Technical Note

An improved meta-Gaussian distribution model for post-processing of precipitation forecasts by censored maximum likelihood estimation

Wentao Li, Qingyun Duan*, Aizhong Ye, Chiyuan Miao

State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

ABSTRACT

Statistical post-processing methods have been applied in hydrometeorological forecasting to correct the bias and spread error in raw forecasts. Among various post-processing methods, the meta-Gaussian distribution model (MGD) is one of the early successful methods for post-processing of precipitation forecasts and has been applied in the National Weather Service's Hydrologic Ensemble Forecast System (HEFS), together with the mix-type meta-Gaussian distribution model (MMGD). However, recent studies have shown that the original MGD cannot yield reliable forecasts especially for sub-daily precipitation forecasts (e.g., 6-hourly). In this paper, we improved the MGD model by applying the censored maximum likelihood estimation (CMLE) method. We conducted experiments using GEFS reforecasts in Huai river basin in China to evaluate its performance. The results show that the proposed method performs better than the original MGD for sub-daily precipitation forecasts. The proposed method also achieves similar forecast skill with the state-of-the-art censored, shifted Gamma distribution-based ensemble MOS (CSGD-EMOS) if both use ensemble mean as the only predictor. The results indicate that the proposed CMLE-MGD can be useful for further applications such as flood forecasting that needs forecasts of high temporal resolution.

1. Introduction

Hydrometeorological ensemble forecasts are vitally important for applications such as flood forecasting and water resource management. However, due to different uncertainty sources including model inputs, initial conditions, model structures and parameters, raw forecasts from meteorological or hydrological models suffer from systematic bias and dispersion errors and need to be corrected (Demargne et al., 2014; Schaake et al., 2010; Schaake et al., 2007b). Various statistical post-processing methods have been developed to correct these errors and achieve sharp forecasts subject to calibration (Gneiting et al., 2007; Gneiting and Katzfuss, 2014; Gneiting and Raftery, 2007). These methods mainly follow the model output statistics (MOS) scheme, namely to fit the statistical models using historical forecasts and corresponding observations, then apply the fitted model to calibrate future forecasts (Wilks, 2011).

The post-processing of precipitation forecasts is challenging. One difficulty is the intermittent nature of precipitation, namely a positive probability for zero precipitation and a skewed distribution for nonzero precipitation (Scheuerer and Hamill, 2015; Shrestha et al., 2015; Zhang et al., 2017). During the past decade, several post-processing methods have been developed to deal with this problem. Some methods apply

data transformations and then develop statistical models for the transformed variables. This type of methods includes logistic regression (LR) models (Messner et al., 2014a; Messner et al., 2014b; Wilks, 2009), meta-Gaussian distribution (MGD) models (Kelly and Krzysztofowicz, 1997; Krzysztofowicz and Evans, 2008; 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; Wang et al., 2012). There are also other models that do not need transformations, such as the censored, shifted Gamma distribution ensemble MOS (CSGD-EMOS) (Scheuerer and Hamill, 2015). Multi-model averaging methods have also been developed particularly for precipitation forecasts, such as the Gamma distribution-based Bayesian model averaging (Raftery et al., 2005; Sloughter et al., 2007).

Moreover, the heteroscedasticity problem also exists for post-processing of precipitation forecasts, which means the forecast uncertainty increases with the magnitude of forecast variables (Scheuerer and Hamill, 2015). To deal with the heteroscedastic problem, the non-homogenous regression scheme can be applied, namely to adjust the variance of the predictive distribution by predictors such as raw forecast mean or spread (Gneiting et al., 2005; Messner et al., 2014b; Scheuerer and Hamill, 2015). Besides, the modeling of the spatio-temporal correlation of precipitation forecasts is important for

---

applications such as hydrological forecasting. To solve this problem, non-parametric reordering methods have been widely used. Notable ensemble reordering methods including Schaake shuffle (Clark et al., 2004), ensemble copula coupling (ECC) (Schefzik et al., 2013) and other variants of the two schemes (Joseph et al., 2017; Schefzik, 2015; Scheuerer et al., 2017; Wu et al., 2018). Parametric methods based on copula or the geostatistical output perturbation model (GOP) have also been developed in recent years (Berrocal et al., 2008; Gel et al., 2004; Möller et al., 2013; Pinson and Girard, 2011). A recent paper did a more detailed review of those post-processing methods (Li et al., 2017).

Among existing post-processing methods, the meta-Gaussian distribution (MGD) model (Schaake et al., 2007a) is one of the early successful methods and it has been applied in the Meteorological Ensemble Forecast Processor (MEFP) for National Weather Service's Hydrologic Ensemble Forecast Service (HEFS) (Demargne et al., 2014), together with the mixed-type meta-Gaussian distribution-based model (MMGD) (Wu et al., 2011). However, Wu et al. (2011) found that the original MGD by Schaake et al. fails to obtain reliable forecasts for 6-hourly precipitation, because the estimation of the correlation coefficients of transformed forecasts and observations in the original MGD cannot yield satisfying results (Wu et al., 2011). This disadvantage restricts the original MGD for applications such as flood forecasting that needs precipitation forecasts at high temporal resolution.

In this paper, we proposed a new version of MGD model by applying maximum likelihood estimation for censored data (CMLE for short) to improve its performance. With the proposed CMLE-MGD model, we treated the zero precipitation values as "left-censored" data. Here the term "left-censored" means it is only known that the precipitation amount is less than the threshold of zero precipitation (e.g., 0.1 mm/day), but the exact precipitation amount is unknown. In this way, we made it feasible to estimate the correlation coefficients between the observations and forecasts including zero precipitation values. This censoring technique has been successfully applied in other post-processing models such as HCLR (Messner et al., 2014a) and BJP (Robertson et al., 2013). We compared the proposed CMLE-MGD with the original MGD and the state-of-the-art CSGD-EMOS to demonstrate the effectiveness of the new method.

The paper is organized as follows. Section 2 introduces the data and methods used in this paper. Section 3 describes the validation of the assumptions for the post-processing model. Section 4 provides the comparison results of the three post-processing methods. Section 5 discusses the advantages and limitations of the proposed method. Finally, Section 6 summarizes the main conclusions.

2. Data and methods

2.1. Study region and data

As shown in Fig. 1, the Huai river basin (30°55′–36°36′N, 111°55′–121°25′E) is situated between the Yangtze and Yellow rivers in Eastern China, with an approximate drainage area of 270,000 km². The mean annual precipitation is approximately 700–1600 mm, decreasing from the south to the north. Huai river basin is under the influence of the Asian monsoon system, so most of the precipitation occurs during the June–August flooding season. The characteristics of the 15 sub-basins used in this study are shown in Table 1 (Liu et al., 2013). The 15 subbasins were divided by the China Meteorological Administration for hydrometeorological forecasting purpose.

