Comparison of Probabilistic Quantitative Precipitation Forecasts from Two Postprocessing Mechanisms

YU ZHANG

National Weather Service, Silver Spring, Maryland

LIMIN WU

Lynker Technologies, and National Weather Service, Silver Spring, Maryland

MICHAEL SCHEUERER

Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, and NOAA Earth System Research Laboratory, Boulder, Colorado

JOHN SCHAAKE

Annapolis, Maryland

CEZAR KONGOLI

Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

(Manuscript received 21 December 2016, in final form 19 June 2017)

ABSTRACT

This article compares the skill of medium-range probabilistic quantitative precipitation forecasts (PQPFs) generated via two postprocessing mechanisms: 1) the mixed-type meta-Gaussian distribution (MMGD) model and 2) the censored shifted Gamma distribution (CSGD) model. MMGD derives the PQPF by conditioning on the mean of raw ensemble forecasts. CSGD, on the other hand, is a regression-based mechanism that estimates PQPF from a prescribed distribution by adjusting the climatological distribution according to the mean, spread, and probability of precipitation (POP) of raw ensemble forecasts. Each mechanism is applied to the reforecast of the Global Ensemble Forecast System (GEFS) to yield a postprocessed PQPF over lead times between 24 and 72 h. The outcome of an evaluation experiment over the mid-Atlantic region of the United States indicates that the CSGD approach broadly outperforms the MMGD in terms of both the ensemble mean and the reliability of distribution, although the performance gap tends to be narrow, and at times mixed, at higher precipitation thresholds (>5 mm). Analysis of a rare storm event demonstrates the superior reliability and sharpness of the CSGD PQPF and underscores the issue of overforecasting by the MMGD PQPF. This work suggests that the CSGD's incorporation of ensemble spread and POP does help enhance its skill, particularly for light forecast amounts, but CSGD's model structure and its use of optimization in parameter estimation likely play a more determining role in its outperformance.

1. Introduction

Ensemble weather and hydrologic forecasts have been playing an increasingly critical role in public warning, disaster preparedness, and resource management (Georgakakos et al. 1998; Ajami et al. 2008; Pagano et al. 2014). At present, ensemble weather forecasts supplied by weather agencies worldwide are created using a combination of techniques, including perturbation of initial conditions, parallel runs of models with different model cores or variants of physical parameterizations, incorporation of stochastic tendency terms, and stochastic physics (Houtekamer et al. 1996; Buizza and Palmer 1998; Du et al. 2003). Notwithstanding these advances, operational ensembles from weather models can still be subject to severe

Corresponding author: Yu Zhang, yu.zhang@noaa.gov

DOI: 10.1175/JHM-D-16-0293.1

bias and large errors. In addition, ensemble members can often be clustered in a narrow window. This phenomenon, commonly referred to as underdispersion, leads to severe underrepresentation of the range of possible outcomes and associated forecast uncertainties.

The errors and underdispersion of ensemble weather forecasts are particularly concerning to hydrologic forecasts, because errors and underdispersion in forecast variables such as precipitation would translate into bias and underdispersion in streamflow forecasts, particularly for large storm events. Systematic ways of addressing these issues have thus been an important area of research (Kelly and Krzysztofowicz 1997; Hamill and Whitaker 2006; Krzysztofowicz and Evans 2008).

A number of statistical postprocessing mechanisms have emerged over the last two decades to address the aforementioned issues in the ensemble forecast based on dynamic models. These techniques share the commonality of seeking to establish empirical relationships between forecast variables and observations, and they use such relationships to predict observations from dynamic model-based forecast variables. One of the earliest of such techniques is Model Output Statistics (MOS; Glahn and Lowry 1972; Carter et al. 1989), an approach developed by the U.S. National Weather Service (NWS) that relies on establishing regression equations to relate the forecast (predictor) and observation (predictand). More recent developments include Bayesian modeling averaging (Raftery et al. 2005), logistic regression and its variants (Hamill and Whitaker 2006), the analog approach (Hamill and Whitaker 2006), the censored shifted Gamma distribution (CSGD; Scheuerer and Hamill 2015), Bayesian methods such as the Bayesian processor of ensemble (BPE; Krzysztofowicz and Evans 2008), and the mixed-type meta-Gaussian distribution (MMGD; Kelly and Krzysztofowicz 1997; Herr and Krzysztofowicz 2005; Wu et al. 2011).

Among the forecast variables, precipitation is perhaps the most challenging to postprocess because of its space–time intermittency and usually large forecast errors. Of the aforementioned postprocessing techniques, two have been noted for producing skillful probabilistic quantitative precipitation forecasts (PQPFs) for heavy precipitation events, namely, the MMGD (Kelly and Krzysztofowicz 1997) model and the CSGD approach (Scheuerer and Hamill 2015). The MMGD mechanism has been an integral part of the National Weather Service's Hydrologic Ensemble Forecast System (HEFS; Demargne et al. 2014). It employs a copula to establish the joint distribution of the ensemble forecast mean and precipitation amounts estimated from observations; it then calculates the conditional distribution of precipitation given a known forecast ensemble mean. CSGD, on the other hand, traces its roots to the MOS approach: it relies on regression, but is distinct from traditional MOS in that the predictands are parameters of a prescribed distribution of precipitation amounts rather than the precipitation amounts. Relative to the analog approach (Hamill and Whitaker 2006), the CSGD mechanism uses the dynamic model forecast and historical observations more efficiently, and its postprocessed forecast was shown to be less susceptible to suppression for heavy precipitation events (Scheuerer and Hamill 2015).

