts and observations, which may not be satisfied in hydrological applications (Khajehei and Moradkhani 2017; Wu et al. 2011). Traditionally, the correlation coefficient between transformed forecasts and observations is assumed to be constant in joint probability models. However, the correlation between raw forecasts and observations may be lower when raw forecasts are extremely high, because the forecast skill generally decreases for extreme events. Traditional joint probability models with constant correlation coefficients lack the flexibility to capture asymmetric dependence. Therefore, we proposed the variable-correlation model, which allows the correlation coefficients between transformed forecasts and observations to decrease when the forecasts become extremely high. In this way, the proposed model is able to characterize asymmetric dependence between the forecasts and observations. The traditional joint probability model becomes a special case of the proposed model.

We proposed the following questions in this study. 1) How does the traditional joint probability model perform for extreme events defined by forecast-based stratification? 2) How can we model asymmetric dependence to improve the forecast performance for extreme events? 3) How much improvement can be obtained by the proposed variable-correlation model, especially for extreme events?

To answer these questions, we verified the traditional joint probability model and the proposed variable-correlation model by experiments in the Huai River basin in China. The structure of the paper is as follows. Section 2 introduces the data and methods used in this paper. Section 3 provides the results of the traditional joint probability model and the proposed variable-correlation model. Section 4 discusses the advantages and limitations of the proposed method and summarizes the main conclusions.

2. Data and methods

a. Study catchment and data

The Huai River basin (30°55'–36°36'N, 111°55'–121°25'E) is located between the Yellow River and the

FIG. 1. Illustration of the Huai River basin.

Yangtze River in China, with an approximate drainage area of $270\,000\text{ km}^2$. It is under the influence of the Asian monsoon system, with mean annual precipitation of 700–1600 mm. The precipitation mainly occurs during the June–August flooding season. The Huai River basin was divided into 15 subbasins by the China Meteorological Administration for hydrometeorological forecasting purpose, as shown in Fig. 1 and Table 1.

The precipitation forecasts used in this study is the Global Ensemble Forecast System (GEFS) reforecasts provided by NOAA's National Centers for Environmental Prediction (Hamill et al. 2013). The raw reforecasts were downloaded at a spatial resolution of $1^\circ \times 1^\circ$ grid. Observations are the $0.5^\circ \times 0.5^\circ$ gridded daily China precipitation obtained from the China Meteorological Administration. The mean areal precipitation forecasts and observations were calculated from gridded GEFS forecasts and observations by inverse distance interpolation method.

b. The invariable-correlation joint probability model

In this subsection, the postprocessing method based on the joint probability model [referred to as invariable-correlation model hereinafter (IC)] is described. The IC model is generally similar to the Bayesian joint probability (BJP) model (Robertson et al. 2013; Shrestha et al. 2015; Wang et al. 2009). The main difference between

the IC model and the BJP model is the method of parameter inference. Because more data are available for parameter inference in short-term forecasting, maximum likelihood estimation (MLE) is used in this research instead of Bayesian inference in the original BJP. There are three steps in the IC model: 1) normalizing data using the log-sinh transformation, 2) modeling the joint distribution, and 3) applying the fitted model for postprocessing of new forecasts.

First, the log-sinh transformation is applied to observations and raw ensemble mean forecasts (refer to raw forecasts for brevity) separately to transform the precipitation variables into normally distributed variates. Let z and w be the raw ensemble mean forecasts and observations in the original space, respectively. Let x and y be the ensemble mean forecasts and observations in the transformed space, respectively. The log-sinh transformation for the raw forecasts is as follows (Wang et al. 2012):

$$x = \frac{1}{\lambda_x} \log[\sinh(\epsilon_x + \lambda_x z)]. \quad (1)$$

The distribution of transformed forecasts is assumed to follow a normal distribution:

$$x \sim N(\mu_x, \sigma_x^2). \quad (2)$$

TABLE 1. Main characteristics of the 15 subbasins of Huai River basin; ID gives the subbasin identifier.

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	1063.96
A2	Xixian upstream of Huaihe to Wangjiaba catchment	32.21°N, 115.02°E	8.8	1009.00
A3	Ruhe and upstream of Honghe catchment	33.04°N, 114.12°E	9.5	904.64
B1	Upstream of Yinghe to Zhoukou catchment	34.07°N, 113.33°E	27.4	687.60
B2	Midstream of Yinghe and Zhoukou to Fuyang catchment	33.47°N, 114.99°E	14.3	824.94
B3	Shihe catchment	32.19°N, 115.73°E	10.6	1130.30
B4	Pihe, downstream of Huaihe and Huaigan catchment	32.03°N, 116.29°E	11.4	1103.45
B5	Wohe, midstream of Huaihe and Huaigan catchment	33.48°N, 116.04°E	28.7	781.07
C0	Bengbu to Hungtse, midstream and downstream of Huaigan and Huihe catchment	33.65°N, 117.40°E	42.3	859.87
D1	Nansihu catchment	35.31°N, 116.32°E	30.8	634.28
D2	Zaozhuang and Xuzhou catchment	34.63°N, 117.78°E	9.2	781.11
D3	Upstream of Yihe catchment	35.62°N, 118.12°E	10.1	719.30
D4	Upstream of Shuhe catchment	35.61°N, 118.84°E	4.4	717.89
D5	Downstream of Yihe and Shuhe catchment	34.41°N, 118.95°E	26.9	864.71
E0	Hungtse to downstream of Huaihe catchment	33.18°N, 119.82°E	30.6	947.35

