A Nonhomogeneous Regression-Based Statistical Postprocessing Scheme for Generating Probabilistic Quantitative Precipitation Forecast

MOHAMMADVAGHEF GHAZVINIAN, YU ZHANG, AND DONG-JUN SEO

Department of Civil Engineering, The University of Texas at Arlington, Arlington, Texas

(Manuscript received 24 January 2020, in final form 20 June 2020)

ABSTRACT: This paper introduces a new, two-part scheme for postprocessing single-valued precipitation forecast to create probabilistic quantitative precipitation forecast (PQPF). This scheme, herein referred to as the mixed-type nonhomogeneous regression (MNHR), combines the use of logistic regression for estimating rainfall intermittency and nonhomogeneous regression for estimation of additional parameters of the conditional distribution. The performance of MNHR is evaluated relative to operational mixed-type meta-Gaussian distribution (MMGD) and the censored, shifted gamma distribution (CSGD) in postprocessing Global Ensemble Forecast System (GEFS) reforecasts averaged over 25 watersheds in the American River basin in California. The results point to superior performance of MNHR relative to MMGD and CSGD in terms of the skill of postprocessed PQPFs at 24- and 96-h accumulation windows. In addition, it is observed that the performance of CSGD tends to trail behind MNHR and MMGD at least for the 24-h window, though the performance differences tend to narrow at higher forecast amounts and longer lead times. Our analyses suggest that CSGD's underperformance arises partly from its tendency to inflate the shift parameter estimates, which is pronounced over the study site possibly because of infrequent rainfall occurrence. By contrast, MNHR's use of logistic regression helps avoid such bias, and its formulation of conditional distribution addresses the lack of skewness of MMGD for higher forecast amounts. Moreover, MHNR-based PQPF exhibits both superior calibration and relatively high sharpness at short lead times and on an unconditional sense, whereas it features lower sharpness relative to the other two suites when conditioned on higher forecast amount. This trade-off between calibration and *conditional* sharpness warrants further research.

KEYWORDS: Ensembles; Forecast verification/skill; Forecasting; Probability forecasts/models/distribution; Short-range prediction; Statistical forecasting

1. Introduction

Generating reliable and skillful ensemble of precipitation forecasts is crucial to water resources management and risk assessment purposes. Ensemble precipitation forecasts from numerical weather prediction (NWP) dynamic models (often referred to *raw forecasts*), generated through perturbing initial conditions and factoring in diverse physical schemes thus far do not fully represent the uncertainty in forecast for the following reasons: 1) the dynamic model forecast can be systematically biased; 2) the raw ensemble tends to exhibit ''underspread,'' i.e., its members closely resemble each other in terms of intensity and geographic location; and 3) the ensemble is often too small in size to infer the unknown underlying distribution. These issues motivated the development of statistical postprocessing mechanisms (Hamill et al. 2017; Wilks 2018). In the past few decades, many such mechanisms have emerged (Li et al. 2017). These include the analog method (Hamill and Whitaker 2006; Hamill et al. 2015), logistic regression and its extended versions (Hamill et al. 2004; Wilks 2009; Messner et al. 2014a,b; Mazrooei and Sankarasubramanian 2017), mixed-type meta-Gaussian model (MMGD; Herr and Krzysztofowicz 2005; Wu et al. 2011), Bayesian joint probability (Robertson et al. 2013; Li et al. 2020), Bayesian model averaging (Sloughter et al. 2007), and methods that fall under the broad category of ensemble model output statistics (EMOS; Gneiting et al. 2005). An overview of strengths and weaknesses of each technique can be found in Wilks (2018).

Among the aforementioned techniques, EMOS, or nonhomogeneous regression schemes, are receiving a lot of attention at present. Recent developments include censored generalized extreme value distribution (Scheuerer 2014; Hemri et al. 2014), censored Gaussian and censored logistic distribution (Messner et al. 2014a; Gebetsberger et al. 2017; Stauffer et al. 2017), and the censored, shifted gamma distribution (CSGD; Scheuerer and Hamill 2015; Baran and Nemoda 2016). These techniques represent the discontinuous–continuous nature of precipitation using left-censored distributions and rely on heteroscedastic distributional regression to derive distribution parameters. Unlike logistic regression and its extended version (Wilks 2009), EMOS variants offer the flexibility of incorporating ensemble attributes such as spread, as well as other forecast variables as additional predictors. Perhaps the most promising of EMOS variants is the CSGD (Scheuerer and Hamill 2015). CSGD accounts for heteroscedasticity explicitly without resorting to variable transformation that is a common feature in many postprocessing methods, and thereby avoids the potential distortion of ensemble attributes as a result of the transformation.

While the prowess of EMOS schemes and their postprocessed probabilistic quantitative precipitation forecast

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-20-0019.s1.

Corresponding author: Mohammadvaghef Ghazvinian, mohammadvaghef.ghazvinian@mavs.uta.edu

DOI: 10.1175/JHM-D-20-0019.1

(PQPF) has been extensively demonstrated (Zhang et al. 2017; Yang et al. 2017), the adoption of these schemes in *hydrologic forecast operation* has been slow. This is partly because of their relatively high computational overhead and the need of storing ensemble statistics beyond the mean both at real time and in retrospective mode. At present, the operational Hydrologic Ensemble Forecast Service (HEFS; Demargne et al. 2014) at U.S. National Weather Service (NWS) relies on the mixed-type meta-Gaussian distribution (MMGD; Wu et al. 2011) as the mechanism for postprocessing precipitation and temperature forecasts. MMGD uses deterministic forecast, or the mean of ensemble forecast as the predictor, and employs a Gaussian copula to relate truncated marginals of forecasts and observations. While its performance has been generally satisfactory (Wu et al. 2011), the reliability of its PQPF is known to be relatively low for heavy to extreme rainfall events. A recent investigation by Zhang et al. (2017) confirms this, and the authors show that the CSGD, even with ensemble mean as the sole predictor, outperforms the MMGD, and the performance gap is especially wide for heavier precipitation events (>25 mm over 6-h intervals). The comparison highlights the need to identify robust, computationally efficient, single-predictor postprocessing mechanisms that can incorporate advanced features of contemporary EMOS schemes in modeling predictive distribution and accounting for heteroscedasticity.

In this study, we explore an alternative PQPF postprocessing scheme which will be henceforth referred to as mixed-type nonhomogeneous regression (MNHR). MNHR models rainfall with truncated rather than censored distribution commonly used in EMOS schemes (Messner et al. 2014a; Scheuerer and Hamill 2015; Gebetsberger et al. 2017). To establish this truncated distribution, it adopts a two-stage estimation strategy that combines the use of logistic regression for estimating the probability mass at zero, and the use of nonhomogeneous regression for determining the parameters of predictive distribution. As in the simple, single-valued form of CSGD, MNHR introduces heteroscedasticity in the predictive distribution by relating dispersion to predicted rainfall amount. We surmise that 1) the flexibility and robustness as afforded by logistic regression would lead to improved prediction of precipitation occurrence relative to MMGD and CSGD, and 2) a conjunctive use of an ensemble mean-based representation of heteroscedasticity and a heavier tailed distribution would yield PQPFs with higher reliability and sharpness for heavy events relative to those derived using MMGD, and that of skills comparable to those based on CSGD. To test these hypotheses, we perform a set of hindcast experiments to compare the performance of MNHR, MMGD, and the simple form of CSGD over 25 watersheds in the American River basin in California. To closely mimic forecast operation, the schemes are applied to basin averaged rainfall, and only forecast ensemble mean is used as the predictor.

The remainder of this article is as follows. In section 2 we will describe the three statistical postprocessing mechanisms in detail, study area, and the data and evaluation metrics. Section 3 presents the results of the comparison, and section 4 summarizes the findings and concludes the paper.

2. Materials and methods

a. Postprocessing schemes

1) Mixed-type meta-Gaussian distribution

MMGD (Herr and Krzysztofowicz 2005; Wu et al. 2011) was developed by the NWS as a component of HEFS for deriving the predictive distribution of precipitation from single-valued precipitation forecast. The scheme seeks to establish the joint cumulative distribution of forecast X and observation Y: $F(X, Y) = P(X \le x, Y \le y)$, which then serves as the basis for deriving the predictive distribution $F_{Y|X}(y|x) = P(Y \le y|X = x)$, i.e., distribution of the true precipitation amounts conditioned on forecast. The term "mixed" refers to the fact that the model accounts for rainfall intermittence by breaking down the distribution under different combinations of dry and wet conditions as indicated in forecast and observations. The bivariate CDF is a mixture of four components:

$$F(x, y) = P_{00} + P_{10}G_X(x) + P_{01}G_Y(y) + P_{11}D(x, y), \qquad (1)$$

where P_{00}, P_{10}, P_{01}, and P_{11} represent the empirical joint probability of forecast–observation being dry–dry, wet–dry, dry–wet, and wet–wet, respectively. The terms G_Y and G_X are conditional marginal CDFs, i.e., $G_Y(y) = P(Y \le y|X = 0, Y > 0)$, $G_X(x) = P(X \le x|X > 0, Y = 0)$, and $D(x, y) = P(X < x, Y < y|X > 0, Y > 0)$ is the joint CDF given the forecast and observation are both positive. The predictive CDFs given dry and wet forecasts are mixtures of dichotomous and continuous components:

$$F_{Y|X}(y|x)|_{X=0} = a + G_Y(y)(1 - a), \qquad (2)$$

$$F_{Y|X}(y|x)|_{X>0} = c(x) + D_{Y|X=x}[1 - c(x)], \qquad (3)$$

where $c(x) = [g_X(x)P_{10}]/[g_X(x)P_{10} + d_X(x)P_{11}]$. Note the lowercase $g()$ and $d()$ denote density functions of $G()$ and $D()$, respectively. In this formulation, the probability of precipitation (PoP) is represented by $[1 - a]$ and $[1 - c(x)]$ given dry and wet forecast, respectively. The conditional distribution $D_{Y|X(y|x)}$ is estimated via the meta-Gaussian distribution theorem (Kelly and Krzysztofowicz 1997) wherein both predicator and predictand undergo normal quantile transformation (NQT). See the appendix for derivation details of Eqs. (2) and (3) and their components.

As described earlier, the predictive CDFs comprise a mass probability of zero precipitation [represented by a and $c(x)$ conditioned on dry and wet forecast], and the respective continuous CDFs $G_Y(y)$ and $D_{Y|X}$. It should be noted that, as the value of predictor x (forecast precipitation amount) increases, probability $c(x)$ declines, and this results in a decline in the skewness of predictive distribution $F_{Y|X}$. This diminished skewness at higher forecast precipitation amounts may have contributed to the reduced skills of MMGD when applied to forecast of heavy rainfall. Another notable limitation of MMGD is that it employs a rather large parameter set, which can be challenging to estimate when historical reforecast is limited, or in dry climate settings where few instances of rainy conditions are present. In practice, empirically determined distributional parameter values are often subject to substantial uncertainties, and these uncertainties can translate in large

errors in the estimates of PoP and predictive distribution as a whole. In addition, identifying suitable parametric families for characterizing highly right skewed variates is a nontrivial task, and our research suggests that none of the existing parametric families used in HEFS accurately describe the conditional distributions $P[X|X>0, Y>0]$ and $P[Y|X>0, Y>0]$, though Pearson type III performs better than others.