In light of the promising performance of CSGD as demonstrated in Scheuerer and Hamill (2015), it is of interest to investigate the relative performance of CSGD versus MMGD, in particular in processing forecasts that indicate heavy precipitation. In this study, we perform a detailed comparison of PQPFs derived using CSGD and MMGD to determine the relative performance of these two mechanisms and the specific features of each mechanism that give rise to the difference, if any, in their performance. The key science questions include 1) how the two mechanisms perform differently in terms of reliability, resolution, and sharpness; 2) how their relative performance varies depending on precipitation intensity, season, and terrain; and 3) the contribution of ensemble spread and probability of precipitation (POP) to the differential performance. In addition, it must be noted that, though the postprocessed ensembles generated via the Meteorological Ensemble Forecast Processor (MEFP) have been evaluated in a number of studies (e.g., Brown et al. 2014), PQPF from MMGD has yet to be rigorously scrutinized. This study complements the extant literature on MEFP by focusing on the accuracy of MMGD-based PQPF. In addition, by performing the evaluation on a high-resolution grid rather than on a basin-average basis, the study helps inform the development of the next generation, high-resolution forcings engine for the National Water Model (NWM; http://water.noaa.gov/documents/OWP-interface-PDD.pdf).

The remainder of the article is structured as follows. Section 2 reviews the MMGD and CSGD frameworks. Section 3 describes the design of postprocessing and validation experiments, the forecast and observation data involved, and evaluation metrics. Section 4 presents the results of intercomparisons and sensitivity analysis. Section 5 summarizes the results and concludes the study.

2. MMGD and CSGD mechanisms

a. *MMGD*

First introduced by Kelly and Krzysztofowicz (1997) to the field of forecast postprocessing, MMGD seeks to establish the joint distribution of the forecast and observation using the meta-Gaussian model and then estimate

the conditional distribution of the observation based on the forecast (i.e., the PQPF). It is currently implemented in the NWS's HEFS (Demargne et al. 2014) as a part of the MEFP (Wu et al. 2011). MMGD ingests the raw ensemble forecast and supplies the postprocessed PQPF, which is subsequently sampled to create postprocessed ensemble members. A brief description of MMGD is provided below, while interested readers can find additional details in Wu et al. (2011).

In MMGD, the cumulative distribution function (CDF) of the joint probability of precipitation forecast X and observation Y is given by

$$F(X,Y) = P(X \le x, Y \le y). \tag{1}$$

Because of the intermittent nature of precipitation, Herr and Krzysztofowicz (2005) decompose the joint distribution by considering four possible conditions, namely, 1) both the forecast and observation are dry, 2) the forecast is positive whereas the observation is dry, 3) the forecast is dry whereas the observation is positive, and 4) both the forecast and observation are positive. Representing the probability associated with these four scenarios as P_{00}, P_{10}, P_{01}, and P_{11}, respectively, $F(X, Y)$ can be expanded using the conditional law:

$$\begin{aligned} F(X,Y) = {} & P_{00} + P_{10}G_X(x) + P_{01}G_Y(y) \\ & + P_{11}D(x,y), \end{aligned} \tag{2}$$

where $G_X(x) = P(X \le x | X > 0,\ Y = 0)$, $G_Y(y) = P(Y \le y | X = 0,\ Y > 0)$, and $D(x, y) = P(Y \le y,\ Y \le y | X > 0,\ Y > 0)$.

Parameters, including P_{xx}, are derived empirically from the reforecast and observations; $G_X(x)$ and $G_Y(y)$ can take various forms. In this study they are assumed to follow the Pearson type III (P3) distribution:

$$\begin{aligned} F_{k,\theta,\delta}(y) &= F_k\left(\frac{y-\delta}{\theta}\right) \\ &= \frac{1}{\Gamma(k)}\int_{\delta}^{y}\frac{(u-\delta)^{k-1}e^{-(u-\delta)/\theta}}{\theta^k}du, \end{aligned} \tag{3}$$

where k, θ, and δ are the shape, scale, and location parameters, respectively. The location parameter δ accounts for the intermittency of precipitation. When δ is 0, P3 reverts to the Gamma distribution, for which the probability of dry conditions $F_k(0)$ is always zero. This is physically unrealistic, as it implies that precipitation occurs on a constant basis. A positive probability of dry conditions $F_k(-\delta/\theta)$ can be attained when parameter δ is set to be negative.