The four parameters ε_x , λ_x , μ_x , and σ_x in Eqs. (1) and (2) are estimated using MLE. Details of the likelihood function for log-sinh transformation are in [appendix Ac](#). Similarly, the log-sinh transformation for the observations is

$$y = \frac{1}{\lambda_y} \log[\sinh(\varepsilon_y + \lambda_y w)]. \quad (3)$$

The distribution of transformed observation is assumed to follow a normal distribution:

$$y \sim N(\mu_y, \sigma_y^2). \quad (4)$$

The four parameters ε_y , λ_y , μ_y and σ_y in Eqs. (3) and (4) are estimated using MLE. The forecasts or observations that are less than or equal to a threshold of 0.1 mm day^{-1} are treated as censored data in the likelihood functions. The 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. Here the term “censored data” means it is only known that the precipitation amount is less than or equal to the censoring threshold, but the precipitation amount is not precisely specified ([Robertson et al. 2013](#); [Shrestha et al. 2015](#); [Wang and Robertson 2011](#)). Then the joint distribution of the transformed forecasts and observations is assumed to be a bivariate normal distribution:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma_y^2 \end{bmatrix} \right). \quad (5)$$

The only additional parameter in Eq. (5) is the correlation coefficient ρ between transformed forecasts

and observations. This parameter can be estimated by MLE (see [appendix Aa](#) for details of the likelihood functions).

Last, the conditional distribution of the observation given a new forecast can be obtained. The new ensemble mean forecast z is transformed to x by the log-sinh transformation. If the transformed new forecast x is larger than the censoring threshold x_c , the predictive samples of observation can be obtained by drawing random samples from the conditional distribution:

$$p(y|x) \sim N \left[\mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x), (1 - \rho^2) \sigma_y^2 \right]. \quad (6)$$

If the new forecast is less than the threshold, “data augmentation” is used to draw a random sample x_{aug} that satisfies $x_{\text{aug}} \leq x_c$ from the marginal distribution of the ensemble mean forecasts ([Robertson et al. 2013](#); [Wang and Robertson 2011](#)). Then the random sample is used to substitute x in Eq. (6) and get one sample of y from the conditional distribution. This process is repeated to get all the postprocessed ensemble members. The ensemble size is set as 1000 according to previous experiments. Last, the inverse of the log-sinh transformation is applied to transform the samples into the original space.

To ensure the ensemble members with suitable spatiotemporal correlation, the Schaake shuffle is applied to the generated ensemble members ([Clark et al. 2004](#)). Because the sample size of 1000 is relatively large, the 1000 samples are divided into four blocks with 250 samples, similar to the Schaake shuffle implemented in [Schepen et al. \(2018\)](#). The 250 random samples in each block are shuffled according to the spatiotemporal

dependence template obtained from historical observations. A total of 25 years of historical observations within a 10-day window centered on the forecast validation date during 1960–84 (before the cross-validation period of 1985–2009) are used to obtain the spatiotemporal dependence template for the 250 samples in each block. In this way, the shuffled ensemble members preserve the spatiotemporal correlation of historical observations (Clark et al. 2004).

c. The variable-correlation model

The traditional IC model assumes a constant correlation coefficient between the transformed raw forecasts and observations. However, as the forecast skill for extreme events is usually low, the correlation should also be lower for extreme events than that for moderate events. Therefore, we propose a variable-correlation (VC) model that allows the correlation coefficient between raw forecasts and observations to decrease when raw forecasts are extremely high. Details of the variable-correlation model are as follows.

First, the log-sinh transformation is applied to the raw forecasts and observations as in section 2b. Then, the conditional distribution of the observations given raw forecasts in the transformed space is assumed as follows:

$$p(y|x) \sim N\left\{\mu'_y + \rho(x)\frac{\sigma'_y}{\sigma_x}(x - \mu_x), [1 - \rho^2(x)](\sigma'_y)^2\right\}, \quad (7)$$

where

$$\rho(x) = \rho_0 \tanh\left[\frac{\sigma_x C}{\max(0, x - \mu_x)}\right], \quad (8)$$

The conditional distribution in Eq. (7) is designed in a similar form to the conditional distribution for the traditional IC model in Eq. (6). Instead of a constant correlation coefficient in the traditional IC model, the correlation coefficient $\rho(x)$ in Eq. (7) is defined as a decreasing function of the transformed forecasts x with parameter ρ_0 and C as shown in Eq. (8). The parameters μ_x and σ_x are still the parameters of the marginal distribution of the transformed raw forecasts, estimated during the fitting of the log-sinh transformation as in section 2b. The parameters μ'_y and σ'_y in Eq. (7) are new parameters and may differ from the parameters for the marginal distribution of the transformed observations.

As will be illustrated later in the result section (Fig. 2), the correlation coefficient defined in Eq. (8) is constant for most of the nonextreme events. It gradually

decreases when the forecasts become high. The decreasing rate is controlled by the parameter C . Smaller C values lead to faster decreasing rates. In fact, when C is 50, $\rho(x)$ is almost a constant for most of the events, which means the proposed model reverts to the traditional joint probability model. In this way, the proposed VC model allows the correlation between forecasts and observations to decrease with the forecasts and is able to characterize asymmetric dependence between forecasts and observations in postprocessing problems.