2) CENSORED, SHIFTED GAMMA DISTRIBUTION

CSGD was first introduced by Scheuerer and Hamill (2015). This technique and its variants have been examined in a number of studies (Baran and Nemoda 2016; Scheuerer et al. 2017; Zhang et al. 2017; Scheuerer and Hamill 2018; Baran and Lerch 2018; Taillardat et al. 2019; Li et al. 2019). The PDF and CDF of CSGD take the following forms:

$$f^0_{k,\theta,\xi}(y) = \frac{(y-\xi)^{k-1}\exp\left[\frac{-(y-\xi)}{\theta}\right]}{\theta^k\Gamma(k)}, \quad y \geq 0, \tag{4}$$

$$F^0_{k,\theta,\xi}(y) = G_{k,\theta}(y-\xi), \quad y \geq 0, \tag{5}$$

where $k>0$ and $\theta>0$ are the shape and scale parameters of the gamma distribution with CDF $G_{k,\theta}(y)$, and $\Gamma(k)$ is the value of gamma function at k. The shift parameter $\xi<0$ guarantees left censoring at zero by extending the PDF to negative hypothetical precipitation values to produce the point mass at zero. Precipitation quantiles of CSGD for any $0 \leq p_i < 1$ can be derived by

$$q_{p_i} = \max[0, \xi + G^{-1}_{k,\theta}(p)]. \tag{6}$$

To generate postprocessed PQPFs, one first derives the three CSGD climatological parameters: $\mu_{\text{cl}} = k_{\text{cl}}\theta_{\text{cl}}$, $\sigma_{\text{cl}} = \sqrt{k_{\text{cl}}}\theta$, and ξ_{cl} using precipitation analysis. Then, one performs nonhomogeneous regression to relate statistics of archived ensemble forecast to parameters (μ and σ) of predictive CSGD. The regression coefficients thus derived would be applied to the ensemble statistics of real-time forecast to obtain predictive CSGD. It must be noted that in the CSGD formulation, the shift parameter ξ of the predictive distribution is not related to real-time forecast and is kept identical to the climatological shift.

In this study, we implement the simple version of CSGD in which ensemble mean serves as the only predictor to be consistent with the predictor selection of both MMGD and the new scheme MNHR. The CSGD regression equations are

$$\mu = \frac{\mu_{\text{cl}}}{a_1}\,\text{log1}p\left[\text{expm1}(a_1)\left(a_2 + a_3\frac{\overline{f}}{\overline{f}_{\text{cl}}}\right)\right],$$
$$\sigma = a_4\sigma_{\text{cl}}\left(\frac{\mu}{\mu_{\text{cl}}}\right)^{a_5}, \tag{7}$$

where $\text{expm1}(x) = \exp(x) - 1$, $\text{log1}p(x) = \log(1+x)$, and $\overline{f}$ and $\overline{f}_{\text{cl}}$ correspond to the raw ensemble mean forecast and their climatological mean, respectively.

A possible weakness of CSGD is its dependence upon the climatological shift parameter. Our analysis indicates that CRPS minimization often produces biased, inflated values of shift parameter that result in a negative bias in PoP. Figure 1 suggests that this bias regardless of basin (headwater and downstream), is persistent in higher accumulation interval (96 h) where the amounts of zero values are relatively small. The underestimation of PoP as a result of shift parameter inflation can also be seen in Fig. 3 of Scheuerer and Hamill (2015). We postulate that this affects the performance of predictive CSGD, because the predictive distribution of precipitation for an upcoming weather regime often departs widely from the climatology, and changes in the scale and location parameters are insufficient to capture this departure. This issue will be revisited in the later part of this paper.

3) MIXED-TYPE NONHOMOGENEOUS REGRESSION

To address the aforementioned limitations of MMGD and potentially CSGD, we explore an alternative approach which we term the mixed-type nonhomogeneous regression. MHNR combines the use of logistic regression for determining probability of precipitation, and nonhomogeneous regression for estimating the predictive distribution. Similar to MMGD, MNHR employs a mixed-type predictive distribution for characterizing the forecast–observation relationship, i.e.,

$$F_{Y|X}(y|x)|_{X>0} = c(x) + P(Y \leq y|X=x, X>0, Y>0) \times [1-c(x)]. \tag{8}$$

Rather than employing the more complex and less robust Bayesian model $[g_X(x)P_{10}]/[g_X(x)P_{10} + d_X(x)P_{11}]$ in Eq. (A9), we compute the mass probability of zero precipitation $c(x)$ as a function of cubic rooted ensemble mean forecast using a simple two-parameter logistic regression model (Nelder and Wedderburn 1972):

$$\text{logit}[c(x)] = \log\frac{c(x)}{1-c(x)} = a_0 + a_1 x^{1/3}. \tag{9}$$

Note that the logistic model here can be considered as an EMOS model with Bernoulli conditional distribution. The mean of this distribution is related to the predictor(s) with a logit link function. It is worth noting that including the pairs with forecast equal to zero degrades the performance of our mixture model, and so we postprocess the real-time zero forecasts using the predictive CDF in Eq. (2). In lieu of the Gaussian copula used in MMGD, the predictive CDF $P(Y \leq y|X=x, X>0, Y>0)$ is estimated through distributional regression by selecting a suitable parametric family defined on $\mathbb{R}^+$ whose parameters are linked to the ensemble mean. As in other EMOS models, a key challenge here is to select a suitable parametric distribution that can accurately reflect the uncertainty of forecast over a range of predicted precipitation amounts. To this end, some EMOS schemes adopt computationally expensive regime switching or distribution combining techniques (see, e.g., Bentzien and Friederichs 2012; Lerch and Thorarinsdottir 2013; Baran and Lerch 2016, 2018). To simplify matters, we explore only two-parameter predictive families including gamma (Sloughter et al. 2007; Bentzien and Friederichs 2012), lognormal and inverse Gaussian (Bentzien and Friederichs 2012), and zero truncated Gaussian and logistic distributions (Messner et al. 2016; Scheuerer and Möller 2015). Test for goodness of fit using generalized Akaike

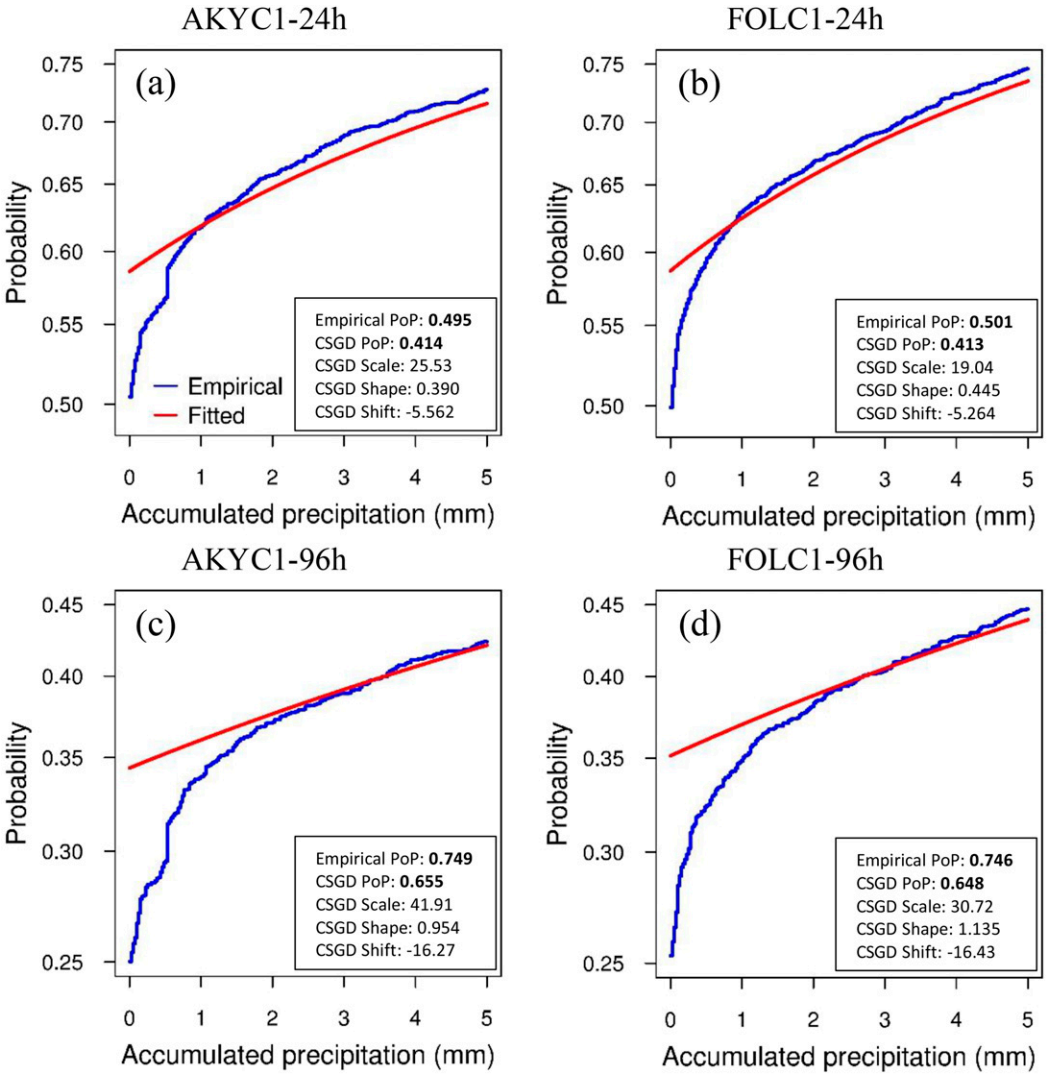

FIG. 1. Empirical vs fitted climatological CDFs of (a),(b) 24- and (c),(d) 96-h accumulation of mean areal precipitation observations over three months (December–February) and over two subbasins of the American River basin, namely, (left) Kyburz, AKYC1 (upper-elevation zone), and (right) Folsom Lake, FOLC1. The CDF values are shown for 0–5-mm accumulated precipitation amounts to magnify discrepancies in estimating the PoP. Also displayed are empirical and CSGD fitted PoP values (in bold), as well as parameters of CSGD. As shown CSGD underestimates the PoP regardless of basin and accumulation intervals.

information criterion (Akaike 1998; Schwarz 1978) shows that zero truncated Gaussian and logistic distributions provide the best fit. We choose the latter since it consistently outperforms in our cross validation, and this outperformance is consistent with the finding of Gebetsberger et al. (2017) that heavier tails of logistic distribution allow it to better represent the occurrence of higher precipitation amounts.

We denote the PDF of zero (left) truncated logistic distribution of the cubic rooted observed precipitation accumulation y by

$$f_{L_0}(y,\mu,\sigma)=\begin{cases}\dfrac{f_L(y;\mu,\sigma)}{1-F_L(0;\mu,\sigma)}, & \text{if} \quad y>0\\ 0, & \text{otherwise}\end{cases}, \tag{10}$$

where $f_L(y;\,\mu,\,\sigma)=(1/\sigma)\{\exp[-(y-\mu)/\sigma]\}\{1+\exp[-(y-\mu)/\sigma]\}^{-2}$ is the PDF of the logistic distribution and F_L is the corresponding CDF.