Parameter $D(x, y)$ is modeled using the bivariate meta-Gaussian distribution. Let Z and W be standard normal variables that are obtained by applying a normal quantile transform to X and Y, that is, $Z = Q^{-1}[F_X(X)]$ and $W = Q^{-1}[F_Y(Y)]$, respectively. Assuming that the joint distribution of Z and W is bivariate Gaussian, $D(x, y)$ can be written as

$$D(x,y;\rho) = B\{Q^{-1}[F_X(x)], Q^{-1}[F_Y(y)]; \rho\}, \tag{4}$$

where ρ is the correlation between Z and W. The density function of B takes the following form:

$$f_B(z,w;\rho) = \frac{1}{2\pi(1-\rho^2)^{1/2}}\exp\left[-\frac{w^2+z^2-2\rho zw}{2(1-\rho^2)}\right]. \tag{5}$$

One of the primary goals of the postprocessing is to obtain the CDF of observations conditional on the forecast:

$$F_{Y|X}(y|x) = P(Y \le y | X = x). \tag{6}$$

Wu et al. (2011) shows that $F_{Y|X}$ can be expressed as a function of the conditional CDF $D_{Y|X}(y|x)$; $g_X(x)$, the probability density function (PDF) of $G_X(x)$; marginal CDF $D(x, \infty)$, where $D(x, \infty) = P(X \le x,\ Y \le \infty | X > 0,\ Y > 0)$; and its PDF, $d_X(x)$:

$$F_{Y|X}(y|x) = c(x) + [1 - c(x)]D_{Y|X}(y|x), \tag{7}$$

where

$$c(x) = \frac{P_{10}g_X(x)}{P_{10}g_X(x) + P_{11}d_X(x)}. \tag{8}$$

Establishing the MMGD requires the estimation of 17 parameters, including the correlation, intermittency, and the parameters for the marginals of the forecast and observations. As indicated earlier, these parameters are estimated empirically. The parameters for the marginals are commonly estimated using the L-moment (Hosking1990), whereas the correlation is estimated using Pearson's correlation. The coefficients P_{xx} (P_{00}, P_{01}, P_{10}, and P_{11}) are estimated from reforecast and observation data using empirical frequency of each of the four scenarios.

b. CSGD

The CSGD method was conceived by Scheuerer and Hamill (2015). This approach relies on the three-parameter censored Gamma distribution to model both observed and forecast precipitation amounts. The CSGD is practically identical to the P3 distribution used in modeling the marginals in MMGD [Eq. (3)]. Although the formulation of P3 permits a negative value of y [Eq. (3)], the precipitation forecast or observation is guaranteed to be nonnegative. Therefore, a negative value in the shift parameter δ would help assign a positive probability associated with zero precipitation.

TABLE 1. Comparison of MEFP and CSGD.

	Parameter estimation	Use of ensemble mean	Use of ensemble spread	Expanded domain
MEFP	Goodness of fit	Yes	No	No
CSGD-P	Regression	Yes	Yes	Yes
CSGD-S	Regression	Yes	No	Yes

A distinct feature of CSGD is that both the climatological (unconditional) distribution of the precipitation observation, and the conditional distribution of the precipitation forecast, follow the P3 distribution. By contrast, MMGD produces a mixed-type distribution.

PQPF generation using CSGD is accomplished in three phases. In phase I, archival observed precipitation is used to establish the parameters for the climatological or unconditional CSGD. Next, in phase II, regression relationships are established using reforecast and observations between three CSGD parameters and the ensemble mean and spread. In phase III, the regressions relationship derived earlier would apply to the real-time ensemble forecast to obtain the adjusted CSGD parameters.

Following Scheuerer and Hamill (2015), let us denote $\log(1+x)$ as $\log1p(x)$ and $\exp(x)-1$ as $\text{expm1}(x)$. The regression equations for obtaining the mean μ_s and standard deviation of predictive CSGD σ_s from the climatological CSGD parameters $\mu_{\text{cl},s}$ and $\sigma_{\text{cl},s}$ are

$$\mu_s = \frac{\mu_{\text{cl},s}}{\alpha_{1,s}}\log1p\left[\text{expm1}(\alpha_{1,s}) \times\left(\alpha_{2,s}+\alpha_{3,s}\text{POP}_{f,s}+\alpha_{4,s}\frac{\overline{f}_s}{\overline{f}_{\text{cl},s}}\right)\right] \quad \text{and} \tag{9}$$

$$\sigma_s = \sigma_{6,s}\sigma_{\text{cl},s}\left(\frac{\mu_s}{\mu_{\text{cl},s}}\right)^{\alpha_{7,s}}+\alpha_{8,s}\text{MD}_{f,s}, \tag{10}$$

where $\alpha_{i,s}$, $i=1,\ldots,8$ are regression coefficients; $\text{POP}_{f,s}$ is the probability of precipitation; $\overline{f}_s$ and $f_{\text{cl},s}$ are the mean of the raw ensemble and the corresponding climatological mean, respectively; and $\text{MD}_{f,s}$ is a measure of ensemble spread. In this study, we follow Scheuerer and Hamill (2015) in computing ensemble mean and spread over a neighborhood with a radius of six Hydrologic Rainfall Analysis Project (HRAP) pixels (approximately 24 km). The ensemble spread MD is computed using a weighted mean absolute error, where the weights are determined by the distance of a particular pixel to the center pixel. The predicted σ_s and μ_s are then used to estimate the CSGD (or P3) parameters k and θ (note that $\mu=\sigma+k\theta$ and $\sigma^2=k\theta^2$ for P3).