The precipitation forecasts used here is the Global Ensemble Forecast System (GEFS) reforecasts provided by NOAA's National Centers for Environmental Prediction (NCEP) (Hamill et al., 2013). The raw reforecasts are downloaded at a spatial resolution of 1° × 1° grid. Observations are the 0.1° × 0.1° gridded hourly China Hourly Merged Precipitation Analysis (CMPA-Hourly) obtained from the China Meteorological Administration (Shen et al., 2014). The precipitation analysis is merged from gauge-based analysis and satellite-based CMORPH

precipitation from NOAA Climate Prediction Center (CPC). As the CMPA-Hourly dataset is only available from 2008, here we used ten years of observations and corresponding GEFS reforecasts in rain season (June–August) during 2008–2017. The mean areal precipitation (MAP) forecasts and observations were calculated from gridded GEFS forecasts and observations by inverse distance interpolation method.

2.2. The original MGD and CMLE-MGD

The original MGD and CMLE-MGD both follow four steps as shown in Fig. 2, namely 1) the Normal quantile transformation (NQT) for both the forecasts and the observations; 2) the estimation of the joint distribution between the transformed observations and forecasts in Normal space; 3) the computation of the conditional distribution of observations given the new forecasts and the inverse NQT; 4) the Schaake shuffle procedure to maintain the spatio-temporal correlation. The difference between the original MGD and CMLE-MGD only exists in the second step.

2.2.1. The Normal quantile transformation procedure

The Normal quantile transformation (NQT) is designed to transform the non-Gaussian precipitation forecasts and observations into standard Normal variables as follows:

$$u = \Phi_N^{-1}(F_x(x)) \tag{1}$$

$$v = \Phi_N^{-1}(F_y(y)) \tag{2}$$

where Φ_N^{-1} is the inverse of the cumulative distribution function (CDF) of the standard Normal distribution, F_x and F_y are the CDFs of the marginal distribution of raw forecasts and observations, respectively. In this study, we used mixed-type distributions for the marginal distribution of forecasts and observations. For example, the marginal distribution of forecasts is as follows:

$$F_x(x) = p_0 + (1 - p_0) \cdot F_{x|x > 0}(x) \tag{3}$$

where p_0 is the probability of the occurrence of precipitation, and $F_{x|x > 0}$ is the CDF of the precipitation amount given the occurrence of precipitation. The distribution for the non-zero precipitation is selected based on the p-value of the Anderson–Darling goodness-of-fit test (Anderson and Darling, 1954) for several parametric distributions including Pearson Type III, Weibull and the generalized Gamma distribution. These three distributions were chosen because they were found to perform well in fitting the distribution for daily or sub-daily precipitation amounts (e.g., Papalexiou and Koutsoyiannis, 2016; Papalexiou et al., 2018; Ye et al., 2018). These parametric distributions also perform well for the study region according to the results of the Anderson–Darling test in previous experiments.

2.2.2. The estimation of the joint distribution of transformed forecasts and observations in Normal space

In the meta-Gaussian model, the transformed forecasts u and observations v are assumed to follow the bivariate standard Normal distribution as follows (Schaake et al., 2007a):

$$f(u, v) \sim N(\mu, \Sigma) \tag{4}$$

where $\mu = (0, 0)$, and $\Sigma = \begin{pmatrix} 1, \rho \\ \rho, 1 \end{pmatrix}$.

The only parameter to be estimated for the joint distribution is the correlation coefficient ρ in Normal space. The difference between the original MGD and the proposed CMLE-MGD is how ρ is estimated. In original MGD, several options are provided for the estimation of ρ. The default option is to substitute ρ in Normal space by that in the untransformed space. There are also other options in the original MGD, but the default option leads to acceptable results in general (Liu et al., 2013; Tao et al., 2014; Ye et al., 2017). However, this substitution is not correct, and may lead to an underestimation in the post-processed
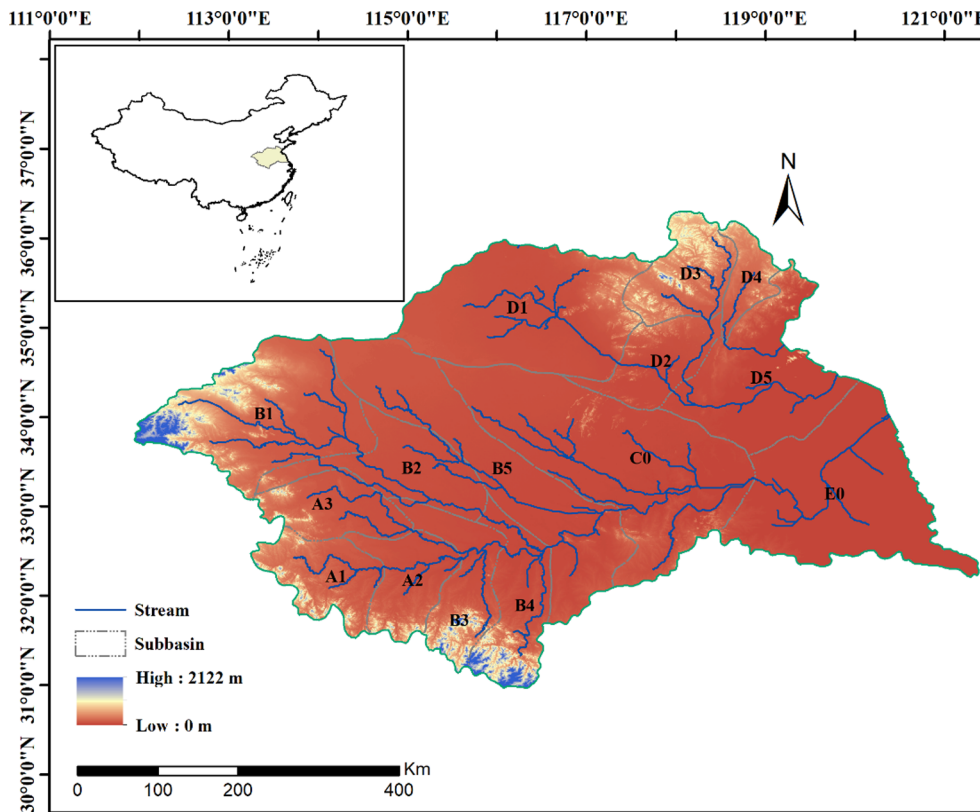
Fig. 1. Illustration of the Huai river basin.

results (Wu et al., 2011).

In this paper, we tried to estimate the actual ρ in Normal space directly by applying the maximum likelihood estimation for censored data (CMLE). We treated the zero precipitation values as "left-censored" at the threshold of zero precipitation (e.g., 0.1 mm/day used in this study). Here the term "left-censored" means the data is known to be less than or equal to the threshold, but unknown precisely. The MLE method for censored data was first developed in statistics and has been successfully used in the post-processing of precipitation forecasts (e.g., Messner et al., 2014a; Robertson et al., 2013).