To constrain the estimated parameter C in Eq. (8) within a suitable range, the transformed forecasts x are standardized. In other words, forecasts are subtracted by the mean μ_x and divided by the standard deviation σ_x . The maximum function in the denominator of Eq. (8) is used to avoid negative values of the denominator. The four parameters μ'_y , σ'_y , ρ_0 , and C in Eqs. (7) and (8) are fitted together by MLE. Note that the joint distribution of the forecasts and observations in the VC model is no longer bivariate normal, so the likelihood functions differ from those in the IC model (see appendix Ab for details).

After fitting the model, the conditional distribution of observations given a new forecast can be obtained according to Eqs. (7) and (8). Then, ensemble members can be generated from the predictive distribution and Schaake shuffle can be applied as in section 2b.

d. Forecast verification

To verify the performance of the two postprocessing models, a 25-fold leave-one-year-out cross validation is conducted by the 25-yr GEFS reforecasts and observations dataset during 1985–2009. Postprocessing models are fitted for each subbasin in Huai River basin on each day during the rainy season (June–August). The training dataset for each day is composed of a 31-day window centered on that day during the training years, thus a training dataset of 31×24 days can be obtained.

Several commonly used verification metrics are used including the bias, the root-mean-square error (RMSE), the mean continuous ranked probability skill score (CRPSS), the Brier skill score (BSS), the probability integral transform (PIT) diagram, and the α index. The sampling uncertainty for the first four verification metrics is estimated by generating 1000 bootstrap samples and calculating confidence intervals from these samples. The method to generate confidence intervals follows the block bootstrap method described in Hamill (1999) to consider the spatial correlation among subbasins. The CRPSS and BSS are computed by taking the postprocessed results of the IC model as the reference forecasts to compare the

FIG. 2. Predicted and empirical conditional quantiles of 5%, 25%, 50%, 75%, and 95% (the median is highlighted in red) in (top) transformed space and (bottom) original space at lead time of 5 days for three postprocessing models [(a),(d) IC; (b),(e) VC; and (c),(f) CLR; solid lines], obtained by pooling samples from 15 subbasins together. The empirical quantiles are shown as crosses.

performance of the proposed model relative to the IC model. Moreover, a permutation paired test (Hamill 1999) is used to evaluate whether the improvements by the proposed VC model over the IC model are significant for CRPS and Brier score. Details of the verification metrics can be found in appendix B and related references (Wilks 2011).

Moreover, the bias, RMSE, CRPS, PIT diagrams, and α index are also calculated specifically for the samples with raw forecast mean larger than the thresholds of 95% or 97.5% quantiles of the raw ensemble mean forecasts, in order to evaluate the performance of the postprocessed results corresponding to the largest 5% or 2.5% of raw forecasts. As will be shown in the PIT diagrams in the results section (Figs. 4 and 5, described in more detail below), the Kolmogorov bands (dashed lines) at the significance of 0.05 are plotted to graphically test the uniformity of the PIT values by

Kolmogorov-Smirnov goodness-of-fit test (Laio and Tamea 2007). If all the PIT values lie within the Kolmogorov band, the uniformity of the PIT values cannot be rejected. In other words, the corresponding forecasts are reliable.

To compare the proposed model with state-of-the-art postprocessing methods, three regression models are applied including the censored logistic regression (CLR), heteroscedastic censored logistic regression (Messner et al. 2014) and the censored, shifted Gamma distribution (CSGD)-based EMOS (Scheuerer and Hamill 2015) in the experiment. The logistic regression models are selected because normalization transformations such as log-sinh transformation can be applied in these models, which makes these models comparable to the joint probability model in this study. CSGD-EMOS is also included in the comparison, because it contains a nonlinear model for the mean parameter and is able

FIG. 3. The fitted correlation coefficients for the invariable-correlation model and the variable-correlation model in subbasin D2 at lead times of (a) 1, (b) 3, and (c) 5 days. The fitted value of parameter C is labeled above each panel. The models were fitted by the samples during the summers of the 25 years.

to capture the asymmetric dependence between forecasts and observations. The CSGD-EMOS used here doesn't incorporate ensemble spread predictors in order to make a fair comparison with the joint probability models. A brief description of these three models is in [appendix C](#).

3. Results

a. Checking the variable-correlation assumption

In this section, the variable-correlation assumption of the VC model is checked. In [Fig. 2](#), the quantiles obtained from the postprocessed predictive distributions (solid lines) are compared with the empirical conditional quantiles of the observations given forecasts (crosses) in untransformed and transformed space. The empirical conditional quantiles are estimated by selecting the observation-forecast pairs with raw forecasts falling within a small window ($x - \varepsilon, x + \varepsilon$) around a series of forecast values x and calculating the quantiles of these observations, similar to the method used for [Fig. 6](#) in [Scheuerer and Hamill \(2015\)](#). To better estimate the empirical conditional quantiles for extreme events, the samples in all the 15 subbasins during the summers of the 25 years are pooled together to increase the sample size, including $92 \text{ days} \times 25 \text{ years} \times 15 \text{ subbasins} = 34\,500$ samples in total.