One critical advantage of the CSGD approach is that a closed form of continuous ranked probability score (CRPS) is available (Scheuerer and Hamill 2015). This allows for an efficient way of estimating a set of optimal coefficients through minimization of CRPS.

Major differences between and MEFP and CSGD are summarized in Table 1. Note that one of the differences is that MEFP relies on the ensemble mean only, whereas CSGD utilizes the ensemble mean, spread, and POP. To assess the impact of incorporating the latter two predictors, we also implemented a simplified version of the CSGD, wherein only the ensemble mean is employed as the predictor in adjusting the μ_s and σ_s:

$$\mu_s = \mu_{\text{cl},s}\left(\alpha_{2,s}+\alpha_{4,s}\frac{\overline{f}_s}{\overline{f}_{\text{cl},s}}\right) \quad \text{and} \tag{11}$$

$$\sigma_s = \sigma_{6,s}\sigma_{\text{cl},s}\left(\frac{\mu_s}{\mu_{\text{cl},s}}\right)^{1/2}. \tag{12}$$

It is also worth noting that CSGD preprocesses the forecast variables prior to conducting the regression. The preprocessing entails constructing a superensemble using pairs within a neighborhood, deriving smoothed forecast variables, and applying a quantile transform to the raw ensemble traces. These measures are introduced to MMGD to make the comparison "fair." The impacts of these measures on MMGD PQPF will also be examined.

3. Experimental design

Our study region is the service area of the NWS Middle Atlantic River Forecast Center (MARFC), located over the east coast of the United States. As shown in the top panel of Fig. 1, this area encompasses Pennsylvania; Delaware; Maryland; New Jersey; and parts of West Virginia, Virginia, and New York. The spatial distribution of precipitation is impacted by the land–ocean boundary and the presence of the Appalachian Mountains (top panel of Fig. 1). This area is chosen for this study for the following reasons. First, multisensor precipitation records over this region are of high quality, which provides a basis for robust calibration and validation of the two postprocessing schemes. Second, a number of heavy storms occurred in the 2010–14 period, which allows a meaningful intercomparison and assessment of the two models for heavy–extreme precipitation events. Third, precipitation-producing mechanisms in the study region are diverse: major players include landfalling

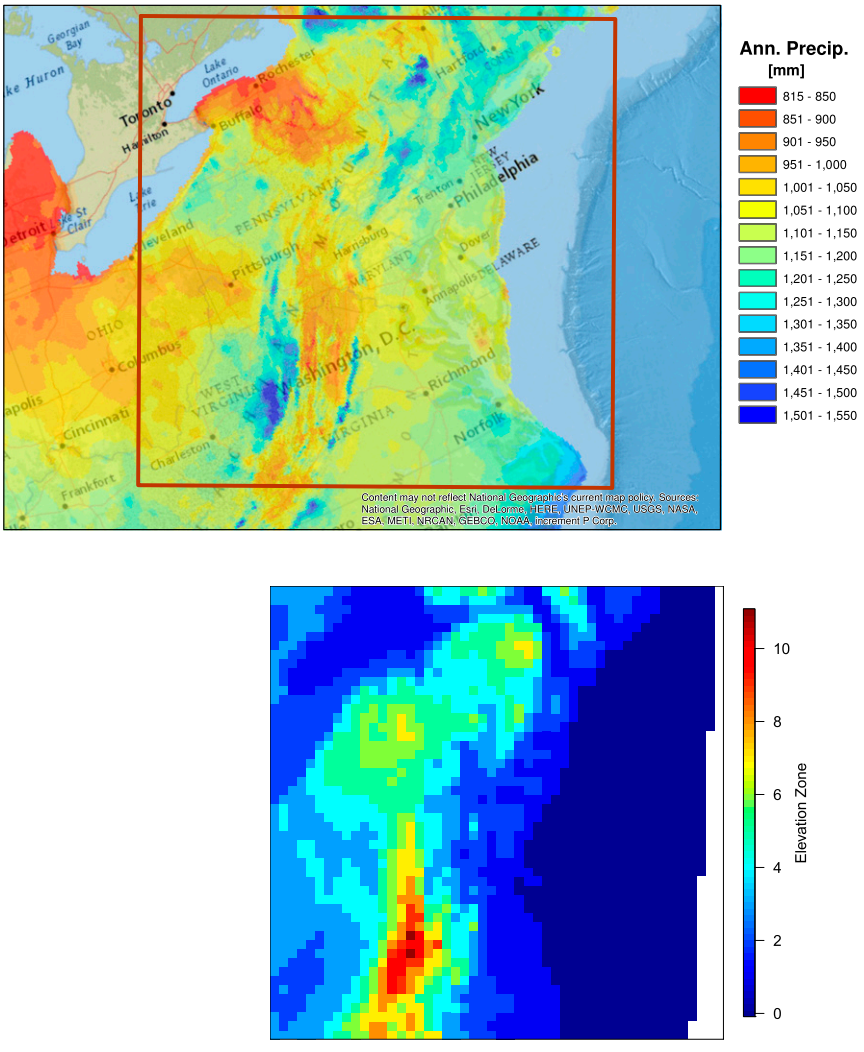

FIG. 1. (top) Mean annual precipitation accumulation for the mid-Atlantic study domain (source: PRISM). (bottom) Elevation zones, where zone i covers the elevation range of $[100i \text{ m}, 100(i + 1) \text{ m}]$.