The details of the likelihood functions are described as follows. The CDF value at the censoring threshold instead of the density should be used in the likelihood function for censored data. Specifically, depending on whether the forecasts and observations are censored or not, the likelihood function in Normal space can be divided into four cases as follows (Schepen et al., 2016):

$$L(u, v; \rho) = \prod_{i=1}^{t} l(t) \tag{5}$$

where,

$$l(t) = \begin{cases} \phi_{BN}(u(t), \ v(t); \rho) & if \ u(t) > u_0, \ v(t) > v_0 \\ \Phi_N(v_0; \mu_{v|u}, \sigma_{v|u}) \cdot \phi_N(u(t)) & if \ u(t) > u_0, \ v(t) \leqslant v_0 \\ \Phi_N(u_0; \mu_{u|v}, \sigma_{u|v}) \cdot \phi_N(v(t)) & if \ u(t) \leqslant u_0, \ v(t) > v_0 \\ \Phi_{BN}(u_0, v_0; \rho) & if \ u(t) \leqslant u_0, \ v(t) \leqslant v_0 \end{cases} \tag{6}$$

u_0 and v_0 are the transformed values for the thresholds of zero precipitation for forecasts and observations, respectively. For the first and last cases, ϕ_{BN} and Φ_{BN} are the density and the CDF for the standard bivariate Normal distribution, respectively. For the second case, is the density for a standard univariate Normal distribution; $\Phi_N(\mu_{v|u}, \sigma_{v|u})$ is the CDF of the conditional distribution of observations given forecasts

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	114.01°E, 32.31°N	16.5	1063.96
A2	Xixian upstream of Huaihe to Wangjiaba catchment	115.02°E, 32.21°N	8.8	1009.00
A3	Ruhe and upstream of Honghe catchment	114.12°E, 33.04°N	9.5	904.64
B1	Upstream of Yinghe to Zhoukou catchment	113.33°E, 34.07°N	27.4	687.60
B2	Midstream of Yinghe and Zhoukou to Fuyang catchment	114.99°E, 33.47°N	14.3	824.94
B3	Shihe catchment	115.73°E, 32.19°N	10.6	1130.30
B4	Pihe, downstream of Huaihe and Huaigan catchment	116.29°E, 32.03°N	11.4	1103.45
B5	Wohe, midstream of Huaihe and Huaigan catchment	116.04°E, 33.48°N	28.7	781.07
C0	Bengbu to Hungtse, midstream and downstream of Huaigan and Huihe catchment	117.40°E, 33.65°N	42.3	859.87
D1	Nansihu catchment	116.32°E, 35.31°N	30.8	634.28
D2	Zaozhuang and Xuzhou catchment	117.78°E, 34.63°N	9.2	781.11
D3	Upstream of Yihe catchment	118.12°E, 35.62°N	10.1	719.30
D4	Upstream of Shuhe catchment	118.84°E, 35.61°N	4.4	717.89
D5	Downstream of Yihe and Shuhe catchment	118.95°E, 34.41°N	26.9	864.71
E0	Hungtse to downstream of Huaihe catchment	119.82°E, 33.18°N	30.6	947.35
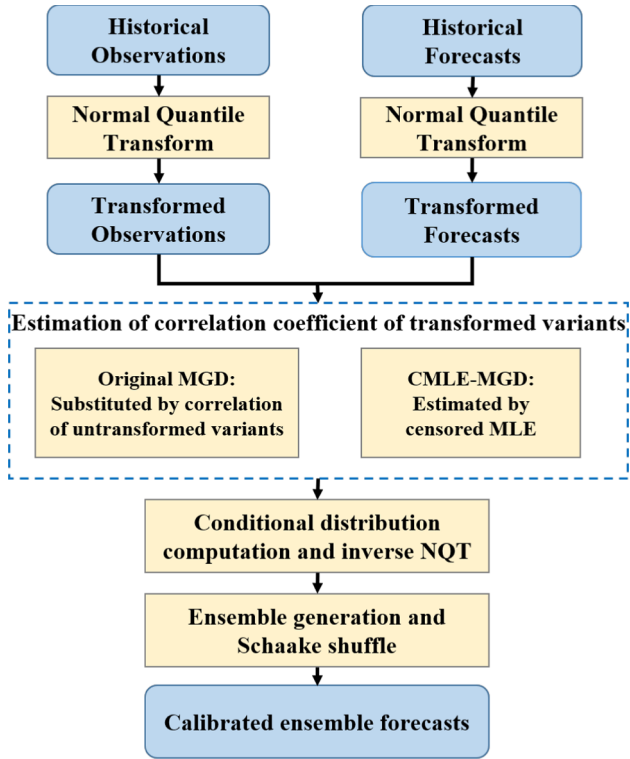
Fig. 2. The flowchart of the original MGD and CMLE-MGD.

with parameters as follows:

$$\mu_{v|u} = \rho \cdot u(t) \tag{7}$$

$$\sigma_{v|u} = 1 - \rho^2 \tag{8}$$

Similarly, for the third case, is the density for a standard univariate Normal distribution; $\Phi_N(\mu_{u|v}, \sigma_{u|v})$ is the CDF of the conditional distribution of forecasts given observations with parameters as follows:

$$\mu_{u|v} = \rho \cdot v(t) \tag{9}$$

$$\sigma_{u|v} = 1 - \rho^2 \tag{10}$$

2.2.3. The computation of the conditional distribution and inverse NQT

After the estimation of the joint distribution, the conditional distribution of observation given a new forecast can be obtained exactly in the same way in the original MGD scheme (Schaake et al., 2007a). Specifically, if the new forecast is larger than zero, the conditional distribution of the observation is a Normal distribution with parameters defined in Eqs. (7) and (8). If the new forecast is less than the threshold of zero precipitation, the CDF of the observation in Normal space can be written as follows (Schaake et al., 2007a):

$$F_{v|u}(v|u, u \leqslant u_0) = \frac{\int_{-\infty}^{v}\int_{-\infty}^{u_0} f(u,v)dudv}{\int_{-\infty}^{\infty}\int_{-\infty}^{u_0} f(u,v)dudv} = \frac{\int_{-\infty}^{v} \Phi_N(u_0; \mu_{u|v}, \sigma_{u|v}) \cdot \phi_N(v)dv}{\int_{-\infty}^{\infty} \Phi_N(u_0; \mu_{u|v}, \sigma_{u|v}) \cdot \phi_N(v)dv} \tag{11}$$

where $\Phi_N(\mu_{u|v}, \sigma_{u|v})$ is the CDF of the conditional distribution of forecasts given observations with parameters defined as Eqs. (9) and (10); is the density for a standard univariate Normal distribution. Then, the observation given a new forecast in original space can be obtained by the inverse NQT as follows (Schaake et al., 2007a):

$$y = F_y^{-1}(\Phi_N(F_{v|u}^{-1}(v|u))) \tag{12}$$

where Φ_N is the CDF of the standard univariate Normal distribution, F_y^{-1} is the inverse of the CDF of the observations.

2.2.4. The ensemble generation using the Schaake shuffle procedure

After obtaining the calibrated predictive distribution, ensemble members can be generated from the predictive distribution by the sampling method used in the original MGD (Schaake et al., 2007a). Firstly, a larger number (e.g., 1000 in this research) of equally spaced samples are drawn from the predictive distribution obtained in Eq. (12). Then, the predictive distribution is partitioned into n equal intervals with probability equal to $1/n$, where n is the ensemble size. An ensemble size of 100 was used in this research according to previous experiments. Finally, the ensemble members are obtained by computing the sample mean within each interval of the predictive distribution.