The regression equations relate scale σ and location μ to ensemble mean:

$$\mu=b_0+b_1x^{1/3}, \tag{11}$$

$$\log(\sigma)=c_0+c_1x^{1/3}, \tag{12}$$

where b_0, b_1, c_0, c_1 are regression coefficients. Equation (12) introduces heteroscedasticity, and the log link function guarantees the nonnegativity of the scale parameter σ. The use of cubic root transformation follows the practice of

TABLE 1. Comparison of the three postprocessing schemes.

	MMGD	MNHR	CSGD
Predictive distribution	Two part	Two part	Censored
Parameter estimation	Goodness of fit	Maximum likelihood	Minimum CRPS
Predictor	Ensemble mean	Ensemble mean	Ensemble mean
Data transformation	NQT	Power (cubic root)	—

Bentzien and Friederichs (2012) and Sloughter et al. (2007), whereas relating scale to ensemble mean the works of Baran and Nemoda (2016) and Lerch and Thorarinsdottir (2013). While Scheuerer and Hamill (2015) illustrated that variable transformation tends to distort the scale of forecast variable, the impacts of distortion are likely muted for MNHR PQPFs since it uses ensemble mean as the sole predictor (M. Scheuerer 2019, personal communication). When employing the logistic predictive distribution, MNHR resembles the heteroscedastic extended logistic regression (HXLR) and its extension heteroscedastic censored logistic regression (HCLR) in form (see, e.g., Messner et al. 2014a,b). The key distinctions of MNHR from the latter two are (i) MNHR uses truncated, rather than censored logistic distribution as adopted in HCLR and (ii) MNHR uses single-predictor by relating scale parameter to ensemble mean, whereas in HXLR and HCLR the scale parameter is a function of ensemble spread.

The logistic and heteroscedastic regression coefficients are estimated separately using the method of maximum likelihood and implemented in R package GAMLSS of Rigby and Stasinopoulos (2005) and Stasinopoulos and Rigby (2007). Table 1 summarizes the major differences between MMGD, CSGD, and MNHR. To isolate the impacts of model structures, we chose not to implement the preprocessing steps that are originally parts of the CSGD (Scheuerer and Hamill 2015), namely, expansion of forecast domain and quantile mapping, in any of the schemes.

b. *Study area and data*

We implement the three schemes, namely, the operational MMGD, CSGD, and MNHR, and cross compare their relative efficacy through a set of cross-validation experiments performed over 25 subbasins (separately for upper- and lower-elevation zones) in the American River basin in the Sierra Nevada (Fig. 2). For each subbasin, we apply each scheme to the areal means of Global Ensemble Forecast System (GEFS) reforecast to derive basin-mean PQPFs that mimic the products of Meteorological Ensemble Forecast Processor (MEFP) of HEFS.

The experimental site for this study, as shown in Fig. 2, comprises watersheds spread along the main stem and major downstream tributaries of the American River. We chose this area as all study basins are located in the service area of the NWS California–Nevada River Forecast Center (CNRFC). CNRFC is an active user of HEFS and its staff routinely produce postprocessed PQPF using MMGD. Assessing the performance of benchmark and competing models can directly inform NWS operation. The subbasins of interest include those whose headwaters are located in the Tahoe and Eldorado National Forests of the Sierra Nevada mountain range. The South Fork American River converges with the combined North and Middle Forks at Folsom Lake east of the city of Sacramento, draining overall 4830 km^2 of upstream area (Seo et al. 2015). Much of the precipitation falls between November and April, with the highest averaged precipitation amount recorded in January. Summers are dry, with the lowest amount of precipitation falling in July (He et al. 2016a,b; Scheuerer et al. 2017). Hereinafter, we will designate November–April as the wet season and May–October as the dry season.

The postprocessing experiment involves applying each scheme to 24- and 96-h (4 day) accumulated precipitation

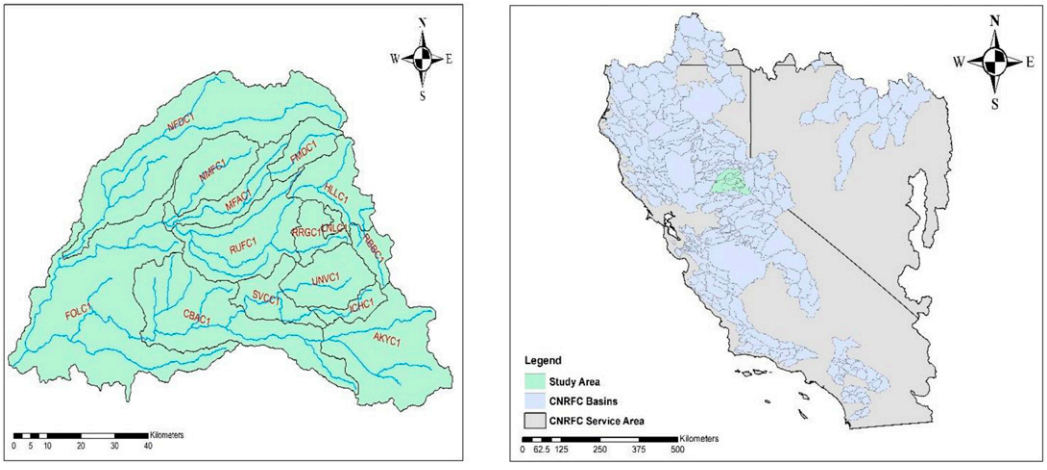

FIG. 2. (left) Map of the study subbasins of American River basin and (right) the locations of American River basin within the service area of CNRFC.

forecast from GEFS. We conduct validation analysis on both 24- and 96-h precipitation accumulation. The choice of 96-h accumulation is based on the fact that the CNRFC also considers skillful prediction of 3–5-day accumulation most important for decision support of reservoir management. Moreover, our analysis shows that the skill in GEFS forecast is the highest at 96-h accumulation window.

A long record of 6-h mean areal precipitation (MAP) analysis is acquired from CNRFC for each of the test basin to serve as ground truth. For forecast, we utilize the GEFS reforecast dataset (Hamill et al. 2013). The GEFS data are generated on a quadratic Gaussian grid ~0.5° resolution for the first 8 days and ~0.67° for 9–16 days ahead forecasts. The reforecasts are composed of 11 ensemble members (10 perturbed and one control) issued every 24 h at 0000 UTC. To be consistent with operational HEFS, we use the nearest GEFS grid node forecasts to basin centroid (Brown et al. 2014; Brown 2015; Seo et al. 2015; Wu 2020). The observation and forecasts were recorded in different time zones (PDT/PST versus UTC). To account for this timing difference, following Brown et al. (2014), we choose observed data whose time stamps closely match those of GEFS forecasts. This approach tolerates a 2-h timing offset between observation and forecast, but we consider the potential impacts of timing discrepancies insignificant as far as 24-/96-h accumulations are concerned. Aggregated observation–ensemble mean forecast data are paired from January 1985 to December 2010 (26 years).

In generating the postprocessed PQPFs within a given month, we follow Wu et al. (2011) in applying a 91-day window centered on the fifteenth of that month of hindcast record to collect forecast–observation pairs. These pairs will serve as the basis for estimating the distributional parameters. The use of moving window helps account for seasonality and expand the sample size for parameter estimation. We fit individual model parameters for each study basin, for each month of the year and up to 7 days ahead (7–10 days in advance for 96-h total accumulation interval), and this yields a validation sample that consists of [(25 basins) × (7 lead times) × (12 months) × (26 years)].

The PQPF produced by each postprocessing scheme is evaluated against the CNRFC observed MAPs using the leave-one-year-out cross validation technique (see, e.g., Hamill et al. 2015; Scheuerer and Hamill 2015; Scheuerer et al. 2017), in which one year of data is set aside for validation and the remainder data are used for parameter estimation. This approach yields 26 years' worth of verified PQPF and corresponding observations for each basin, each month and lead time.

c. Evaluation metrics

1) Proper scoring rules

Proper scoring rules are synthetic measures of sharpness and calibration of probabilistic forecasts (Gneiting et al. 2007). Raw ensemble of forecasts can be sharper (more concentrated) than postprocessed forecast but may not be calibrated (see, e.g., Hamill et al. 2015). If forecasts are perfectly calibrated, observations and random samples of predictive distributions should be exchangeable (Mascaro et al. 2010; Scheuerer 2014; Taillardat et al. 2016; Rasp and Lerch 2018). The proper scoring rules that we apply in this study include continuous ranked probability score (CRPS; Matheson and Winkler 1976; Hersbach 2000), and Brier score (BS; Brier 1950). CRPS is a common measure of overall predictive performance

$$\mathrm{CRPS}(F_i, y_i) = \int_{-\infty}^{\infty} [F_i(x) - I\{y_i \le x\}]^2 \, dx, \tag{13}$$

where F_i denotes the CDF of PQPF at the ith forecast instance and y_i is the verifying observation. The term $I(.)$ is the indicator (step) function that takes the value of 1 if $x \ge y_i$ and 0 elsewhere. BS is useful to assess the performance of probabilistic forecast in predicting binary events

$$\mathrm{BS}_{\tau}(F_i, y_i) = [F_i(\tau) - I\{y_i \ge \tau\}]^2, \tag{14}$$

where $F_i(\tau)$ is the probability of PQPF exceeding the threshold value τ, and $I(.)$ is the indicator (step) function that takes the value 1 if the ith verifying observation exceeds the threshold value and 0 otherwise. To compare the relative performance of PQPFs from three schemes, we produce the continuous ranked probability skill score $[\mathrm{CRPSS} = 1 - (\overline{\mathrm{CRPS}}/\overline{\mathrm{CRPS}}_{\mathrm{ref}})]$ and Brier skill score $[\mathrm{BSS} = 1 - (\overline{\mathrm{BS}}/\overline{\mathrm{BS}}_{\mathrm{ref}})]$ with the MMGD-based forecast as the reference. Positive values of skill score indicate improvement over the MMGD model. As the magnitude of MAPs varies with the size of drainage and month, rather than prescribing "hard" threshold precipitation amounts in calculating BS, we instead choose 97% and 99% quantiles of observed MAPs as thresholds (denoted by q97 and q99, respectively). These thresholds are calculated for each month and basin.

2) Probability integral transform (PIT)

Calibration of PQPFs are assessed using histograms of probability integral transform $\mathrm{PIT}_i = F_i(y_i)$, where F is the predictive CDF of an individual mechanism at the forecast instance i and y_i is the corresponding verifying observation. As indicated earlier in a perfectly probabilistically calibrated PQPF, PIT should follow a standard uniform distribution $U(0, 1)$ with $E(\mathrm{PIT}) = 1/2$ and $\mathrm{Var}(\mathrm{PIT}) = 1/12$ yielding a flat histogram. In addition to visual inspection, we follow Taillardat et al. (2016, 2019) and compare the sample bias and dispersion of probabilistic forecast against corresponding statistics for uniform distribution. For example a $[\overline{\mathrm{PIT}} > (1/2)]$ (negative bias) indicates that the observations on average tend to be larger than the median PQPFs or similarly $[S^2(\mathrm{PIT}) > (1/12)]$ shows that predictive distribution is underdispersive which indicates overpopulation of lower and higher PITs.