[Figure 2](#) shows the empirical and predictive conditional quantiles of the IC, VC and CLR models at the lead time of 5 days. The empirical conditional quantiles of observations given forecasts in normal space (crosses in [Figs. 2a–c](#)) exhibit a nonlinear relationship with the transformed forecasts for extreme events. The linear quantile lines of the IC model ([Fig. 2a](#)) or the CLR model ([Fig. 2c](#)) fail to capture the nonlinear relationship in normal space and lead to overestimation when raw forecasts become extremely high. On the contrary,

the predictive quantile lines of the VC model ([Fig. 2b](#)) can be nonlinear in normal space and generally approximate well with the empirical conditional quantiles of observations. There is still some discrepancy between the empirical and predictive quantiles of the VC model due to sampling error. The predictive quantile lines of the VC model also correspond generally well to the empirical quantiles in the original space ([Fig. 2e](#)), while the predictive quantiles of the IC or CLR models ([Figs. 2d or 2f](#)) lead to overestimation when the raw forecasts become extremely high. The results above show the assumption of variable correlation in a joint probability model is appropriate. If the correlation coefficient is constant, the predictive quantile lines in normal space are linear ([Fig. 2a](#)) and cannot capture the nonlinear relationship between forecasts and observations. Note that the nonlinearity may depend on lead times. As shown in [Fig. 2](#) in the paper and [Figs. S4–S6](#) in the online supplemental materials, the nonlinearity is remarkable at lead times of 3 or 5 days, but the nonlinearity is not so obvious at lead time of 1 or 7 days. The predictive quantiles obtained from the proposed VC model are able to approximate well with the empirical quantiles in both cases.

b. Detailed results for one subbasin

In this section, the detailed results of one subbasin (Subbasin D2) are presented as an example to show the advantage of the proposed variable-correlation model. [Figure 3](#) shows the correlation coefficients of the VC and IC models fitted by the samples in the summers of all the 25 years at lead times of 1, 3 and 5 days. As shown in [Fig. 3](#), the fitted correlation coefficients of the VC model (red curve) are almost constant when raw forecasts are light or moderate rain (less than 25 mm day^{-1}), slightly higher than those by the IC model (dashed blue line). Then, the correlation coefficients of the VC model gradually

FIG. 4. The PIT diagrams for the invariable-correlation model and the variable-correlation model at thresholds of (a)–(c) 0%, (d)–(f) 95%, and (g)–(i) 97.5% quantiles of raw forecasts in subbasin D2 at lead times of (left) 1, (center) 3, and (right) 5 days.

decrease when forecasts become extremely high. The correlation coefficients for the VC model at lead times of 3 or 5 days decrease faster than those at the lead time of 1 day.

Figure 4 shows the PIT diagrams of the postprocessed results corresponding to raw forecasts larger than the thresholds of 0% (Figs. 4a–c), 95% (Figs. 4d–f) and 97.5% (Figs. 4g–i) quantiles of raw forecasts at three lead times (1, 3, and 5 days) in Subbasin D2. In PIT diagrams of all the samples (Figs. 4a–c), the PIT values

of both the IC and VC models align well with the diagonal line, which indicates both models can achieve overall reliable forecasts. However, the PIT values of the IC model (blue points) for higher thresholds are below the diagonal line and even exceed the Kolmogorov 5% significance bands, especially at lead times of 3 and 5 days (Figs. 4e,f,h,i), which exhibits IC model may suffer from overestimation for extreme events. The PIT values of the VC model (red points) are much closer to the diagonal line than those of the IC

FIG. 5. The PIT diagrams by pooling the samples of all 15 subbasins together for the invariable-correlation model and the variable-correlation model at thresholds of (a)–(c) 0%, (d)–(f) 95%, and (g)–(i) 97.5% quantiles of raw forecasts at lead times of (left) 1, (center) 3, and (right) 5 days.

model for extreme events, which indicates the overestimation problem of the IC model can be alleviated by the VC model for these extreme events.

Figures 3 and 4 show that when the IC model suffers from overestimation (e.g., the lead times of 3 or 5 days in Fig. 4) the fitted correlation coefficients by the VC model will decrease faster than those for other cases to alleviate the overestimation problem. Figures 3 and 4 show that the function form of the correlation

coefficients in the VC model is appropriate regardless of whether there is obvious overestimation for the IC model.

c. Verification for all subbasins

In this section, the reliability, bias, RMSE, CRPSS and BSS of the postprocessed forecasts are evaluated for all 15 subbasins. Detailed results for each of the subbasins can be found in Figs. S1–S3 in the online supplemental materials. Figure 5 shows the stratified PIT diagrams of

FIG. 6. The bias for the raw forecasts and the postprocessed results for 15 subbasins. The bias is computed by (a) all samples and (b) samples corresponding to raw forecasts larger than the 95% quantile. The postprocessing models include the invariable-correlation (IC) model, the variable-correlation (VC) model, the censored logistic regression (CLR), the heteroscedastic censored logistic regression (HCLR), and the censored-shifted Gamma distribution (CSGD)-based EMOS with ensemble mean as the only predictor. The 90% confidence intervals by bootstrapping are shown by the error bars.

the postprocessed results corresponding to raw forecasts larger than at the thresholds of 0% (Figs. 5a–c), 95% (Figs. 5d–f) and 97.5% (Figs. 5g–i) quantiles of raw forecasts in all 15 subbasins. Figure 5 generally exhibits similar patterns with Fig. 4. While the PIT values of VC and IC models are both close to diagonal lines at threshold of 0% quantile (Figs. 5a–c), the PIT values of the IC model (blue points) are below diagonal lines and exceed the Kolmogorov 5% significance band at higher thresholds, especially at lead times of 3 and 5 days (Figs. 5e,f,h,i). On the contrary, the PIT values for the VC model (red points) are close to the diagonal lines and are generally within the 5% significance band at high thresholds. The results in Fig. 5 show that the VC model achieves better reliability than the IC model for extreme events defined by the largest 2.5% or 5% of raw forecasts.