Then, Schaake shuffle is applied to generate ensemble series with appropriate spatio-temporal correlation (Clark et al., 2004). Firstly, the current ensemble members and the historical observations on similar dates are both ranked. Then the ensemble members are reordered to match the rank of the historical observations. In this way, the shuffled ensemble members represent the spatio-temporal correlation of historical observations. For more details on Schaake shuffle, please refer to Clark et al. (2004).

2.3. Forecast experiments and verification strategies

To verify the performance of the proposed method, we conducted a 10-fold leave-one-year-out cross-validation using the 10-year GEFS reforecasts and corresponding observations. In other words, we chose nine years of data for training and used the other one year of data for verification, and repeated this process for ten times. Post-processing models were fitted for each month using a training dataset composed from a 91-day window centered on the 15th of each month during the training years. For example, the training dataset for the post-processing model in June is the 9-year data within the time window of 1st May –30th July, thus a training dataset of 91 × 9 days can be obtained. In this study, we focused on 6-hourly accumulated precipitation amounts within the lead time of five days, because the forecast skill for 6-hourly precipitation is limited beyond five days.

Three post-processing methods including the original MGD, CMLE-MGD and CSGD-EMOS were applied to the MAP of each subbasin. The CSGD-EMOS is briefly described in the Appendix. The post-processed forecasts were verified both for all the 15 subbasins together and for each subbasin individually. The verification for all 15 subbasins was implemented after pooling the samples from the 15 subbasins and 10-fold cross-validations together. The verification for each subbasin was implemented after pooling the samples of 10-fold cross-validations together. Several commonly used verification metrics were applied, including the relative mean error (RME), the mean continuous ranked probability score (CRPS) (Hersbach, 2000) and the Brier skill score (BSS) (Brier, 1950). Reliability is the statistical consistency of forecasts and observations, and here we used the stratified probability integral transform (PIT) histogram (Dawid, 1984; Gneiting et al., 2007; Bellier et al., 2017) to assess the reliability. As for discrimination, we used ROC score computed from the relative operating characteristic (ROC) curve (Wilks, 2011). The sampling uncertainty of metrics including RME, CRPS, BSS and ROC score was estimated by the stationary block bootstrap technique (Politis and Romano, 1994) for 1000 times by the software of Ensemble Verification System (EVS, Brown et al., 2010). Details of the verification metrics can be found in the Appendix and related references (e.g., Wilks, 2011).

The stratified PIT histogram is able to show the reliability for precipitation forecasts of different strata. In this work, three strata are divided by the 85% and 95% quantiles of raw forecast mean. In other words, the three strata are corresponding to the samples with raw forecast mean within the range of 0–85% quantiles (light rain), 85%–95% quantiles (moderate rain), and 95%–100% quantiles (heavy rain). Then, the PIT histograms for each stratum are plotted and stacked together, as will be shown in Fig. 5. It should be noted that the stratification should be based on forecasts (e.g., raw forecast mean as we

used here), instead of observations to ensure calibration for each stratum (Lerch et al., 2017; Bellier et al., 2017).

3. Validation of the model assumption

In this section, we applied the proposed CMLE-MGD and the original MGD method to the full dataset described in Section 2 to validate the bivariate Normal distribution assumption for the joint distribution of transformed forecasts and observations. Due to the existence of zero precipitation, it is not easy to directly check the bivariate Normal distribution assumption, so we compared the estimated correlation coefficients of transformed variants with the "optimal" correlation coefficients. The "optimal" correlation coefficients are defined as the correlation coefficients that minimize the CRPS of the post-processed ensemble forecasts. As we haven't got the closed form of the CRPS for MGD model, we simply searched the optimal correlation coefficients by traversing a series of correlation coefficients within the 0–1 interval. We computed the CRPS of the ensemble forecasts from the MGD model with each correlation coefficient value. Then, the correlation coefficients with the lowest CRPS were selected as the optimal correlation coefficients. The other parameters in MGD were kept the same during the searching for the optimal correlation coefficients.

In Fig. 3, we plotted the optimal correlation coefficients in transformed space (horizontal axis) versus the estimated correlation coefficients (vertical axis) by CMLE-MGD (red dots) and original MGD (blue circles) for the results of all the subbasins. As shown in Fig. 3, the correlation coefficients estimated by the original MGD (blue circles) are obviously lower than the optimal correlation coefficients. The results show that the original MGD underestimates the correlation coefficients of the transformed variants. On the contrary, the correlation coefficients estimated by CMLE-MGD (red dots) are much closer to the 1:1 line than those by the original MGD, which indicates better correlation coefficients can be obtained by the proposed CMLE-MGD. The superior results of CMLE-MGD also indirectly suggest that the assumption of bivariate Normal distribution for transformed forecasts and observations is generally appropriate.
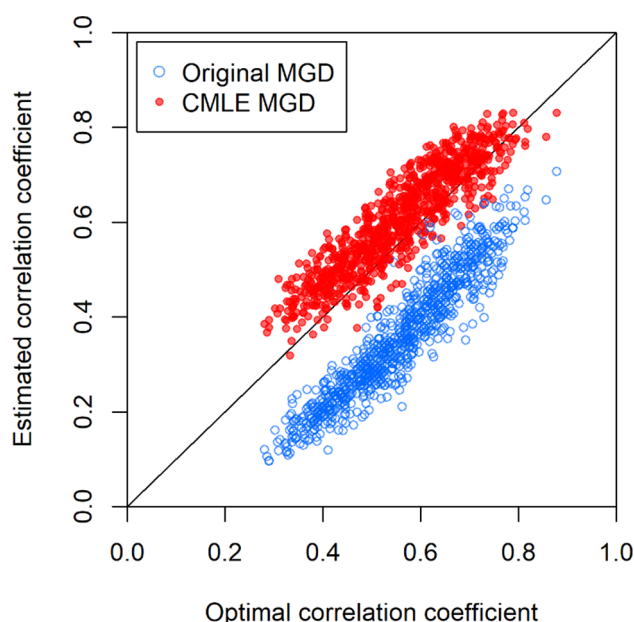

Fig. 3. The scatterplot of the correlation coefficients of transformed forecasts and observations estimated by the original MGD (blue circles) and CMLE-MGD (red dots) against the optimal correlation coefficients (defined in Section 3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Verification results

In this section, we evaluated the predictive performance of the GEFS forecasts and the results of the post-processing methods using the 10-fold cross-validation method described in Subsection 2.3. The verification results for all the 15 subbasins together are shown in Subsections 4.1 and 4.2. The verification results for individual subbasins are presented in the supplementary materials.

4.1. The overall forecast performance

In Fig. 4, the relative mean error and the mean CRPS (hereafter referred to as CRPS for brevity) of all 15 subbasins for raw GEFS forecasts and the post-processed forecasts are plotted versus lead times. The 90% confidence intervals are obtained through bootstrapping for each metric by EVS. As shown in Fig. 4a, raw GEFS forecasts (black lines) suffer from an obvious positive bias of 20%–60%. After post-processed by three methods, the relative mean error is much lower, which means that the post-processing methods are able to correct the systematic bias in raw GEFS forecasts. The relative mean error of CMLE-MGD (red lines) is 10%–20%, higher than that of the original MGD (blue lines), which is 0%–10%. The reason for the overestimation problem of CMLE-MGD will be further analyzed in Section 5.