We also employ the reliability index (RI; Delle Monache et al. 2006) to quantify the deviation of PIT histograms from uniformity: $\mathrm{RI} = \sum_{i=1}^{M} |f_i - (1/M)|$ where M denotes the number of categories in which the PIT values are populated, and f_i is the relative frequency of PITs in each category. Therefore, perfectly calibrated PQPF should result in RI = 0. It should be noted that, to account for the intermittent nature of precipitation, we estimate a randomized PIT value (Sloughter et al. 2007)

in the cases of zero precipitation observation by generating a random uniform number in the interval $[0, F_i(0)]$.

3) RELIABILITY DIAGRAM

We further investigate the calibration of PQPF for exceeding a range of thresholds graphically using reliability diagrams. Following the practice of previous studies (Bröcker and Smith 2007; Scheuerer et al. 2017) we lump forecasts from all basins and all months. In this study we use 10 probability bins (0–0.1, 0.1–0.2, etc.) to aggregate the forecasts. We further decompose the Brier score (Murphy 1973) to quantify the contribution of reliability and resolution. Recall that the resolution of PQPF depicts its ability to discriminate between events and is identical to sharpness for perfectly reliable forecasts (Jolliffe and Stephenson 2012). In addition, we examine frequency of forecast probabilities for each category which can be utilized to assess the sharpness of PQPF for specific thresholds graphically. Note that sharp forecast is characterized by higher frequencies for the forecast probabilities close to either 0 or 1.

4) PREDICTION INTERVAL WIDTH

The overall sharpness of PQPF can be quantified using average values of prediction interval width (PIW; see, e.g., Sloughter et al. 2007). We choose the 90% PIW which represents the range between the 5th and 95th percentiles of PQPF generated by the real-time forecast. To estimate this value, we evaluate the predictive CDF at the designated quantiles. It should be noted that narrower PIW corresponds to greater sharpness. To facilitate the comparison, we compute relative improvement of the sharpness over that of MMGD using the expression $[1 - (\overline{\text{PIW}}/\overline{\text{PIW}}_{\text{ref}})]$. Obviously, positive values indicate improvement in sharpness.

5) THE DIEBOLD–MARIANO TEST

To assess whether differences in forecast performances are statistically significant, we utilize two-sided Diebold–Mariano (DM; Diebold and Mariano 1995) statistical test. Let us denote $\overline{\Delta} = \overline{S_{F_1} - S_{F_2}}$ the mean of proper scoring rule S differences from two different forecast sources F_1 and F_2 over cross validated forecasts with length n where $\hat{\sigma}_\Delta$is the estimator of asymptotic standard deviation of Δ. Under standard regularity conditions, the test statistic $t_n = \sqrt{n}(\overline{\Delta}/\hat{\sigma}_\Delta)$ follows the standard Gaussian distribution under the null hypothesis of equal predictive performance of two forecast sources. Forecast source F_1 out/underperforms F_2 if the test statistic is negative/positive. Following past studies (Baran and Lerch 2016; Baran and Nemoda 2016; Lerch et al. 2017) we use sample autocovariance of score differences up to lag $k - 1$ as an estimator of $\hat{\sigma}^2_\Delta$ for the k step-ahead forecasts.

3. Results

In this section we present verification results, which are aggregated over all subbasins and months. Detailed results of proper scoring rules based on different seasons and individual subbasins can be found in the online supplemental material. To facilitate the comparisons among the three schemes and highlight the performance of MNHR and CSGD relative to MMGD, we use the PQPF produced by MMGD rather than climatology as the reference. In validating the PQPF suites produced at the 96-h window, we use running (overlapping) windows ending at time increments of 1 day starting from the first 4 days (i.e., the lead times will be +96 h, +120 h, +144 h, …), and thus the outcomes are not entirely independent. As the qualitative differences among the PQPFs produced by the three schemes are broadly similar at 24- and 96-h accumulation intervals, we choose to focus our comparisons on the former and present the results for the 96-h interval only on a limited basis.

a. *CRPSS*

Comparisons of CRPSS based on 24- and 96-h accumulation intervals, aggregated over all subbasins and months are shown in Figs. 3a and 3b, respectively, where the results are stratified by the lead time. In the 24-h accumulation interval, MNHR consistently shows better performance across lead times while CSGD underperforms MMGD. The differences among the three schemes seems to be small in longer lead times but it is still conspicuous. In the 96-h accumulation interval, both MNHR and CSGD perform better than that in the 24-h accumulation interval relative to MMGD. MNHR still generates the most skillful PQPF; however, its gap with MMGD narrows in longer lead times. Interestingly, CSGD compares well with MMGD across lead times and only slightly deviates from the horizontal line. To assess the overall relative performances when events of significance are predicted to occur by the dynamic mode, following Bellier et al. (2017) and Li et al. (2020), we compute CRPSS using a subsample of PQPFs for instances where moderate rainfall is predicted to occur by GEFS. In our work, we use 97% quantile of observed MAP (q97) as the threshold of moderate rainfall for a given month and subbasin. For example, q97 of 24-h MAPs ranges from 0.63 mm day^{-1} in July for the American River Folsom Lake basin, FOLC1, to 61 mm day^{-1} in January for upper-elevation basin of North Fork American River Foresthill, NMFC1. The corresponding q97 threshold values for 96-h MAPs for the same basins range from 14 to 151 mm.

The resulting CRPSS for the subsample are shown in Figs. 3c and 3d for 24- and 96-h windows, respectively. These results confirm that MNHR remains the best performer among the three over the subsample, and, for both windows, the outperformance tends to be more pronounced over the subsample. The CSGD PQPF underperforms the other two by a wide margin and this underperformance is more appreciable at the 96-h window (Fig. 3d). A possible explanation of this is that the MMGD tends to perform well when the distribution of precipitation amounts conditioned on forecast is likely less skewed. This is the case for longer accumulation windows. With good performance of MMGD, the marginal outperformance of MNHR declines, at least for the shorter lead times, whereas the underperformance by CSGD is magnified.

Pairwise two-sided DM test results, calculated based on mean CRPS and over all samples, are summarized in Tables 2 and 3 for 24- and 96-h accumulation intervals, respectively. These include those based on cross validated forecast and observation pairs from all subbasins and months. In this study we consider predictive performance differences as significant

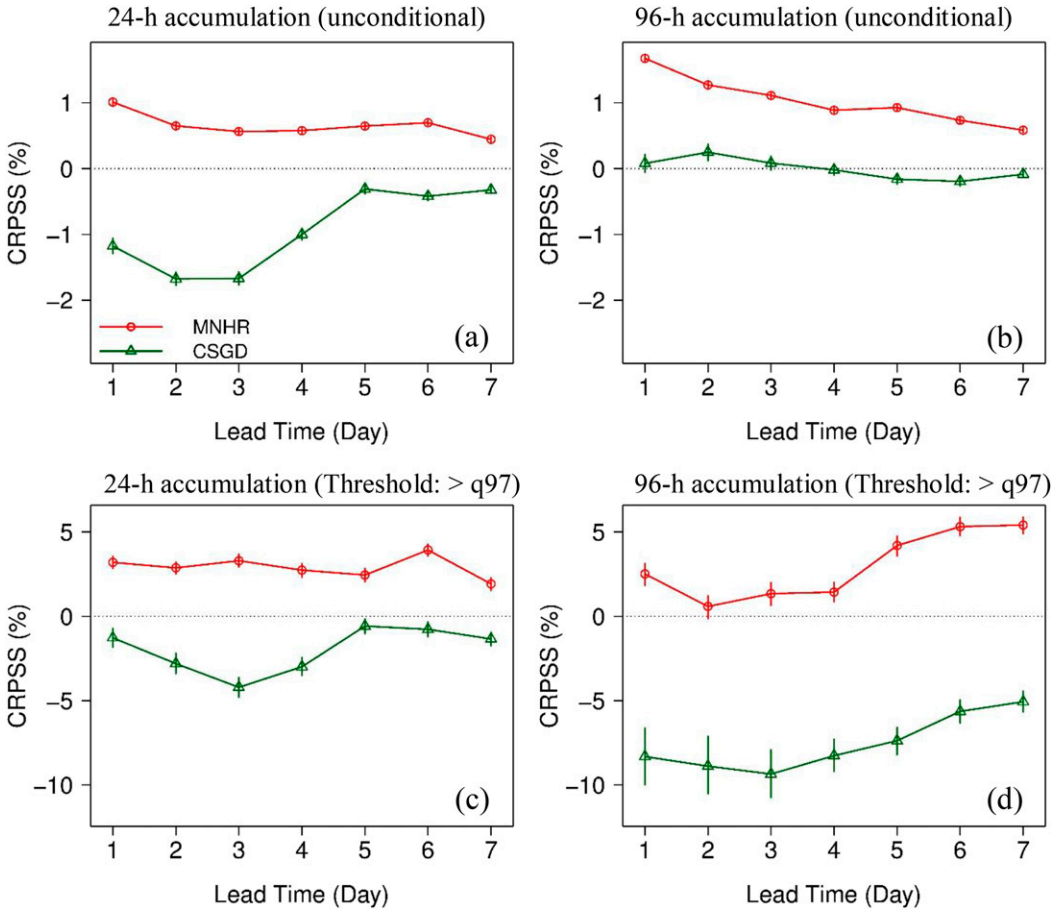

FIG. 3. CRPSS (with 95% bootstrap confidence intervals) averaged over all subbasins and months as a function of lead time calculated based on all observation–forecast pairs for (a) 24-h accumulation interval and (b) 96-h accumulation interval and based on a subsample corresponding raw GEFS mean > 97% quantile of observations for (c) 24-h accumulation interval and (d) 96-h accumulation interval. MMGD is used as the reference forecast.

at p value < 0.05. The results suggest that, in general the outperformance of MNHR relative to MMGD and CSGD is statistically significant in both accumulation windows. For the results at the 96-h window, the test outcome indicates that the outperformance of CSGD over MMGD at 1-day lead time (i.e., first four days forecast) is not significant at 5% level and MMGD significantly outperforms CSGD at 7-day lead time (7–10 days ahead forecasts) but the difference is less significant than that with regard to MNHR.

b. BSS

The BSS values for the three PQPFs and for 24-h accumulation interval are compared in Fig. 4. At the zero threshold (Fig. 4a), it is evident that the MNHR outperforms MMGD and CSGD and this outperformance is more striking in the longer lead times. This confirms our hypothesis that logistic regression is a superior, more robust alternative to the MMGD's Bayesian model in predicting precipitation occurrence. The CSGD trails the other two schemes in performance, likely a result of bias in the shift parameter that was discussed earlier. However, it catches up with MMGD at longer lead times. At higher thresholds (q97 and q99) the following features are clear: 1) MNHR clearly outperforms the other two schemes across lead times (Figs. 4b,c) and 2) both q97 and q99 show similar patterns, but CSGD's underperformance is more striking at the q99 threshold (extreme events).