Figure 6 shows the bias of the raw forecasts and the postprocessed results of the five models including IC, VC, CLR, HCLR, and CSGD-EMOS (the acronyms are defined in the caption). As shown in Fig. 6a, raw forecasts (black bars with crosses) suffer from overestimation for the lead time of one day and suffer from underestimation for lead times of 6–7 days. The bias of the five postprocessed results is near to zero if the results are evaluated for all the samples (Fig. 6a). As shown in Fig. 6b, the raw forecasts are positively biased if they are evaluated by the samples corresponding to the largest 5% of raw forecasts. The results of the IC, CLR and HCLR models still suffer from overestimation for these cases. On the contrary, the postprocessed results of the VC model and the CSGD-EMOS model are unbiased when they are evaluated for these extreme events.

Figure 7 shows the RMSE for the raw forecasts and the five postprocessing models. All five postprocessing

models achieve lower RMSE than the raw forecasts. The VC model and the CSGD-EMOS achieve the smallest RMSE among the five postprocessing models, especially for the threshold of 95% quantiles. The RMSE of the IC model is larger than that of the VC and CSGD-EMOS models, but smaller than that of two logistic regression models.

Figure 8 shows the CRPSS of the postprocessed results of the VC, CLR, HCLR, and CSGD-EMOS models. The postprocessed results of the IC model are used as a reference to compute the skill score. Although the CRPSS of the four models are generally similar when the results are evaluated by all the samples (Fig. 8a), remarkable differences can be seen when the results are evaluated for the samples corresponding to the largest 5% or 2.5% of raw forecasts (Figs. 8b,c). The VC model and CSGD-EMOS perform the best among these postprocessing models. The VC model significantly outperforms the IC model at lead times of 2–5 days (shown in red filled triangle markers). The CRPSS of the CLR and HCLR models are negative at lead times of 2–5 days, which indicates these two models are worse than the IC model for these lead times.

The Brier skill score of the postprocessing models is shown in Fig. 9 at three thresholds of 85%, 95%, and 97.5% quantiles. The postprocessed results of the IC model are still used as a reference. Figure 9 generally exhibits similar patterns with Fig. 8. The CSGD-EMOS and the VC model still perform the best among these postprocessing models at most of the lead times. The VC model performs significantly better than the IC model in terms of Brier score at most of the lead times for high thresholds (shown as red filled triangles, such as lead times of 3 and 5 days in Fig. 9b and lead times

FIG. 7. As in Fig. 6, but for the RMSE.

of 1–3 and 5 days in Fig. 9c). The BSS for the two logistic regression models falls below zero at several lead times for high thresholds, which indicates that their performance is worse than the IC model for these situations.

4. Discussion and conclusions

As shown in the result section, although the traditional joint probability model performs well in terms of overall forecast skill and reliability, it may suffer from overestimation for extreme events defined by the largest 2.5% or 5% of raw forecasts. The reason can be attributed to the limitation of the bivariate normal distribution assumption in the traditional joint probability model. The model checking results in section 3a show the relationship between transformed forecasts and observations in transformed space can be nonlinear, which indicates asymmetric dependence between

transformed forecasts and observations. The correlation between transformed forecasts and observations should be lower for extreme events than that for non-extreme events, because the forecast skill is generally lower for extreme events. However, the constant correlation coefficient in the traditional joint probability model cannot capture the asymmetric dependence and may lead to overestimation for extreme events. In fact, the bivariate normal distribution assumption for the transformed forecasts and observations in traditional IC model may not be valid in postprocessing problems (Wu et al. 2011). Khajehei and Moradkhani (2017) also found the traditional bivariate normal distribution-based model is inferior to copula-based models for postprocessing of extreme events in monthly precipitation forecasts.

To improve the forecast performance for extreme events, we developed the variable-correlation model in this study. The form of the conditional distribution

FIG. 8. The CRPSS for the postprocessed results for 15 subbasins. The postprocessed results from the IC model are used as the reference to compute the skill score. The CRPSS is computed by (a) all samples, (b) samples corresponding to raw forecasts larger than 95% quantile, and (c) samples corresponding to raw forecasts larger than 97.5% quantile. The 90% confidence intervals by bootstrapping are shown by error bars. Red filled triangles indicate that the VC model significantly outperforms the IC model at a significance level of 5% as indicated by the permutation test.

FIG. 9. As in Fig. 8, but for the BSS using thresholds of (a) 85% quantile, (b) 95% quantile, and (c) 97.5% quantile of raw forecasts.

of observations given forecasts in the VC model is generally similar to that in the IC model, but the correlation coefficient in the VC model is a decreasing function of the transformed forecasts. The decreasing correlation coefficients make the postprocessed forecasts in the VC models lower than those in IC models when raw forecasts become extremely high. In this way, the VC model alleviates the overestimation problem of the IC model and achieves more reliable forecasts than the traditional IC model for extreme events.

Moreover, the proposed VC model can be seen as a flexible extension of the traditional joint probability model. The variable correlation coefficients in the VC model make the model more flexible to capture asymmetric dependence between forecasts and observations. The proposed model reverts to the traditional model when the fitted correlation coefficient is not decreasing significantly. The traditional IC model is only a special case of the proposed VC model. Note that extreme events were defined by the largest 2.5% or 5% of raw forecasts instead of the observations in this study. The stratification for extreme events should be based on forecasts instead of observations, because only the former can ensure reliable forecasts (Lerch et al. 2017; Bellier et al. 2017).