In Fig. 4b, the CRPS results show that all three post-processing methods obtained much lower CRPS values than those of the raw GEFS forecasts. Among the three post-processing methods, EMOS performs the best, followed by CMLE-MGD and the original MGD, but the differences between the CRPS of the three post-processing methods are not obvious in general.

In Fig. 5, the stratified PIT histograms show the reliability of the post-processed forecasts from the original MGD and CMLE-MGD for three strata (light, moderate and heavy rain based on raw forecasts) at the four lead times within the first day. As shown in Fig. 5a–d, the histograms of the original MGD exhibit obvious underestimation (upslope "/"-shape), especially for moderate and heavy rain (the histograms in moderate blue and dark blue color). On the contrary, the histograms of all the samples for CMLE-MGD (Fig. 5e–h) are generally flat, which indicates CMLE-MGD can achieve generally reliable forecasts. Fig. 5 shows that the proposed CMLE-MGD outperforms the original MGD in term of overall reliability. It should also be noted that the PIT histograms of CMLE-MGD for heavy rain (in dark blue color) are still not perfect and indicate a bit overestimation (downslope "\"-shape), especially for lead time of 12–18 h (Fig. 5g) and 18–24 h (Fig. 5h). The corresponding reason will be further analyzed in Section 5. The PIT histograms for other lead times are generally similar to the results in Fig. 5 and are only shown in the supplemental material.

4.2. The forecast performance for dichotomous events

The forecast performance of the original MGD and CMLE-MGD for dichotomous events at different precipitation thresholds are assessed through the Brier skill score and the ROC score in this subsection. Three thresholds, namely the 85% quantile (0.6 mm/6h), 95% quantile (6.5 mm/6h) and 97.5% quantile (13.3 mm/6h) of observed 6-hourly precipitation amounts during summers of 2008–2017 are used. As shown in Fig. 6a–c, the BSS results of CMLE-MGD (red lines) are better than those of original MGD (blue lines) at the threshold of 85% and 95% quantiles. For the threshold of 97.5% quantile, the BSS results of CMLE-MGD are slightly better than those of the original MGD for events within the first 72-h lead times. The BSS results show that CMLE-MGD outperforms the original MGD mainly for light and moderate rain.

Fig. 6d–f show the ROC scores of the post-processed methods. The ROC scores for the two methods are generally similar for the first 72-h lead times. For several longer lead times, the CMLE-MGD outperforms the original MGD, such as for lead times beyond 96 h in Fig. 6e and for lead times beyond 60 h in Fig. 6f. The results show that CMLE-MGD also
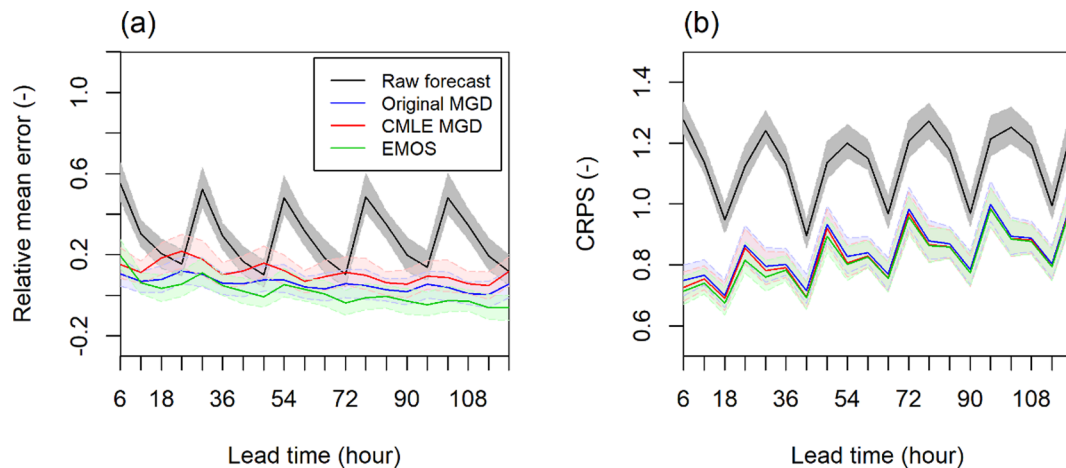
Fig. 4. Overall evaluation for the raw GEFS forecasts and post-processed forecasts by the original MGD, CMLE-MGD and EMOS. (a) Relative mean error, (b) CRPS. Both metrics are negative oriented (smaller values are better). The 90% confidence intervals by bootstrapping are shown for each metric.

performs better than the original MGD in terms of discrimination for longer lead times.

In summary, the results from all the 15 subbasins demonstrate that the proposed CMLE-MGD performs better than the original MGD in terms of Brier skill score, especially for light and moderate rain. There are also improvements in discrimination in terms of ROC score for several lead times. The detailed results for each of the 15 subbasins are shown in the supplementary materials, which also indicate improvements in terms of BSS for most of the subbasins.

5. Discussion

As illustrated in Section 4, the proposed CMLE-MGD is generally superior to the original MGD in terms of forecast skill and overall reliability. The original MGD cannot achieve reliable forecasts for 6-hourly precipitation, which confirms the results in Wu et al. (2011). The

reason of the improvements of CMLE-MGD can be attributed to the improved estimation of correlation coefficients between the transformed forecasts and observations by the CMLE method (Fig. 3), as it is the only difference between the two methods.