TABLE 2. DM test statistic (t_n) values based on pairwise mean CRPS differences of PQPFs from postprocessing schemes aggregated over all subbasins and months, for 24-h accumulation interval and for both lead times of 1 and 7 days. Negative/positive values show that PQPF from the model in the corresponding row/column significantly performs better. Values in bold show that the differences are significant at 5% level.

	1 and 7 days (24-h accumulation interval)		
Model	MMGD	MNHR	CSGD
MMGD	—	**23.82**	**−11.86**
MNHR	**−34.85**	—	**−39.32**
CSGD	**19.24**	**38.06**	—

TABLE 3. As in Table 2, but for 96-h accumulation interval. The null hypothesis of equal predictive performance of PQPF from MMGD and CSGD for 1-day lead time cannot be rejected at 5% level ($t_n = -1.13$, p value > 0.05).

	1 and 7 days (96-h accumulation interval)		
Model	MMGD	MNHR	CSGD
MMGD	—	**31.76**	**−3.33**
MNHR	**−54.54**	—	**−31.67**
CSGD	−1.13	**22.97**	—

c. Calibration and sharpness

The overall calibration of 24-h accumulated PQPFs is assessed using PIT histograms and corresponding reliability index and statistics. As populating rank histograms with cross-validated PIT values from all subbasins and all months violates the assumption of independence of samples as suggested by Hamill (2001), we instead develop PIT histograms for each basin, lead time and month separately. Figure 5 displays the PIT for January at 1- and 7-day lead times for two watersheds: 1) a head water basin of south fork American River Kyburz (AKYC1, upper-elevation zone) and 2) the one situated over the most downstream portion of the basin, namely, the Folsom Lake basin (FOLC1). It appears that PQPFs based on MNHR are better calibrated than those from MMGD and CSGD. This performance difference is visible graphically where the histograms deviate only slightly from uniformity, and is also confirmed quantitatively by the deviations in the RI and statistics [$\overline{\text{PIT}}$ and S^2(PIT)] from optimal values indicated earlier. PQPF produced by CSGD on the other hand underperforms, as evidenced by an overall negative bias and overdispersion.

We further assess the calibration of PQPFs globally by examining the distribution of mean RI over all subbasins as a function of lead time (Fig. 6a). Indeed, the mean RI values are lower for both MNHR and MMGD than CSGD reflecting flatter PIT histograms on average. The lead-time dependence of RI varies among the PQPFs from the three schemes. While the RI for MMGD PQPF changes only slightly at long lead time, the performance of CSGD PQPF, interestingly, exhibits a clear downward trend in mean values of RI. It appears that overall calibration of a PQPF suite does not necessarily deteriorate at longer lead times. This feature is also evident in FOLC1 PIT histograms (Fig. 5), where PQPFs based on MNHR and CSGD are better calibrated in 7-day lead time than 1-day lead time in terms of flatness and dispersion.

Figures 6b and 6c display the mean values of $\overline{\text{PIT}}$ and S^2(PIT) where the horizontal lines mark perfect calibration of PQPF in terms of bias and dispersion of (see section 2). A notable feature is that MNHR on average generates the least biased PQPFs across lead times. In general, MMGD consistently generates PQPF with negative bias and CSGD's performance varies among lead times. The mean values of S^2(PIT) shows that MNHR PQPF exhibit better dispersion mostly since the values are closer to the perfectly calibrated population variance Var(PIT) = 1/12. The results also show that MMGD PQPF is underdispersed and this feature becomes more pronounced at longer lead times. Whereas CSGD PQPF exhibits overdispersion at shorter lead time but this tends to diminish at longer lead times.

Reliability diagrams for the resulting PQPFs produced using three schemes at 1- and 7-day lead times and 24-h accumulation are depicted in Figs. 7 and 8, respectively. Three thresholds, i.e., 0 mm, q97, and q99 of MAP observations, are used in computing the reliability diagrams. Following Zhang et al. (2017), we omit the uncertainty information from the decomposition of Brier score since it is independent from the forecast source. For 1-day lead time, MNHR PQPF performs the best across all three thresholds as previously measured by BSS (Fig. 4), and this outperformance over the two competing schemes are attributed to

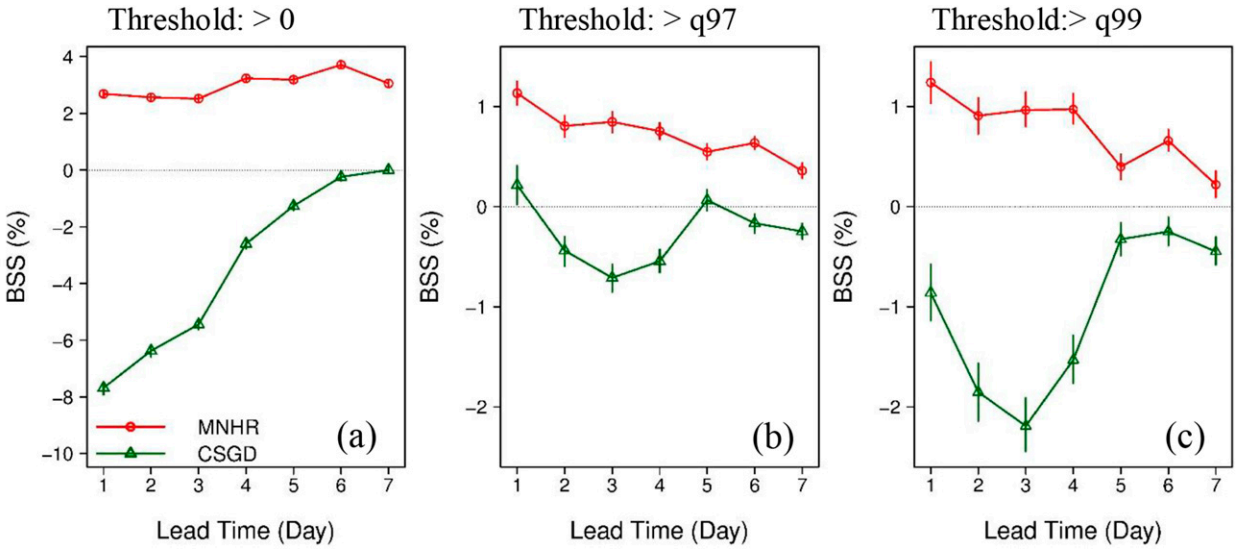

FIG. 4. As in Fig. 3, but for BSS and for 24-h accumulation interval at three different thresholds: (a) >0 (mm) i.e., PoP, (b) >97% quantile of observations and (c) >99% quantile of observations. MMGD is used as the reference forecast.

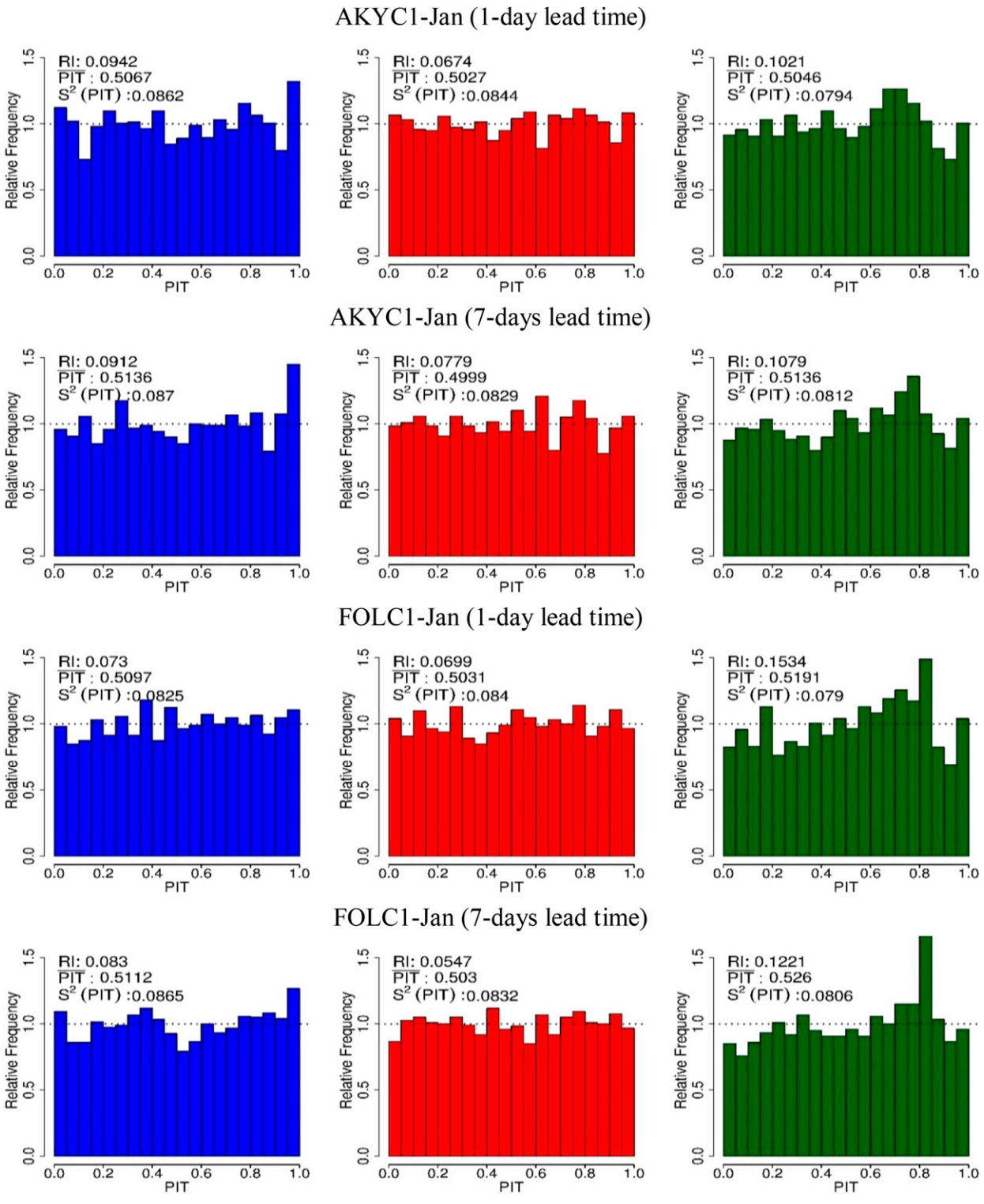

FIG. 5. PIT histograms of postprocessed forecasts at 1- and 7-day lead times (January), for AKYC1 (upper-elevation zone) and FOLC1 and for 24-h accumulation interval. The results for MMGD, MNHR, and CSGD are shown in blue, red, and green, respectively. We follow Sloughter et al. (2007) in using a uniformly distributed random number in the interval $[0, F(0)]$ to calculate the corresponding PIT values for zero observations. Reliability index (RI) and PIT statistics are shown in the upper-left corner of each panel.