tropical storms, extratropical cyclones, and, to a lesser extent, summertime convective storms and orographic systems. In particular, the central Appalachians region is known for heavy storms whose formation is influenced by orographic lifting (Barros and Kuligowski 1998; Smith et al. 2011). This diversity implies that the outcome of the experiment would be representative of the entire eastern United States. To examine the potential influence of terrain on the relative performance of the two mechanisms, we created a map of elevation zones on the grid mesh of the postprocessed forecast, with zone i encompassing the elevation range of $[100i \text{ m}, 100(i+1) \text{ m}]$. This map is shown in the bottom panel of Fig. 1.

For this study region, 6-h precipitation accumulations from reforecasts of the Global Ensemble Forecast System (GEFS; Wei et al. 2008; Hamill et al. 2011) and observations over 1997–2009 were retrieved and serve as the basis of parameter estimation, whereas corresponding data for the period of 2010–14 are used for validation. This training–validation strategy is chosen to mimic the real-time operation of the NWS hydrologic prediction service where parameters are derived over a long historical archive and would be in use without any update for an extended period of time. In this study, we focus our assessment at a relatively narrow range of lead times: 24–72 h given the following considerations. First, shorter-range (0–19 h) precipitation forecasts from convective-allowing numeric weather models are increasingly employed in the NWS in short-term hydrologic forecasts; accuracy of a GEFS forecast and its postprocessed versions over this range is less of a concern. Second, forecasters in the NWS typically issue flood warnings/watches within a lead time of 3 days, and

the skill of PQPFs is most relevant to river forecast operations within this window.

The GEFS reforecast was created and maintained at the Earth System Research Laboratory (Hamill et al. 2013) using version 9.0.1 of GEFS. Over the lead time windows used in this study, the reforecast data are on a quadratic Gaussian grid mesh that is approximately 40 km in resolution. The reforecasts were issued at a 3-h time step for lead times within 72 h every 24 h. The reforecast at each time step consists of 11 members, one being control and the rest perturbed. To facilitate parameterization, the reforecast precipitation for each pixel is accumulated on 6-h intervals ending at synoptic hours, and on a polar stereographic grid mesh with grid spacing of approximately 19 km to be consistent with that of the observation.

The precipitation observation used in this study is an adjusted version of the hourly Multisensor Quantitative Precipitation Estimates (MQPE) produced at the MARFC. The earlier MQPEs (prior to 2002) were produced using the Stage III algorithm (Zhang et al. 2011), and the later data were based on the Multisensor Precipitation Estimator (MPE; Seo et al. 2011; Zhang et al. 2011; Kitzmiller et al. 2013). The MQPEs are on a polar stereographic grid mesh [the so-called HRAP grid; Reed and Maidment 1999], with spatial resolution of approximately 4.7 km. The original MQPE underwent adjustment using a monthly gauge-based product as a reference to mitigate temporally varying biases, and the resultant product was then accumulated onto 6-h intervals and aggregated onto a coarser grid mesh (each grid cell consists of 4×4 HRAP pixels, or ~19 km in resolution) that is identical to that used for the processed GEFS reforecast dataset.

The parameters for the bivariate meta-Gaussian distribution are estimated using the data over 1997–2009, and forecast data for 2010–14 are used to derive the conditional distribution $F_{Y|X}(y|x)$ in Eq. (6). The training for CSGD entails deriving the parameters of the climatological CSGD (CSGD-C) and the regression coefficients [Eqs. (9) and (10) for the full CSGD (CSGD-P) and Eqs. (11) and (12) for the simplified version of CSGD (CSGD-S)], the latter of which are estimated using observation and reforecast for the training period. For the validation period, the regression coefficients derived for the earlier period are then used to adjust the climatological CSGD on the basis of the attributes of raw GEFS forecast. Our evaluation first focuses on the accuracy of ensemble means using bias and root-mean-square error (RMSE). Next, we assess the conditional distributions through Brier skill score (BSS), reliability, and ranked probability skill score (CRPSS). The definitions of BSS and CRPSS are provided in the appendix.

For the reference Brier score and CRPS, we chose to use the climatological CSGD derived over the 1997–2009 period. In addition to MMGD and CSGD-P, we compute and show the validation statistics for CSGD-S. The relative performance of MMGD and CSGD is further illustrated through the study of a rare storm event in the fall of 2010. Here, the sensitivity of MMGD results to quantile mapping and spatial smoothing of ensemble means is examined.