In fact, the estimation of correlation coefficients for the transformed variates with censored data needs special treatments. If the default option in the original MGD is used, the correlation coefficient of the transformed variates is substituted by the correlation coefficient in the untransformed space. However, this substitution method in the original MGD will lead to underestimation of the correlation coefficients, which can be proved theoretically by investigating the relationship between the correlation of untransformed variants and the correlation of transformed variants (Papalexiou, 2018). Another estimation option of the correlation coefficients in the original MGD is a weighted sum of the correlation coefficients for untransformed variants and those for transformed variants, but the results are still not satisfying (Wu et al.,
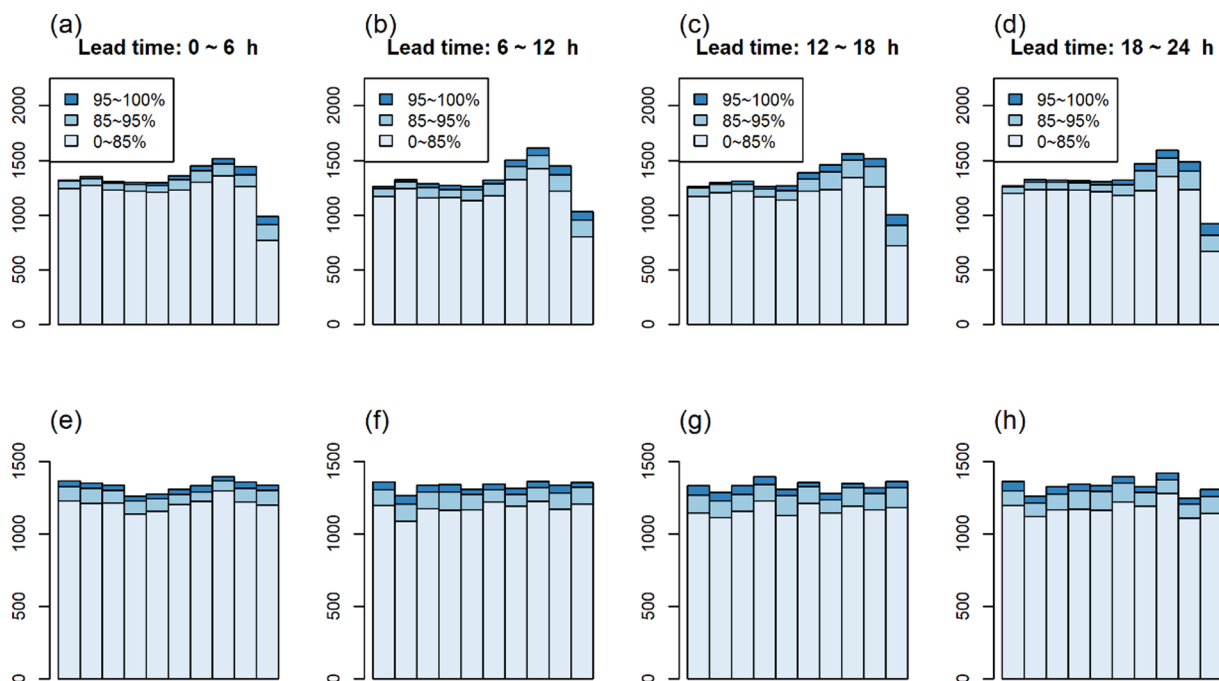

Fig. 5. The stratified PIT histogram of the post-processed forecasts at lead time of 0–6 h, 6–12 h, 12–18 h and 18–24 h for (a)–(d) original MGD and (e)–(h) CMLE-MGD. The PIT histograms in color of light blue, moderate blue and dark blue are the PIT histograms for the three strata, corresponding to the samples with raw forecast mean within the 0–85% quantiles (light rain), 85%–95% quantiles (moderate rain), and 95%–100% quantiles (heavy rain), respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
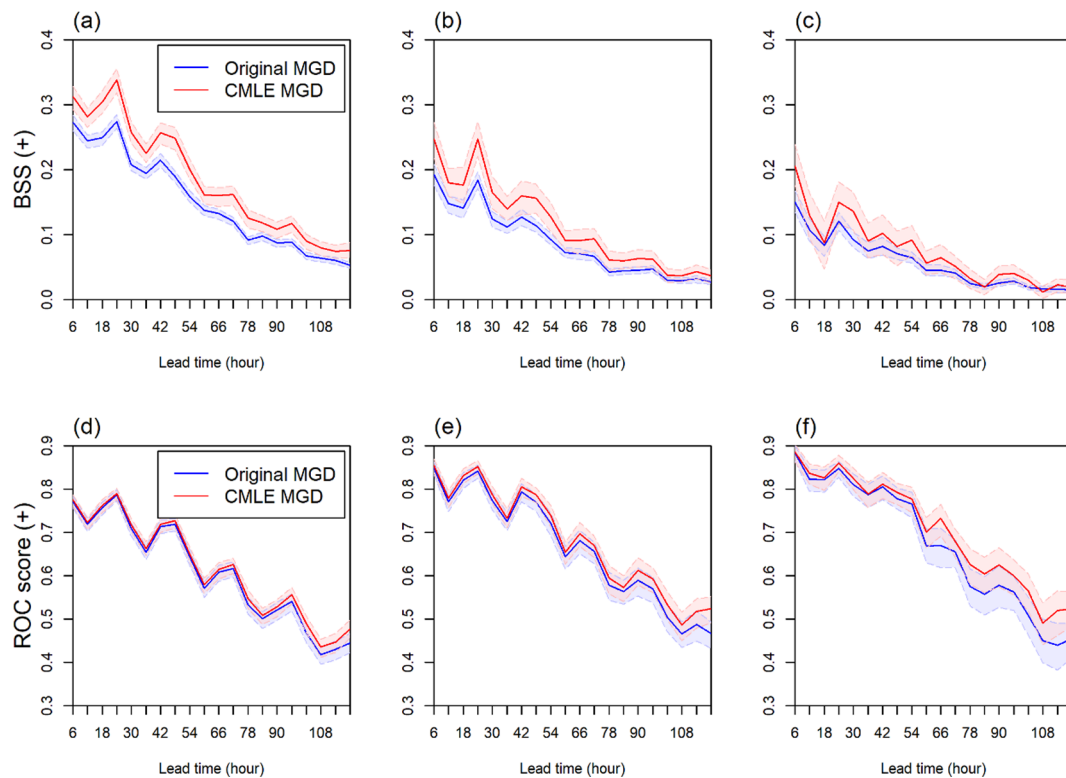
Fig. 6. (a)–(c) The Brier skill score for the original MGD and CMLE-MGD at thresholds of (a) 85%, (b) 95% and (c) 97.5% quantiles of observations. (d)–(f) The ROC score of the original MGD and CMLE-MGD at thresholds of (d) 85%, (e) 95% and (f) 97.5% quantiles of observations. The 90% confidence intervals by bootstrapping are shown for each metric.

2011).

On the contrary, the correlation coefficients can be well estimated by the CMLE method, because zero precipitation values are modeled as left-censored data at the threshold of zero precipitation in CMLE-MGD. The CMLE method has been successfully applied in other studies for post-processing of precipitation forecasts, such as in HCLR (Messner et al., 2014a) and BJP (Robertson et al., 2013). The CMLE-MGD's improvement over the original MGD for sub-daily precipitation events is meaningful in hydrological applications, because accurate and reliable precipitation forecasts of high temporal resolution are very important in applications such as flood forecasting.

There are still limitations in the proposed CMLE-MGD model, which may lead to the overestimation of CMLE-MGD in terms of bias (Fig. 4a) and reliability for heavy rain (Fig. 5). As Fig. 3 shows, the estimated correlation coefficients by CMLE-MGD are slightly higher than the "optimal" correlation coefficients that minimize CRPS, which may result in the overestimation of CMLE-MGD. A possible reason for the imperfect estimation of the correlation coefficients is the joint distribution of transformed forecasts and observations may not be bivariate Normal distribution (Wu et al., 2011). In the future, more flexible tools such as copulas can be applied to model the joint distribution of forecasts and observations (Khajehei and Moradkhani, 2017).

Moreover, the current version of CMLE-MGD is based on the parametric NQT in this study. However, the performance of the parametric NQT is influenced by the goodness-of-fit of the marginal distribution. If the fitting of the marginal distribution is not satisfying, the transformed variables may not follow standard Normal distribution, which will further deteriorate the modeling of the joint distribution in transformed space. In the future, we will test other transformations or non-parametric NQT to achieve better performance. Besides, Wu et al. (2011) demonstrated that the mixed-type MGD performed better than the original MGD developed by Schaake et al. (2007a). The comparison of

the mixed-type MGD and the proposed CMLE-MGD is also needed in the future.