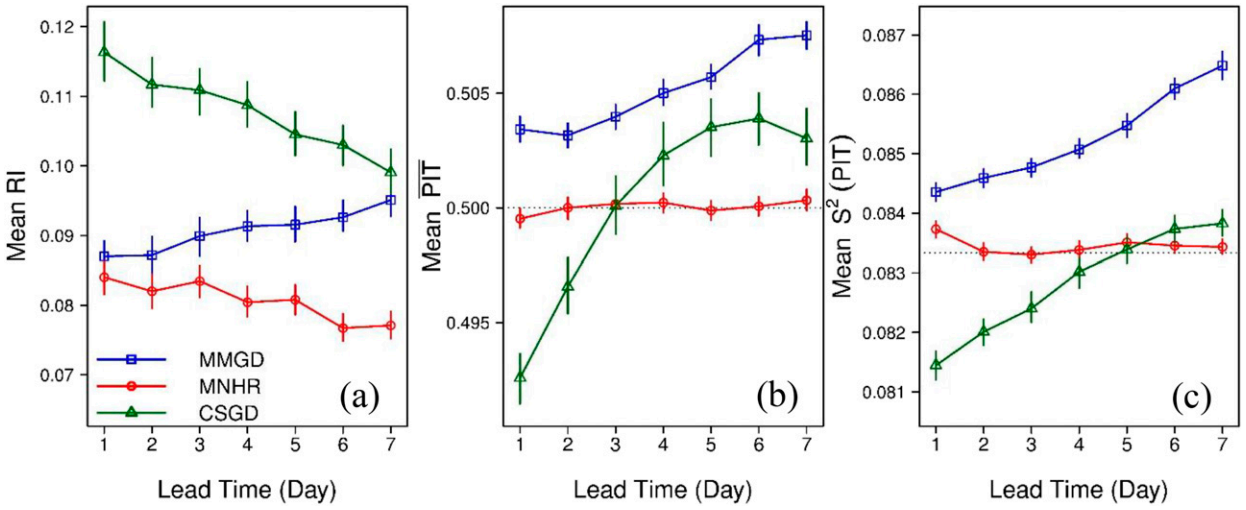

FIG. 6. Average values of (a) RI, (b) $\overline{\text{PIT}}$, and (c) S^2(PIT) (with 95% bootstrap confidence intervals) as a function of lead time and for 24-h accumulation interval. Results are based on individual monthly PIT histograms averaged across subbasins and months. The lower mean RI values of MNHR (closer to zero) indicate that the associated PIT histograms on average are much closer to a perfectly calibrated forecast. PQPF from MNHR is less biased in (b) and relatively well dispersed in (c).

higher reliability and resolution (Figs. 7d–f). Interestingly, among three postprocessed PQPF suites, the CSGD PQPF exhibits the lowest reliability (highest reliability value as designated by "REL") over the 0-mm threshold (Fig. 7g). At this threshold, the reliability curves for CSGD PQPF are consistently above the diagonal at medium–high forecast probability categories. This feature, which indicates that frequency of positive rainfall in each subsample exceeds the forecast probability, is evidently related to the underestimation of PoP by CSGD as discussed earlier. At the q97 threshold, the CSGD PQPF exhibits higher reliability (lower REL) and resolution than that of MMGD PQPF (Figs. 7b,h), thus it is more skillful than MMGD PQPF (Fig. 4b). At the q99 threshold, CSGD and MMGD perform comparably, with both exhibiting a tendency to underforecast at the middle probability range (0.2–0.6). MMGD PQPF is less reliable but exhibits better resolution than that of CSGD (Figs. 7c,i).

At the 7-day lead time, MHNR still outperforms MMGD and CSGD in terms of reliability, resolution, and overall BSS across all thresholds (Figs. 8d–f). MHNR PQPF is nearly perfectly reliable at the 0-mm threshold (Fig. 8d). As pointed out by Scheuerer et al. (2017), by increasing the threshold values, the uncertainty in relative frequencies increases due to a reduction in sample size. Therefore, we exclude the categories with very few instances (e.g., 20) where the prescribed probability threshold is exceeded. At the q97 threshold (Fig. 8b), MMGD produces the sharpest forecasts (highest frequency for forecast probabilities of 0.9–1). It is also interesting to note that CSGD performs similar to MMGD in terms of reliability (Figs. 8b,h). Finally, at the highest threshold (i.e., q99), while MMGD produces the sharpest forecasts (Fig. 8c), these forecasts are again less reliable than those based on CSGD (Fig. 8i), but this lack of reliability is compensated by the higher resolution and the net outcome is higher BSS (Fig. 4c).

Figure 9 compares the overall sharpness of PQPFs produced via MMGD, MNHR, and CSGD as function of lead times based on 24-h accumulation analysis with sharpness defined earlier (see section 2). To facilitate comparisons, the values of *relative sharpness*, i.e., difference in sharpness against MMGD, are shown. As it was done in the CRPSS comparison, we compute two sets of relative sharpness, one using the PQPFs for the entire sample (Fig. 9a), and the other relying only subset corresponding GEFS mean > q97 (Fig. 9b).

In general, as the forecast lead time extends, the uncertainty in PQPF expands and therefore the sharpness declines. Among the three PQPF suites, the CSGD PQPF shows the highest sharpness for 1–3 days lead times (Fig. 9a), but its sharpness drops more rapidly than that of MHNR PQPF, and it is the least sharp beyond 4-day lead time. While CSGD PQPF exhibits better overall sharpness (narrower PIWs) for lead times within the first 3 days, it lacks the necessary probabilistic calibration which leads to poorer overall skill as measured by CRPSS. These observations also suggest that the overall superior predictive performance of MHNR relative to MMGD beyond 3-day lead time is mainly due to its better calibration which more than compensates for the lack of sharpness. As shown in Fig. 9b, the relative sharpness among the schemes clearly changes once only the instances where moderate rainfall is expected are considered. At the q97 threshold, MNHR produces the least sharp PQPF across lead times. CSGD PQPF is the sharpest for shorter lead times (i.e., 1–4 day). Beyond 4 days, MMGD PQPF is consistently the sharpest. The findings corroborate the fact that, although MNHR PQPF is the least sharp (has widest PIWs) for the events with higher forecasts, it is the best calibrated; and this superior calibration outweighs the lack of sharpness to produce higher CRPSS (see Fig. 3c).

4. Discussion and concluding remarks

In this study we propose MNHR, a single-predictor PQPF postprocessing scheme that is based on a truncated predictive distribution. The development of this scheme was motivated by the need to identify single-predictor-based, computationally

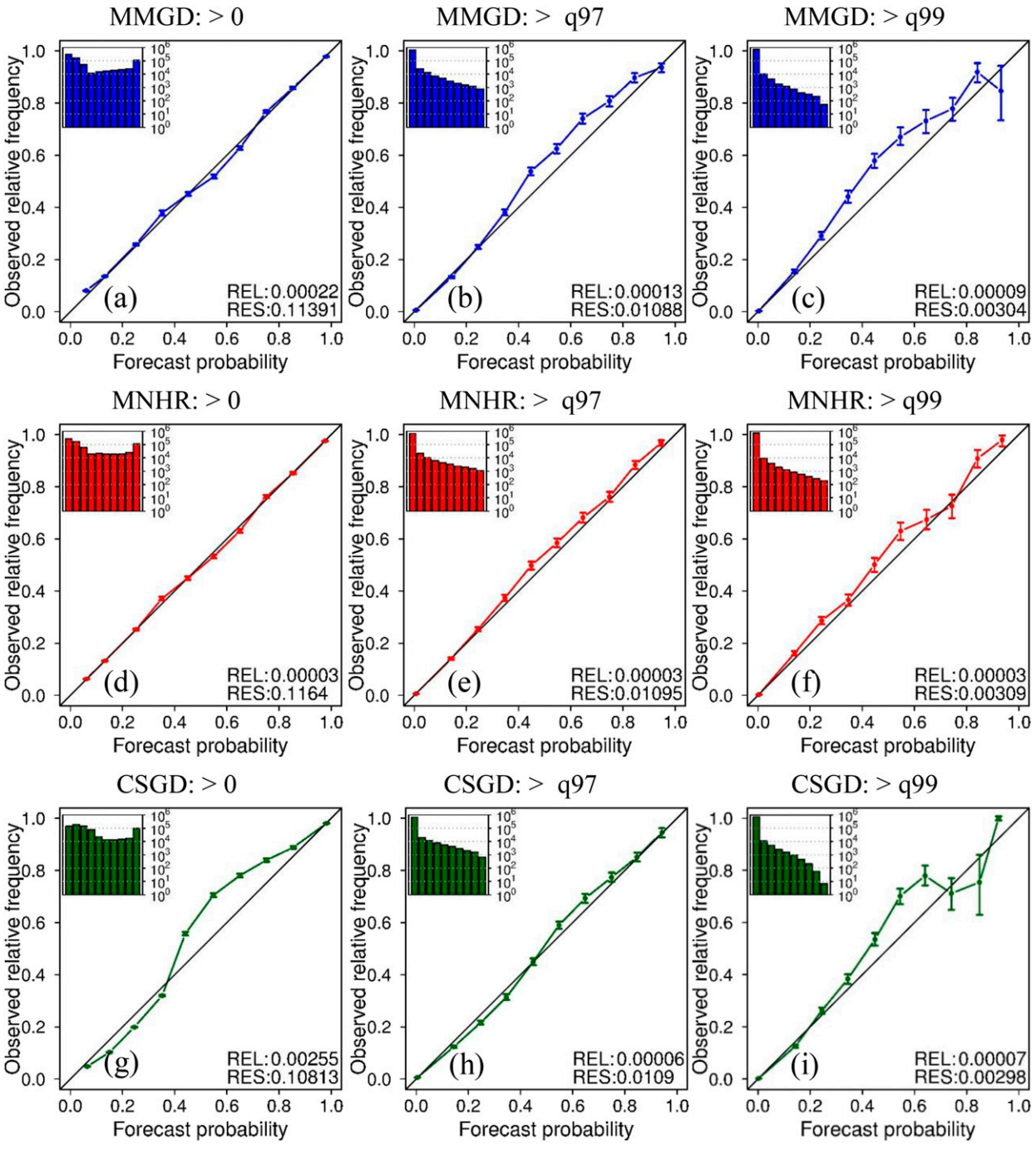

FIG. 7. Reliability diagrams of MMGD, MNHR, and CSGD for the exceedance of three thresholds, i.e., (0 mm, 97% and 99% quantile of observations). Results are based on observation–forecast pairs of all months and cross-validated years aggregated over all subbasins but for 1-day lead time and 24-h accumulation interval. Reliability and resolution values are given in each panel. The inset histograms show the frequencies for each of 10 forecast probability bins in $\log_{10}$ scale and the bars indicate 95% confidence intervals of observed frequencies from bootstrap resampling.

efficient postprocessing methods that can address performance issues of operational MMGD that were highlighted in Zhang et al. (2017). These include a lack of calibration for forecasts of heavy to extreme rainfall amounts, and biases in predicting the rainfall occurrence. These issues can trace their roots to the bivariate normality assumption of MMGD, its rather large parameter set, and the uncertainties embedded in the parametric representation of probability mass at zero.