4. Results

Prior to the assessment of postprocessed PQPF, the seasonal accuracy of the raw GEFS ensemble mean over the 2010–14 period as a function of season is briefly characterized. Figures 2a–d show the percentage of 6-h intervals where observed or forecast precipitation exceeds four thresholds, namely, 0.25, 5, 25, and 50 mm (per 6 h; note the amounts are those for 6-h accumulations throughout the paper), over each month. Also shown is the percentage of intervals where 6-h precipitation accumulations from both the GEFS mean and observations are beyond each threshold (referred to as overlapping hours). It is evident that 1) seasonal variations of precipitation as depicted by GEFS closely mimic that of analysis; 2) GEFS has a wet bias (Fig. 2a); 3) the GEFS mean tends to underrepresent the hours with light–moderate to heavy rainfall, and this becomes more severe at higher thresholds (Figs. 2b–d); 4) light to moderate precipitation is more common during spring (Fig. 2a), whereas most heavy rain occurs mostly during late summer and fall (Fig. 2d), likely as a result of landfalling tropical cyclones; and 5) accuracy of the GEFS ensemble mean tends to be poor as judged by the percentage of overlapping hours, and particularly for heavier precipitation (Figs. 2a–d). The underrepresentation of heavy precipitation is unsurprising, as some of the heavy precipitation fell during convective events that cannot be captured at GEFS's resolution.

a. Evaluation of PQPF means

The means of 6-h precipitation accumulations from postprocessed PQPF via MMGD and two CSGD schemes are used to compute averaged bias and RMSE for each lead time and month, and the results are shown in Figs. 3a–d. Among the PQPFs, bias for the raw GEFS-based PQPFs is consistently close to neutral; two CSGD schemes both exhibit negative bias, and the bias is worse for CSGD-S (Fig. 3a). MMGD, by contrast, exhibits a sharp positive bias across lead times. A close examination of the MMGD-based conditional distribution reveals that, for a large number of instances where the forecast is zero or nearly zero, the conditional mean is positive. Further analysis suggests that spatial smoothing plays a

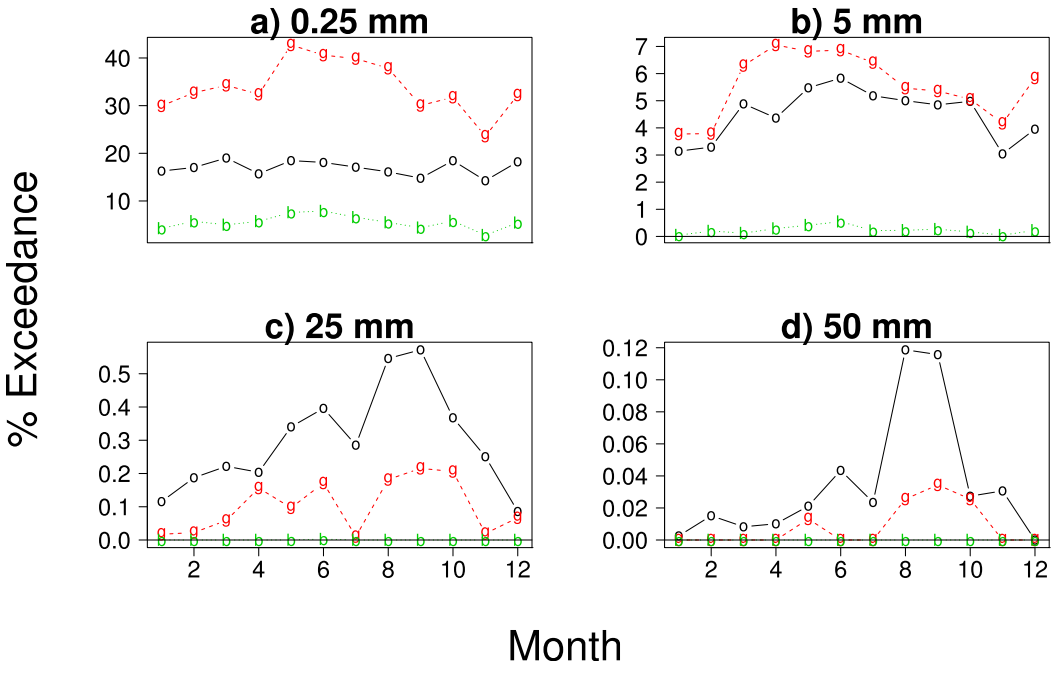

FIG. 2. Percentage of 6-h intervals where precipitation from observations, GEFS (at 24-h lead time), and both are beyond the threshold of (a) 0.25, (b) 5, (c) 25, and (d) 50 mm.

large role in inducing this bias, by inflating the POP and shifting the probability mass in $F_X(x)$ [Eq. (2)] toward the right. For RMSE (Fig. 3b), both CSGD schemes fare better than MMGD. MMGD does improve upon the result of GEFS, but the RMSE associated with its results is consistently higher than that for CSGD-P and CSGD-S. It is also clear that introducing ensemble spread and POP as predictors in the CSGD framework has limited impacts on the RMSE.