In addition, we only used ensemble mean as predictor in this work. There are several methods that utilize all the ensemble members or the ensemble spread as predictors, such as the member-by-member (MBM) (Van Schaeybroeck and Vannitsem, 2015) and EMOS (Gneiting et al., 2005; Scheuerer and Hamill, 2015). We will improve the MGD models to make better use of the information from the ensemble forecasts to address the heteroscedasticity problem in the future.

6. Summary and conclusions

Raw forecasts from numerical weather models (NWP) can be improved significantly through post-processing to remove the systematic bias and obtain better forecast skill. The meta-Gaussian distribution (MGD) model is one of the early successful methods. However, recent studies have shown that the original MGD may suffer from reliability problems for 6-hourly precipitation forecasts. In this paper, we proposed an improved version of MGD model, namely the censored maximum likelihood estimation-based MGD (CMLE-MGD). We evaluated the proposed method using the GEFS reforecasts of 6-hourly precipitation. The results show that the proposed CMLE-MGD outperforms the original MGD in terms of Brier skill score and overall reliability. The proposed method can achieve similar performance with the state-of-the-art CSGD-EMOS if both use ensemble mean as the only predictor. The proposed CMLE-MGD can achieve reliable sub-daily precipitation forecasts, and it can be used in applications such as flood forecasting that needs precipitations forecasts of high temporal resolution. More studies are still needed to test the performance of the new method in different climates.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Appendix A. Verification metrics

A.1. Relative mean error

The relative mean error (RME) measures the average difference between the ensemble mean forecast $\bar{f_i}$ and corresponding observation o_i as a fraction of the average observation, namely,

$$\text{RME} = \sum_{i=1}^{n} (\bar{f_i} - o_i)/ \sum_{i=1}^{n} o_i \tag{A.1}$$

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

A.2. Brier score and Brier skill score

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

$$\text{BS} = \frac{1}{n} \sum_{i=1}^{n} (F_{f_i}(q) - F_{o_i}(q))^2 \tag{A.2}$$

where $F_{f_i}(q)$ is the probabilities of 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, namely,

$$F_{o_i}(q) = \begin{cases} 1, & o_i > q \\ 0, & otherwise \end{cases} \tag{A.3}$$

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

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{ref}}} \tag{A.4}$$

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

A.3. Continuous ranked probability 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,

$$\text{CRPS} = \int_{-\infty}^{\infty} (F_f(q) - F_o(q))^2 dq \tag{A.5}$$

Then, the mean CRPS of all samples can be computed from the average of the CRPS of all forecast and observation pairs.

A.4. Stratified PIT histogram

The probability integral transform (PIT) is the predictive cumulative distribution function (CDF) F_t at the corresponding observation y_t as follows,

$$\text{PIT}_t = F_t(y_t) \tag{A.6}$$

The PIT for reliable forecasts follows a uniform distribution (Gneiting et al., 2007), so the corresponding PIT histogram for reliable forecasts should be flat. When the observations are equal or below the threshold of zero precipitation (0.1 mm/day in this study), a pseudo-PIT value is generated from a uniform distribution with the range of $[0, F_t(y_{threshold})]$, where $y_{threshold}$ is the threshold of zero precipitation for observation (Wang and Robertson, 2011).

A.5. 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,

$$\text{ROCscore} = \frac{\text{AUC} - \text{AUC}_{\text{ref}}}{1 - \text{AUC}_{\text{ref}}} \tag{A.7}$$

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

Appendix B. CSGD-EMOS model

CSGD-EMOS is designed for post-processing of precipitation forecasts based on the EMOS scheme (Gneiting et al., 2005; Scheuerer and Hamill, 2015). It firstly fits the climatology distribution of observation using the censored, shifted Gamma distribution (CSGD). Then, the CSGD-based regression model can be fitted using forecasts and corresponding observations. To make a fair comparison with MGD, here we use ensemble mean as the only predictor in the regression model. The mean and standard deviation of the conditional distribution of observations given forecasts are modeled in CSGD-EMOS as follows (Scheuerer and Hamill, 2015),

$$\mu = \frac{\mu_{cl}}{\alpha_1} \log 1p \left[\operatorname{expm1}(\alpha_1) \left(\alpha_2 + \alpha_3 \frac{\bar{f}}{\bar{f}_{cl}} \right) \right] \tag{B.1}$$

$$\sigma = \alpha_4 \sigma_{cl} \sqrt{\frac{\mu}{\mu_{cl}}} \tag{B.2}$$

where $\alpha_1, \cdots, \alpha_4$ are the regression parameters, $\bar{f}$ is the ensemble mean of raw forecasts, $\bar{f}_{cl}$ 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 $\operatorname{expm1}(x) = \exp(x) - 1$. More details about CSGD-EMOS can be found in Scheuerer and Hamill (2015).

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2019.04.073.

References

Anderson, T.W., Darling, D.A., 1954. A test of goodness of fit. J. Am. Stat. Assoc. 49, 765–769.

Bellier, J., Zin, I., Bontron, G., 2017. Sample stratification in verification of ensemble forecasts of continuous scalar variables: potential benefits and pitfalls. Mon. Wea. Rev. 145, 3529–3544.

Berrocal, V.J., Raftery, A.E., Gneiting, T., 2008. Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. Ann. Appl. Stat. 2 (4), 1170–1193.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Wea. Rev. 78 (1), 1–3.

Brown, J.D., Demargne, J., Seo, D.-J., Liu, Y., 2010. The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ. Modell. Softw. 25 (7), 854–872.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., Wilby, R., 2004. The Schaake shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. J. Hydrometeorol. 5 (1), 243–262.

Dawid, A.P., 1984. Statistical theory: the prequential approach. J. Roy. Stat. Soc. A 147, 278–292.

Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y., 2014. The science of NOAA's operational hydrologic ensemble forecast service. Bull. Am. Meteorol. Soc. 95 (1), 79–98.

Gel, Y., Raftery, A.E., Gneiting, T., 2004. Calibrated probabilistic mesoscale weather field forecasting. J. Am. Stat. Assoc. 99 (467), 575–583.

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annu. Rev. Stat. Appl. 1, 125–151.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102 (477), 359–378.

Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Wea. Rev. 133 (5), 1098–1118.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. 69 (2), 243–268.

Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau, T.J., Zhu, Y., Lapenta, W., 2013. NOAA's second-generation global medium-range ensemble reforecast dataset. Bull. Am. Meteorol. Soc. 94 (10), 1553–1565.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Wea. Forecasting 15 (5), 559–570.

Joseph, B., Guillaume, B., Isabella, Z., 2017. Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting. Water Resour. Res. 53 (12), 10085–10107.

Kelly, K.S., Krzysztofowicz, R., 1997. A bivariate meta-Gaussian density for use in hydrology. Stochastic Hydrol. Hydraul. 11 (1), 17–31.