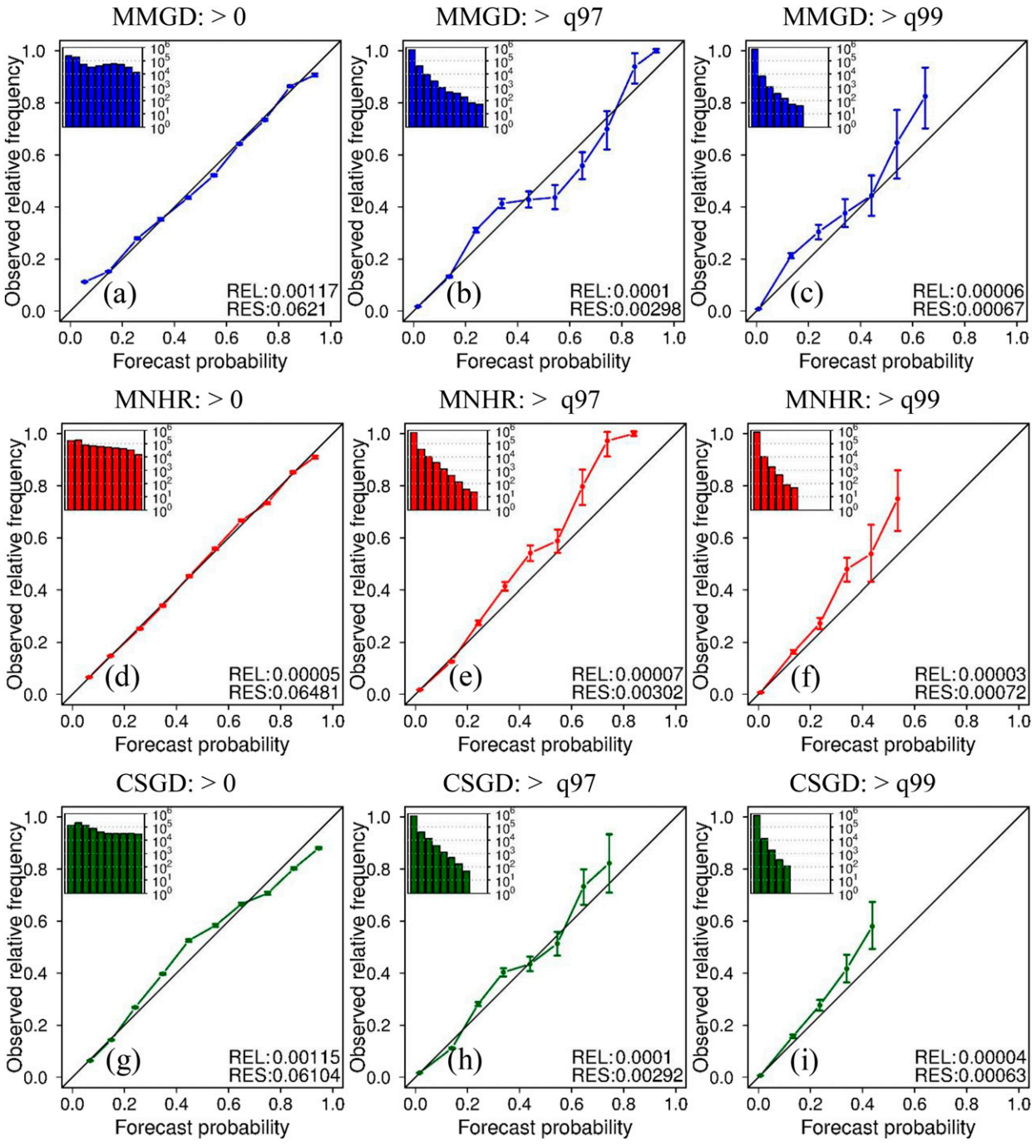

FIG. 8. As in Fig. 7, but for 7 days lead time.

This new scheme retains the two-part estimation framework of MMGD, but it employs logistic regression to derive the probability of precipitation (PoP) and combines it with a truncated logistic distribution to estimate the conditional, or predictive distribution. This combination takes advantage of the simplicity and robustness of logistic regression in predicting categorical events (in case rainy versus dry conditions), and the ability of distributional regression to establish the predictive distribution efficiently and accurately. As in CSGD, another widely known contemporary EMOS scheme, MNHR explicitly models heteroscedasticity in forecast–observation relation by relating dispersion to forecast ensemble mean. As our experience suggests, CSGD's current estimation approach of retaining climatological shift parameter tends to produce a negative bias in PoP estimates. We anticipate that the two-part approach of MNHR makes it less likely for the postprocessed PQPFs to suffer this bias.

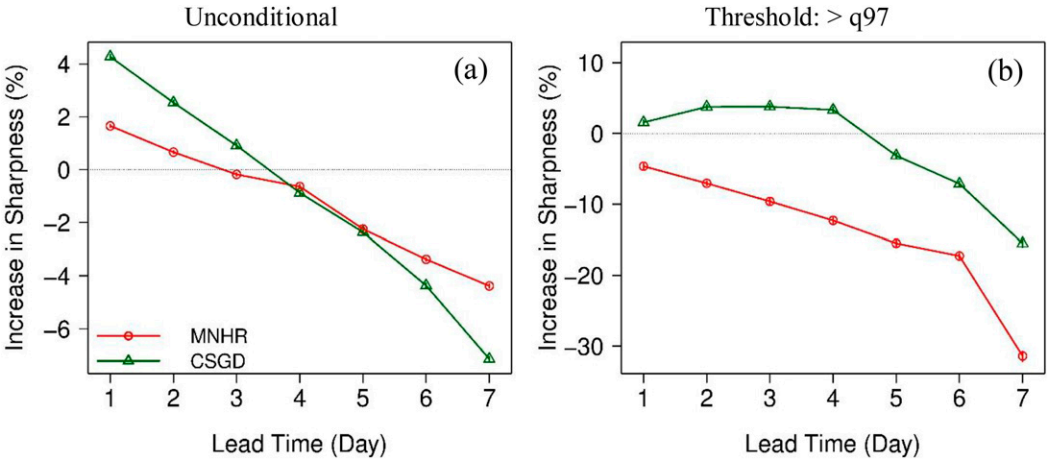

FIG. 9. Increase in sharpness (%) relative to MMGD (dashed line) assessed by average of 90% PIWs from predictive distributions as a function of lead time and based on 24-h accumulation interval. Results are aggregated over all months and subbasins and based on (a) all forecast–observation pairs and (b) subsample corresponding raw GEFS mean > 97% quantile of observations. Given higher raw forecasts, MNHR generates the least sharp PQPF.

We evaluated the performance of MNHR versus MMGD and CSGD through a set of hindcast experiments over the subbasins of American River located in California. We employed the ensemble mean of GEFS precipitation forecast averaged over each subbasin to generate PQPFs using all of the models. The resulting PQPFs were comparatively assessed for their overall skills as measured by CRPSS, calibration, and sharpness. The results show that, in terms of overall skills, the MNHR outperforms MMGD and CSGD across lead times and the performance difference is statistically significant, though the performance gap tends to narrow at longer lead times. Further analysis shows that MHNR yields on average the least biased and the most probabilistically calibrated PQPFs as judged by PIT indices and associated statistics. By contrast, CSGD PQPF shows rather poor calibration, and this to an extent is related to the bias in the PoP that is related to the inflation in the shift parameter. Moreover, MHNR leads in performance in terms of both reliability and resolution for all three precipitation thresholds (i.e., 0 mm, 97% and 99% quantiles), whereas CSGD slightly outperforms MMGD in terms of reliability for heavy to extreme events, though its PQPF has somewhat lower resolution than that from MMGD.

An interesting finding relates to the trade-off between sharpness and calibration. MHNR PQPF is better dispersed, as judged by the variance of PIT, and less biased than the other two PQPF suites. On the other hand, it is not as sharp (assessed by 90% PIW) as those produced using MMGD and CSGD. Another example is that, at shorter lead times (<4 days), 24-h CSGD PQPF features the highest sharpness but the lowest calibration among the three. At longer lead times, the CSGD PQPF becomes less sharp relative to the PQPF by MMGD, and in the meantime its CRPSS approaches that of the latter. Moreover, when the verification is done on subsamples stratified by the magnitude of GEFS forecast, it is shown that the overall outperformance of MNHR, as measured by CRPSS, tends to widen for subsamples with higher forecast amounts. Apparently, improved calibration is often achieved at the expense of reduced sharpness, though in limited cases sharpness and calibration may both improve (e.g., MHNR PQPF being unconditionally sharper at day 1–2 and more reliable; Figs. 6a and 9a).

Our analysis confirms our first postulation that the use of logistic regression in MNHR could produce more skillful PoP estimates relative to those based on MMGD and CSGD. By employing logistic regression, one can relate transformed forecast mean directly to PoP and thereby avoids the complications stemming from the errors in the estimation of, say, marginal distribution as in MMGD. It is also preferable to CSGD for which the inflation of shift parameter is clearly a structural issue that to an extent compromises its robustness. The broad overperformance of MHNR as judged by CRPSS, BSS, and reliability also corroborates our second hypothesis that the joint use of heteroscedastic regression, which links dispersion to ensemble mean, and a heavier tailed predictive distribution would produce PQPFs superior to those based on MMGD. Interestingly, CSGD underperforms both MHNR and MMGD, at least over shorter lead times (say, <4 days), though it also employs heteroscedastic regression to establish a heavy tailed, right skewed distribution. It is yet unclear what contributes to this underperformance, which contrasts starkly to its clear outperformance against MMGD over Mid-Atlantic region of the United States as illustrated by Zhang et al. (2017). A number of factors could have given rise to the differences, including different rainfall characteristics, the omission of preprocessing steps such as quantile mapping as suggested in Scheuerer and Hamill (2015). The climate over the study region over California is much drier than the eastern United States: the city of Sacramento, situated near the outlet of American River basin, has on average of 63 rainy days per year, compared to 110 days over Washington, D.C. There is a possibility that the scarcity of wet forecast–observation pairs, together with inflated shift parameter, result in a distortion in the distributional parameters as derived through CRPS

minimization. In addition, CSGD implemented accounts for heteroscedasticity by relating the dispersion to the distributional mean [Eq. (7)], rather to the ensemble mean itself. As we exclude the ancillary predictor PoP in our implementation, any degradation in the estimates of distribution mean as a result of this exclusion would adversely impact the estimates of dispersion. Our analysis (not shown here) shows that modeling the scale parameter directly as an affine function of power transformed GEFS ensemble mean improves the current CSGD. Moreover, as mentioned earlier, the direct use of climatological shift parameter results in an inflation of shift which further impacts the accuracy of estimates of scale and shape parameters.

To summarize, MHNR, while structurally simple, has been demonstrated to be a potentially robust, and computationally efficient alternative to MMGD, and even CSGD for postprocessing precipitation forecast. We argue that the use of truncated distributions is advantageous to the use of censored distributions in the context of precipitation postprocessing. In the near future, we plan to further assess the performance of MNHR versus CSGD, as well as other established mechanisms that rely on censored distributions, over other parts of CONUS, and on both a gridded and basin-mean basis. We hope such comparisons can shed light on the connection between the relative performance and distribution types (truncated versus censored), rainfall climatology, preprocessing steps (e.g., quantile mapping), and predictor selections, and thereby facilitate the operational adoption of methods for routine hydrometeorological forecasts in the United States.