The seasonal variations in bias and RMSE are illustrated in Figs. 3c and 3d. The conspicuous positive bias in MMGD illustrated in Fig. 3a is consistent throughout the year (Fig. 3c), though it is the highest in late spring–early summer. The timing of this seasonal peak is consistent with that for GEFS. Both are related to overforecasts by GEFS for light precipitation events over this period (see Figs. 2a and 2b). As indicated earlier, the large positive bias in MMGD-based PQPF is primarily an outcome of spatial smoothing performed on the forecast ensemble. PQPFs based on the two CSGD schemes exhibit much lower bias, and the seasonality of bias differs drastically from that for GEFS and MMGD (Fig. 3c). Between the two sets of PQPFs, the PQPF based on CSGD-P exhibits nearly neutral bias, whereas the bias for CSGD-S is overall negative and most severely so in July. This seasonal cycle is almost exactly out of phase with that of the GEFS forecast, whose bias attains a peak (positive) over the summer. Closer examination reveals a similar negative bias in the CSGD climatology (Fig. 3c), which most likely arose from discrepancies in the seasonality of precipitation between the calibration (1997–2009) and validation (2010–14) periods. Between the two CSGD schemes, CSGD-S, which incorporates only the ensemble mean as a predictor, suffers from more severely negative bias, and the magnitude of bias closely resembles that of the CSGD climatology.

The contrasting seasonality in the bias of MMGD and CSGD-based PQPFs underscores the differing reliance on the GEFS ensemble mean and climatology by the three schemes. Evidently, the seasonality in bias in MMGD PQPF is heavily influenced by that in the GEFS mean. By contrast, though CSGD schemes also employ the MMGD ensemble mean as a predictor, the seasonality of bias is modulated primarily by the climatology. As shown in Fig. 3c, the positive bias in the GEFS mean is unable to offset the seasonal bias of climatology through either of the CSGD mechanisms. Introducing POP and ensemble spread as predictors in CSGD proves helpful in mitigating the bias, possibly because of the enhanced POP and spread over the summer season.

RMSEs for postprocessed PQPFs exhibit a clear seasonal cycle with a peak in the summer (Fig. 3b). This seasonal cycle closely resembles that in the raw GEFS ensemble. PQPFs from all three schemes outperform the raw GEFS ensemble in RMSE. Among the three schemes, the two CSGD schemes outperform the

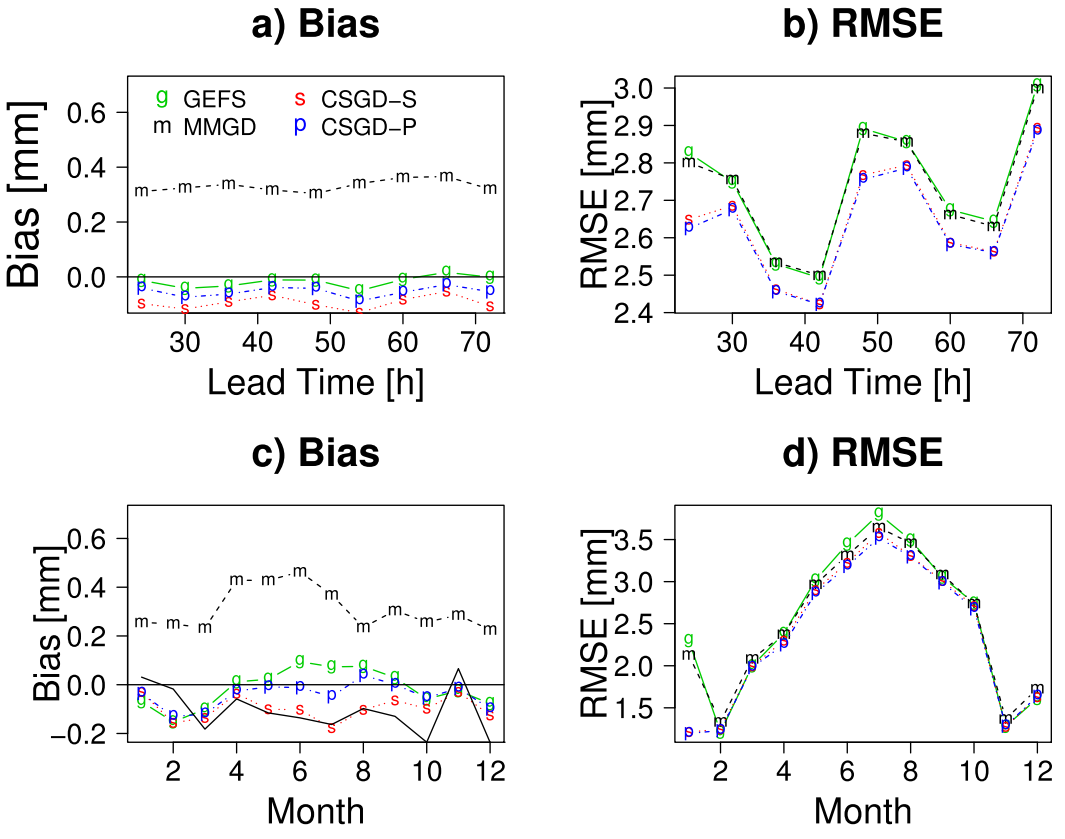

FIG. 3. Comparisons of averaged bias and RMSE of the raw GEFS ensemble mean (g; green), mean of PQPF based on MMGD (m; black), CSGD-S (s; red), and CSGD-P (p; blue): (a) bias as a function of lead time, (b) RMSE as a function of lead time, (c) monthly variation in bias at 24-h lead time, and (d) monthly variation in RMSE at 24-h lead time. The solid line in (c) illustrates the bias of the PQPF based on the CSGD climatology.