Although we mainly extended a joint probability model in this work, our results are also meaningful to other postprocessing models. Transformation-based regression models such as CLR or HCLR are usually based on the assumption of linearity between transformed forecasts and observations. The model checking results (Fig. 2) exhibit that the relationship between forecasts and observations can be nonlinear even in transformed space. Existing models such as CLR or HCLR which are based on a linear assumption for transformed variates cannot capture such a nonlinear relationship. Further improvements might be made to state-of-the-art models such as CLR or HCLR by allowing a nonlinear relationship of forecasts and

observations in transformed space. The VC model has a similar effect with the nonlinear CSGD-EMOS developed by Scheuerer and Hamill (2015). As shown in the result section, the proposed VC model generally performs as well as the CSGD-EMOS for cases when the raw forecasts become extremely high. The slightly inferior performance of the VC model relative to CSGD-EMOS might be attributed to the fact that the parameters are estimated by CRPS minimization in CSGD-EMOS instead of MLE, which improves the performance of CSGD-EMOS in terms of CRPS or BSS.

The proposed VC model still has limitations. First, ensemble mean is used as the only predictor in the current VC model. However, researchers have found that the incorporation of other predictors from ensemble forecasts will enhance the forecast skill. For example, the predictor of ensemble spread will improve the quantifying of the forecast uncertainty (Scheuerer and Hamill 2015; Zhang et al. 2017). Moreover, the probability of precipitation also provides useful information about the occurrence of precipitation (Gebetsberger et al. 2018). How to add these predictors to the proposed model will be investigated in the future.

Acknowledgments. We are grateful to the valuable comments from the editor and anonymous reviewers. We are also thankful for Dr. Yating Tang for providing useful comments on the early version of the paper. 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), the National Key Research and Development Program of China (2018YFE0196000), and the Special Fund for Meteorological Scientific Research in Public Interest (GYHY201506002; CRA-40: The 40-Year CMA Global Atmospheric Reanalysis). The first author is supported by China Scholarship Council.

APPENDIX A

The Details of the Likelihood Functions*a. The likelihood function for the invariable-correlation model*

The likelihood function for the IC model can be divided into four cases depending on whether the

transformed forecasts and observations are censored, given a data series with n forecast–observation pairs, as follows (Schepen et al. 2016):

$$L(x, y; \rho) = \prod_{t=1}^n l(t), \quad (\text{A1})$$

where

$$l(t) = \begin{cases} \phi_{\text{BN}}[x(t), y(t); \rho] & \text{if } x(t) > x_c, \ y(t) > y_c \\ \Phi_N[y_c; \mu_{y|x}, \sigma_{y|x}] \times \phi_N[x(t); \mu_x, \sigma_x] & \text{if } x(t) > x_c, \ y(t) \leq y_c \\ \Phi_N(x_c; \mu_{x|y}, \sigma_{x|y}) \times \phi_N[y(t); \mu_y, \sigma_y] & \text{if } x(t) \leq x_c, \ y(t) > y_c \\ \Phi_{\text{BN}}(x_c, y_c; \rho) & \text{if } x(t) \leq x_c, \ y(t) \leq y_c \end{cases}. \quad (\text{A2})$$

For the first and fourth case in Eq. (A2), ϕ_{BN} and Φ_{BN} are the density and cumulative distribution function (CDF) for the bivariate normal distribution defined in Eq. (5), respectively. For the second case, $\Phi_N(y_c; \mu_{y|x}, \sigma_{y|x})$ is the CDF value at the censoring threshold of observations y_c for the conditional distribution defined in Eq. (6); $\phi_N(\mu_x, \sigma_x)$ is the density of the marginal distribution of the transformed forecasts. For the third case, $\Phi_N(x_c; \mu_{x|y}, \sigma_{x|y})$ is the CDF value at the censoring threshold of forecasts x_c for the conditional distribution

$$p(x|y) \sim N\left[\mu_x + \rho \frac{\sigma_x}{\sigma_y}(y - \mu_y), (1 - \rho^2)\sigma_x^2\right]; \quad (\text{A3})$$

$\phi_N(\mu_y, \sigma_y)$ is the marginal distribution of the transformed observations.

b. The likelihood function for the variable-correlation model

The likelihood function for the VC model can also be divided into four cases depending on whether the transformed forecasts and observations are censored as follows:

$$l(t) = \begin{cases} \phi_N[y(t); \mu_{y|x}, \sigma_{y|x}] \times \phi_N[x(t); \mu_x, \sigma_x] & \text{if } x(t) > x_c, \ y(t) > y_c \\ \Phi_N(y_c; \mu_{y|x}, \sigma_{y|x}) \times \phi_N[x(t); \mu_x, \sigma_x] & \text{if } x(t) > x_c, \ y(t) \leq y_c \\ \Phi_N(x_c; \mu_{x|y}, \sigma_{x|y}) \times \phi_N[y(t); \mu'_y, \sigma'_y] & \text{if } x(t) \leq x_c, \ y(t) > y_c \\ \Phi_{\text{BN}}\{x_c, y_c; \rho[x(t)]\} & \text{if } x(t) \leq x_c, \ y(t) \leq y_c \end{cases}. \quad (\text{A4})$$

For the first two cases for which the forecasts are larger than the censoring threshold, $\phi_N(y(t); \mu_{y|x}, \sigma_{y|x})$ and $\Phi_N(y_c; \mu_{y|x}, \sigma_{y|x})$ are the density and CDF of the conditional distribution of observations given forecasts defined in Eqs. (7) and (8). The conditional distribution function is used for the first case in Eq. (A4) instead of the joint distribution in Eq. (5), because the joint distribution of the transformed forecasts and observations is no longer bivariate normal for the first case in the VC model.