Khajehei, S., Moradkhani, H., 2017. Towards an improved ensemble precipitation forecast: a probabilistic post-processing approach. J. Hydrol. 546, 476–489.

Krzysztofowicz, R., Evans, W.B., 2008. Probabilistic forecasts from the national digital forecast database. Wea. Forecasting 23 (2), 270–289.

Lerch, S., Thorarinsdottir, T.L., Ravazzolo, F., Gneiting, T., 2017. Forecasters dilemma: extreme events and forecast evaluation. Stat. Sci. 32, 106–127.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., Di, Z., 2017. A review on statistical post-processing methods for hydrometeorological ensemble forecasting. Wiley Interdisciplinary Rev. Water 4, e1246.

Liu, Y., Duan, Q., Zhao, L., Ye, A., Tao, Y., Miao, C., Mu, X., Schaake, J.C., 2013. Evaluating the predictive skill of post-processed NCEP GFS ensemble precipitation forecasts in China's Huai river basin. Hydrol. Processes 27 (1), 57–74.

Messner, J.W., Mayr, G.J., Zeileis, A., Wilks, D.S., 2014b. Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. Mon. Wea. Rev. 142 (1), 448–456.

Messner, J.W., Mayr, G.J., Wilks, D.S., Zeileis, A., 2014a. Extending extended logistic regression: extended vs. separate vs. ordered vs. censored. Mon. Wea. Rev. 142 (8), 3003–3013.

Möller, A., Lenkoski, A., Thorarinsdottir, T.L., 2013. Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. Q. J. Roy. Meteor. Soc. 139 (673), 982–991.

Papalexiou, S.M., 2018. Unified theory for stochastic modelling of hydroclimatic processes: preserving marginal distributions, correlation structures, and intermittency. Adv. Water Resour. 115, 234–252.

Papalexiou, S.M., AghaKouchak, A., Foufoula-Georgiou, E., 2018. A diagnostic framework for understanding climatology of tails of hourly precipitation extremes in the United States. Water Resour. Res. https://doi.org/10.1029/2018WR022732.

Papalexiou, S.M., Koutsoyiannis, D., 2016. A global survey on the seasonal variation of the marginal distribution of daily precipitation. Adv. Water Resour. 94, 131–145.

Pinson, P., Girard, R., 2011. Evaluating scenarios of short-term wind power generation. Wind Energy 12 (1), 51–62.

Politis, D., Romano, J.P., 1994. The stationarybootstrap. J. Am. Stat. Assoc. 89, 1303–1313.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Wea. Rev. 133 (5), 1155–1174.

Robertson, D.E., Shrestha, D.L., Wang, Q.J., 2013. Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. Hydrol. Earth Syst. Sci. 17 (9), 3587–3603.

Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., Seo, D.J., 2007a. Precipitation and temperature ensemble forecasts from single-value forecasts. Hydrol. Earth Syst. Sci. Discuss. 4, 655–717.

Schaake, J.C., Hamill, T.M., Buizza, R., Clark, M., 2007b. HEPEX: the hydrological ensemble prediction experiment. Bull. Am. Meteorol. Soc. 88 (10), 1541–1547.

Schaake, J., Pailleux, J., Thielen, J., Arritt, R., Hamill, T., Luo, L., Martin, E., McCollor, D., Pappenberger, F., 2010. Summary of recommendations of the first workshop on postprocessing and downscaling atmospheric forecasts for Hydrologic applications held at Meteo-France, Toulouse, France, 15–18 June 2009. Atmos. Sci. Lett. 11 (2), 59–63.

Schefzik, R., 2015. A similarity-based implementation of the Schaake shuffle. Mon. Wea. Rev. 144 (5), 1909–1921.

Schefzik, R., Thorarinsdottir, T.L., Gneiting, T., 2013. Uncertainty quantification in complex simulation models using ensemble copula coupling. Stat. Sci. 28 (4), 616–640.

Schepen, A., Wang, Q.J., Robertson, D.E., 2016. Application to post-processing of meteorological seasonal forecasting. In: Duan, Q. (Ed.), Handbook of Hydrometeorological Ensemble Forecasting. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–29.

Scheuerer, M., Hamill, T.M., 2015. Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. Mon. Wea. Rev. 143 (11), 4578–4596.

Scheuerer, M., Hamill, T.M., Whitin, B., He, M., Henkel, A., 2017. A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. Water Resour. Res. 53 (4), 3029–3046.

Shen, Y., Pan, Y., Yu, J., 2014. A high spatiotemporal gauge-satellite merged precipitation analysis over China. J. Geophys. Res. Atmos. 119 (6), 3063–3075.

Shrestha, D.L., Robertson, D.E., Bennett, J.C., Wang, Q.J., 2015. Improving precipitation forecasts by generating ensembles through postprocessing. Mon. Wea. Rev. 143 (9),

3642–3663.

Sloughter, J.M., Raftery, A.E., Gneiting, T., Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon. Wea. Rev. 135 (9), 3209–3220.

Tao, Y., Duan, Q., Ye, A., Gong, W., Di, Z., Xiao, M., Hsu, K., 2014. An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai river basin. J. Hydrol. 519 (Part D), 2890–2905.

Van Schaeybroeck, B., Vannitsem, S.p., 2015. Ensemble post-processing using member-by-member approaches: theoretical aspects. Q. J. Roy. Meteor. Soc. 141 (688), 807–818.

Wang, Q.J., Robertson, D.E., Chiew, F.H.S., 2009. A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. Water Resour. Res. 45 (5), 1–18.

Wang, Q.J., Robertson, D.E., 2011. Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. Water Resour. Res. 47, W02546.

Wang, Q.J., Shrestha, D.L., Robertson, D.E., Pokhrel, P., 2012. A log-sinh transformation for data normalization and variance stabilization. Water Resour. Res. 48 (5), 1–7.

Wilks, D.S., 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. Meteorol. Appl. 16, 361–368.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences, third ed. Elsevier.

Wu, L., Seo, D., Demargne, J., Brown, J., Cong, S., Schaake, J., 2011. Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. J. Hydrol. 399 (3–4), 281–298.

Wu, L., Zhang, Y., Adams, T., Lee, H., Liu, Y., Schaake, J., 2018. Comparative evaluation of three Schaake shuffle schemes in postprocessing GEFS precipitation ensemble forecasts. J. Hydrometeorol. 19 (3), 575–598.

Ye, A., Deng, X., Ma, F., Duan, Q., Zhou, Z., Du, C., 2017. Integrating weather and climate predictions for seamless hydrologic ensemble forecasting: a case study in the Yalong River basin. J. Hydrol. 547, 196–207.

Ye, L., Hanson, L.S., Ding, P., Wang, D., Vogel, R.M., 2018. The probability distribution of daily precipitation at the point and catchment scales in the United States. Hydrol. Earth Syst. Sci. Discuss. 22 (12), 6519–6531.

Zhang, Y., Wu, L., Scheuerer, M., Schaake, J., Kongoli, C., 2017. Comparison of probabilistic quantitative precipitation forecasts from two postprocessing mechanisms. J. Hydrometeorol. 18 (11), 2873–2891.