Acknowledgments. The authors thank the editor and anonymous reviewers for their valuable suggestions to improve the article. The first author was financially supported by the faculty startup package for Dr. Yu Zhang from UT Arlington and NOAA Grant NA18OAR4590370-01. These supports are graciously acknowledged here. We are indebted to Dr. Michael Scheuerer at ESRL who shared the CSGD code and documentation, and offered many insights on CSGD and other EMOS schemes. We would also like to thank CNRFC staff for making available the archived GEFS and MAP analysis which made the study possible.

APPENDIX

Derivation of MMGD Predictive CDF

Let X and Y denote the random variables of single-valued precipitation forecast and observed precipitation amount, respectively. The predictive CDF given dry and wet forecasts can be derived using the Bayes theorem as follows.

a. Dry forecasts

$$F_{Y|X}(y|x)|_{X=0} = P(Y=0|X=0) + P(Y>0, Y\le y|X=0) \tag{A1}$$

$$= \frac{P(Y=0, X=0)}{P(X=0, Y=0) + P(X=0, Y>0)} + \frac{P(Y>0, Y\le y, X=0)}{P(X=0, Y=0) + P(X=0, Y>0)} \tag{A2}$$

$$= \frac{P_{00}}{P_{00}+P_{01}} + \frac{P(Y\le y|Y>0, X=0)}{P_{00}+P_{01}} \tag{A3}$$

$$= \frac{P_{00}}{P_{00}+P_{01}} + \frac{G_Y(y)\times P_{01}}{P_{00}+P_{01}} \tag{A4}$$

$$= a + G_Y(y)(1-a). \tag{A5}$$

b. Wet forecasts

$$F_{Y|X}(y|x)|_{X>0} = P(Y\le y|X=x, X>0) \tag{A6}$$

$$= \frac{P(Y\le y, X=x, X>0, Y=0) + P(Y\le y, X=x, X>0, Y>0)}{P(X=x, X>0, Y=0) + P(X=x, X>0, Y>0)} \tag{A7}$$

$$= \frac{P(X=x, X>0, Y=0) + P(Y\le y|X=x, X>0, Y>0)P(X=x, X>0, Y>0)}{P(X=x|X>0, Y=0)P(X>0, Y=0) + P(X=x|X>0, Y>0)P(X>0, Y>0)} \tag{A8}$$

$$= \frac{g_X(x)P_{10} + D_{Y|X}d_X(x)P_{11}}{g_X(x)P_{10} + d_X(x)P_{11}} \tag{A9}$$

$$= c(x) + D_{Y|X=x}[1-c(x)]. \tag{A10}$$

The conditional distribution $D_{Y|X(y|x)} = P(Y\le y|X=x, X>0, Y>0)$ is estimated via meta-Gaussian distribution theorem (Kelly and Krzysztofowicz 1997) by applying a normal quantile transformation (NQT) to continuous marginal variates $[X|X>0, Y>0]$and $[Y|X>0, Y>0]$ with strictly increasing cumulative distribution functions $D_X(x)$ and $D_Y(y)$, respectively, as follows:

$$D_{Y|X(y|x)} = Q\left[\frac{Q^{-1}[D_Y(y)] - \rho Q^{-1}[D_X(x)]}{\sqrt{1-\rho^2}}\right], \tag{A11}$$

where Q and Q^{-1} denote CDF and quantile functions of standard normal distribution, respectively, and ρ is the sample

Pearson correlation coefficient between standard normal variates $U = Q^{-1}[D_X(x)]$ and $V = Q^{-1}[D_Y(y)]$.

REFERENCES

Akaike, H., 1998: Information theory and an extension of the maximum likelihood principle. *Selected Papers of Hirotugu Akaike*, E. Parzen, K. Tanabe, and G. Kitagawa, Eds., Springer, 199–213, https://doi.org/10.1007/978-1-4612-1694-0.

Baran, S., and S. Lerch, 2016: Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, **27**, 116–130, https://doi.org/10.1002/env.2380.

——, and D. Nemoda, 2016: Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, **27**, 280–292, https://doi.org/10.1002/env.2391.

——, and S. Lerch, 2018: Combining predictive distributions for statistical post-processing of ensemble forecasts. *Int. J. Forecasting*, **34**, 477–496, https://doi.org/10.1016/j.ijforecast.2018.01.005.

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.

Bentzien, S., and P. Friederichs, 2012: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Wea. Forecasting*, **27**, 988–1002, https://doi.org/10.1175/WAF-D-11-00101.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, https://doi.org/10.1175/WAF993.1.

Brown, J. D., 2015: An evaluation of the minimum requirements for meteorological reforecasts from the Global Ensemble Forecast System (GEFS) of the U.S. National Weather Service (NWS) in support of the calibration and validation of the NWS Hydrologic Ensemble Forecast Service (HEFS). Tech. Rep., 120 pp., http://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/publications_presentations/HSL_LYNT_DG133W-13-CQ-0042_SubK_2013_1003_Task_3_Deliverable_04_report_FINAL.pdf.

——, L. Wu, M. He, S. Regonda, H. Lee, and D. J. Seo, 2014: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. *J. Hydrol.*, **519**, 2869–2889, https://doi.org/10.1016/j.jhydrol.2014.05.028.

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2006: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.*, **111**, D24307, https://doi.org/10.1029/2005JD006917.

Demargne, J., and Coauthors, 2014: The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1.

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263, https://doi.org/10.1080/07350015.1995.10524599.

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.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, 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.

——, ——, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, 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.

——, M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, https://doi.org/10.1175/MWR-D-15-0004.1.

——, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. National Blend of Models for statistical post-processing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, **145**, 3441–3463, https://doi.org/10.1175/MWR-D-16-0331.1.

He, M., M. Russo, and M. Anderson, 2016a: Predictability of seasonal streamflow in a changing climate in the Sierra Nevada. *Climate*, **4**, 57, https://doi.org/10.3390/cli4040057.

——, and Coauthors, 2016b: Verification of ensemble water supply forecasts for Sierra Nevada watersheds. *Hydrology*, **3**, 35, https://doi.org/10.3390/hydrology3040035.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden, 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205, https://doi.org/10.1002/2014GL062472.

Herr, H. D., and R. Krzysztofowicz, 2005: Generic probability distribution of rainfall in space: The bivariate model. *J. Hydrol.*, **306**, 234–263, https://doi.org/10.1016/j.jhydrol.2004.09.011.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. John Wiley & Sons, 292 pp., https://doi.org/10.1002/9781119960003.

Kelly, K. S., and R. Krzysztofowicz, 1997: A bivariate meta-Gaussian density for use in hydrology. *Stochastic Hydrol. Hydraul.*, **11**, 17–31, https://doi.org/10.1007/BF02428423.

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.

——, ——, A. Ye, and C. Miao, 2019: An improved meta-Gaussian distribution model for post-processing of precipitation forecasts by censored maximum likelihood estimation. *J. Hydrol.*, **574**, 801–810, https://doi.org/10.1016/j.jhydrol.2019.04.073.

——, Q. J. Wang, and Q. Duan, 2020: A variable-correlation model to characterize asymmetric dependence for postprocessing short-term precipitation forecasts. *Mon. Wea. Rev.*, **148**, 241–257, https://doi.org/10.1175/MWR-D-19-0258.1.

Mascaro, G., E. R. Vivoni, and R. Deidda, 2010: Implications of ensemble quantitative precipitation forecast errors on distributed streamflow forecasting. *J. Hydrometeor.*, **11**, 69–86, https://doi.org/10.1175/2009JHM1144.1.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1096, https://doi.org/10.1287/mnsc.22.10.1087.

Mazrooei, A., and A. Sankarasubramanian, 2017: Utilizing probabilistic downscaling methods to develop streamflow forecasts from climate forecasts. *J. Hydrometeor.*, **18**, 2959–2972, https://doi.org/10.1175/JHM-D-17-0021.1.

Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014a: Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Wea. Rev.*, **142**, 3003–3014, https://doi.org/10.1175/MWR-D-13-00355.1.

——, ——, A. Zeileis, and D. S. Wilks, 2014b: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, https://doi.org/10.1175/MWR-D-13-00271.1.

——, ——, and ——, 2016: Heteroscedastic censored and truncated regression with CRCH. *R J.*, **8**, 173–181, https://doi.org/10.32614/rj-2016-012.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Nelder, J., and R. Wedderburn, 1972: Generalized linear models. *J. Roy. Stat. Soc.*, **135A**, 370–384, https://doi.org/10.2307/2344614.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Rigby, R. A., and D. M. Stasinopoulos, 2005: Generalized additive models for location, scale and shape. *J. Roy. Stat. Soc.*, **54**, 507–554, https://doi.org/10.1111/j.1467-9876.2005.00510.x.

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.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, https://doi.org/10.1002/qj.2183.

——, and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1.

——, and D. Möller, 2015: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.*, **9**, 1328–1349, https://doi.org/10.1214/15-AOAS843.

——, and T. M. Hamill, 2018: Generating calibrated ensembles of physically realistic, high-resolution precipitation forecast fields based on GEFS model output. *J. Hydrometeor.*, **19**, 1651–1670, https://doi.org/10.1175/JHM-D-18-0067.1.

——, ——, B. Whitin, M. He, and A. Henkel, 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**, 3029–3046, https://doi.org/10.1002/2016WR020133.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464, https://doi.org/10.1214/aos/1176344136.

Seo, D.-J., and Coauthors, 2015: On improving ensemble forecasting of extreme precipitation using the NWS Meteorological Ensemble Forecast Processor (MEFP). *2015 Fall Meeting*, San Francisco, CA, Amer. Geophys. Union, Abstract H51P-08, https://agu.confex.com/agu/fm15/meetingapp.cgi/Paper/81958.

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.

Stasinopoulos, D. M., and R. A. Rigby, 2007: Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Software*, **23**, 1–46, https://doi.org/10.18637/jss.v023.i07.

Stauffer, R., N. Umlauf, J. W. Messner, G. J. Mayr, and A. Zeileis, 2017: Ensemble postprocessing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. *Mon. Wea. Rev.*, **145**, 955–969, https://doi.org/10.1175/MWR-D-16-0260.1.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

——, A. Fougères, P. Naveau, and O. Mestre, 2019: Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Wea. Forecasting*, **34**, 617–634, https://doi.org/10.1175/WAF-D-18-0149.1.

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.

——, 2018: Univariate ensemble postprocessing. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. Messner, Eds., Elsevier, 9–89, https://doi.org/10.1016/B978-0-12-812372-0.00003-0.

Wu, L., 2020: Tuning the bivariate meta-Gaussian distribution conditionally in quantifying precipitation prediction uncertainty. *Forecasting*, **2**, 1–19, https://doi.org/10.3390/forecast2010001.

——, D. J. Seo, J. Demargne, J. 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.

Yang, X., S. Sharma, R. Siddique, S. J. Greybush, and A. Mejia, 2017: Postprocessing of GEFS precipitation ensemble reforecasts over the U.S. Mid-Atlantic region. *Mon. Wea. Rev.*, **145**, 1641–1658, https://doi.org/10.1175/MWR-D-16-0251.1.

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.