For the third and fourth cases when the transformed forecasts are less than or equal to the censoring threshold, as the correlation coefficients become almost constant ρ_0 for these two cases according to Eq. (8), the joint probability distribution of transformed forecasts and observations can be approximated by a bivariate normal distribution as follows:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left\{\begin{bmatrix} \mu_x \\ \mu'_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho_0 \sigma_x \sigma'_y \\ \rho_0 \sigma_x \sigma'_y & (\sigma'_y)^2 \end{bmatrix}\right\}, \quad (\text{A5})$$

where ρ_0 , μ'_y , and σ'_y are defined in Eqs. (7) and (8).

Then, the likelihood functions similar to those in Eq. (A2) can be used for the third and fourth cases in Eq. (A4). The difference is the newly fitted parameters defined in Eqs. (7) and (8) should be used in Eq. (A4). Specifically, the conditional distribution $\Phi_N(\mu_{x|y}, \sigma_{x|y})$ for the third case in Eq. (A4) is defined as follows:

$$p(x|y) \sim N\left[\mu_x + \rho_0 \frac{\sigma_x}{\sigma'_y}(y - \mu'_y), (1 - \rho_0^2)\sigma_x^2\right]; \quad (\text{A6})$$

$\phi_N(\mu'_y, \sigma'_y)$ is the density for the univariate normal distribution as follows:

$$y \sim N[\mu'_y, (\sigma'_y)^2]. \quad (\text{A7})$$

For the fourth case, Φ_{BN} in Eq. (A4) is the CDF for the bivariate normal distribution defined in Eq. (A5).

c. The likelihood function for the log-sinh transformation

The likelihood function for the log-sinh transformation for the 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 (Wang and Robertson 2011; Wang et al. 2012):

$$L(z; \varepsilon_x, \lambda_x, \mu_x, \sigma_x) = \prod_{i=1}^n l[z(t); \varepsilon_x, \lambda_x, \mu_x, \sigma_x], \quad (\text{A8})$$

where

$$l[z(t); \varepsilon_x, \lambda_x, \mu_x, \sigma_x] = \begin{cases} \coth[\varepsilon_x + \lambda_x z(t)] \times \phi\{f[z(t)]; \varepsilon_x, \lambda_x, \mu_x, \sigma_x\} & \text{if } z(t) > z_c, \\ \Phi(x_c; \varepsilon_x, \lambda_x, \mu_x, \sigma_x) & \text{if } z(t) \leq z_c \end{cases} \quad (\text{A9})$$

where ϕ and Φ are the density and CDF of the distribution of transformed forecasts defined in Eq. (2), respectively; $f(z)$ is the log-sinh transformation for forecasts, defined in Eq. (1); x_c is the censoring threshold in the transformed space. The likelihood function of the log-sinh transformation for observations is similar and is omitted here.

APPENDIX B

The Verification Metrics

a. Bias

The bias measures the average difference between the ensemble mean forecast $\bar{f}_i$ and corresponding observation o_i :

$$\text{bias} = \frac{1}{n} \sum_{i=1}^n (\bar{f}_i - o_i), \quad (\text{B1})$$

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

b. Continuous rank probability skill score

The CRPS measures the integral square difference between the CDF of a forecast and corresponding CDF of observation as follows:

$$\text{CRPS} = \int_{-\infty}^{\infty} [F_f(q) - F_o(q)]^2 dq. \quad (\text{B2})$$

Then, the mean CRPS of all samples can be computed from the average of the CRPS of all forecast and observation pairs. The CRPSS can be computed as follows:

$$\text{CRPSS} = 1 - \frac{\overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{ref}}}, \quad (\text{B3})$$

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.

c. Brier skill score

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

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n [F_{f_i}(q) - F_{o_i}(q)]^2, \quad (\text{B4})$$

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:

$$F_{o_i}(q) = \begin{cases} 1, & o_i > q \\ 0, & \text{otherwise} \end{cases}. \quad (\text{B5})$$

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

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{ref}}}. \quad (\text{B6})$$

The BSS is positively oriented. The BSS for the perfect forecast is 1, and the BSS is less than 0 for forecasts with no skill relative to the reference.

d. The PIT diagram and the α index

The PIT is the value of the predictive CDF F_f at the corresponding observation o_i as follows:

$$\text{PIT}_i = F_f(o_i). \quad (\text{B7})$$

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 or 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 $[0, F_y(y_c)]$, where y_c is the censoring threshold for observation (Robertson et al. 2013).

The reliability can also be summarized by the α index as follows (Renard et al. 2010):

$$\alpha = 1 - \frac{2}{n} \sum_{i=1}^n \left| \text{PIT}_i^* - \frac{i}{n+1} \right|, \quad (\text{B8})$$

where PIT_i^* is the sorted PIT_i values in increasing order; n is the total number of forecast–observation pairs. The α index ranges from 0 (worst reliability) to 1 (perfect reliability). The α index is an overall reliability index and cannot be used to diagnose the specific over/underestimation or over/underdispersion problems of forecasts.

APPENDIX C

The Logistic Regression and CSGD-EMOS Models

We compared two versions of the censored logistic regression models with the proposed joint probability model. To make a fair comparison with the joint probability model, we also used the log–sinh transformation in logistic regression models, and estimated the parameters by MLE. The transformed observations y_t is assumed to follow a left-censored normal distribution as follows, similar to the model used in Gebetsberger et al. (2017):

$$y_t \sim \text{Normal}(\mu_t, \sigma_t), \quad (\text{C1})$$

$$\mu_t = \beta_0 + \beta_1 \bar{x}_t, \quad \text{and} \quad (\text{C2})$$

$$\sigma_t^2 = \gamma_0 + \gamma_1 (\text{MD}_{x_t})^2. \quad (\text{C3})$$

The scale parameter σ_t is predicted by ensemble spread predictors to form an HCLR, as shown in Eq. (C3). In this study, we chose a quadratic function as the link function of the scale submodel according to our previous experiments. 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:

$$\text{MD}_{x_t} = \frac{1}{m^2} \sum_{j,j'=1}^m |x_t^j - x_t^{j'}|, \quad (\text{C4})$$

where m is the ensemble size and x_t^j and $x_t^{j'}$ are two ensemble members at time t . Another version of the logistic regression model used in this study is to assume the scale parameter σ_t to be constant; then the logistic regression model can be named as censored logistic regression. For more details of CLR or HCLR models, please see relative references (e.g., Messner et al. 2014).

The CSGD-EMOS is a regression model developed in the untransformed space. The observed precipitation is assumed to follow a censored, shifted Gamma distribution with location and scale parameters as follows (Scheuerer and Hamill 2015):

$$\mu = \frac{\mu_{\text{cl}}}{\alpha_1} \log 1p \left[\exp m 1(\alpha_1) \left(\alpha_2 + \alpha_3 \frac{\bar{f}}{\bar{f}_{\text{cl}}} \right) \right] \quad \text{and} \quad (\text{C5})$$

$$\sigma = \alpha_4 \sigma_{\text{cl}} (\mu / \mu_{\text{cl}})^{1/2}, \quad (\text{C6})$$

where $\alpha_1, \dots, \alpha_4$ are the regression parameters, $\bar{f}$ is the ensemble mean of raw forecasts, $\bar{f}_{\text{cl}}$ is the climatology of the ensemble mean, μ_{cl} and σ_{cl} are the parameters for the climatology distribution of observation, $\log 1p(x) = \log(1 + x)$, and $\exp m 1(x) = \exp(x) - 1$. More details about CSGD-EMOS can be found in Scheuerer and Hamill (2015).

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>.
- 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).
- 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, Eds., 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>.
- 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 non-homogeneous 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>.
- Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, <https://doi.org/10.1146/annurev-statistics-062713-085831>.
- Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, [https://doi.org/10.1175/1520-0434\(1999\)014<0155:HTFENP>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2).
- , G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarnau, Y. Zhu, and W. Lapenta, 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 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>.
- Khajehei, S., and H. Moradkhani, 2017: Towards an improved ensemble precipitation forecast: A probabilistic post-processing approach. *J. Hydrol.*, **546**, 476–489, <https://doi.org/10.1016/j.jhydrol.2017.01.026>.
- 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>.
- Lerch, S., and T. L. Thorarinsdottir, 2013: Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus*, **65A**, 21206, <https://doi.org/10.3402/tellusa.v65i0.21206>.
- , —, 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>.
- Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014: 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>.
- 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>.
- Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks, 2010: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resour. Res.*, **46**, W05521, <https://doi.org/10.1029/2009WR008328>.
- 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. C., 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>.
- Schefzik, R., 2016: A similarity-based implementation of the Schaake shuffle. *Mon. Wea. Rev.*, **144**, 1909–1921, <https://doi.org/10.1175/MWR-D-15-0227.1>.
- , T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, <https://doi.org/10.1214/13-STS443>.
- Schepen, A., Q. J. Wang, and D. E. Robertson, 2016: Application to post-processing of meteorological seasonal forecasting. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan et al., Eds., Springer, 1–29.
- , T. Zhao, Q. J. Wang, and D. E. Robertson, 2018: A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments. *Hydrol. Earth Syst. Sci.*, **22**, 1615–1628, <https://doi.org/10.5194/hess-22-1615-2018>.
- Scheuerer, M., 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>.
- 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>.
- Taillardat, M., A.-L. Fougères, P. Naveau, and O. Mestre, 2019: Forest-based and semiparametric methods for the post-processing of rainfall ensemble forecasting. *Wea. Forecasting*, **34**, 617–634, <https://doi.org/10.1175/WAF-D-18-0149.1>.
- 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, Eds., 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, 362 pp.
- 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**, W05407, <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>.
- 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>.
- Wilks, D. S., 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.
- 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>.
- , Y. Zhang, T. Adams, H. Lee, Y. Liu, and J. Schaake, 2018: Comparative evaluation of three Schaake Shuffle schemes in postprocessing GEFS precipitation ensemble forecasts. *J. Hydrometeor.*, **19**, 575–598, <https://doi.org/10.1175/JHM-D-17-0054.1>.
- Zhang, Y., L. Wu, M. Scheuerer, J. Schaake, and C. Kongoli, 2017: Comparison of probabilistic quantitative precipitation forecasts from two postprocessing mechanisms. *J. Hydrometeor.*, **18**, 2873–2891, <https://doi.org/10.1175/JHM-D-16-0293.1>.