MMGD, and this outperformance is more pronounced over the warm season. Between the two CSGD schemes, CSGD-P features slightly lower RMSE. Evidently, despite being more severely biased, in terms of RMSE CSGD-S is comparable to that based on the full-fledged CSGD-P.

The reason that MMGD degrades the bias of GEFS, but at the same time improves the RMSE, is explored in Fig. 4, where the bias and the mean square error (MSE) are computed over six precipitation categories and are then weighted by the percentage of instances (samples) in each category. Applying this weighting allows for quantification of the contribution to bias and MSE from each precipitation category. It is clear in Fig. 4a that the positive bias induced by MMGD is confined to the lowest category, whereas at higher categories, the bias drops and sometimes becomes more negative. These instances in the lowest category play a major role in MMGD's exacerbation of the positive bias, but their contribution to RMSE is limited due to the small precipitation magnitude. As shown in Fig. 4b, MMGD produces a large reduction in weighted MSE over the middle categories (3–5). Also of note is that all three schemes degrade MSE at the highest two categories (Fig. 4b), although the degradation from CSGD schemes appears milder.

It appears that all three postprocessing schemes are able to suppress the overforecast, but at the expense of reducing the forecast amount for events that indeed occur. The relative efficacy of the three schemes is apparently mixed. For instances where GEFS indicates moderate and heavy precipitation, CSGD schemes tend to perform slightly better by yielding smaller degradation in MSE. For a light precipitation forecast, MMGD performs slightly better in MSE. These features are consistent with the early finding that applying MMGD results in a large increase in the positive bias (Fig. 3), and it is clear that this degradation in bias is the most severe over light precipitation cases. The two CSGD schemes yield relatively small changes to the metrics at the lower thresholds, though CSGD-P appears to perform better in POD, suggesting a limited

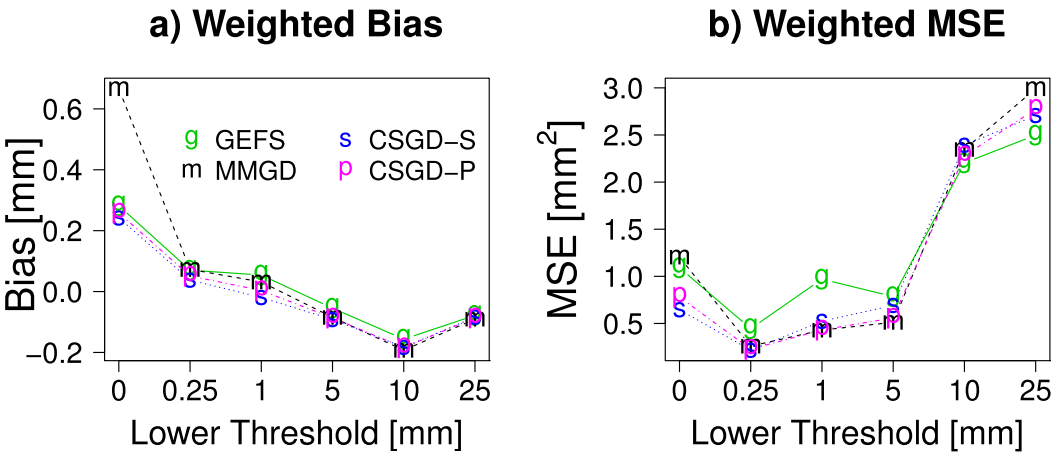

FIG. 4. (a) Weighted bias and (b) MSE computed at six categories of observed precipitation: 0–0.25, 0.25–1, 1–5, 5–10, 10–25, and 25–50 mm. A weighted variable is computed by multiplying the variable over a category with the percentage of instances within this category.

benefit of incorporating ensemble mean and POP in the CSGD framework.

b. BSS, reliability, and CRPSS

The accuracy of PQPFs is further assessed through comparison of the BSS (Figs. 5 and 6), decomposition of the Brier score (Fig. 7), reliability diagrams (Fig. 8), and CRPSS (Fig. 9).

Figure 5 shows the average BSS from GEFS, MMGD, and two CSGD schemes (CSGD-S and CSGD-P) over lead times at four selected thresholds: 0.25, 5, 10, and 25 mm. At the lowest threshold (0.25 mm; Fig. 5a), CSGD-P shows consistently the highest BSS across lead times, followed by CSGD-S and then MMGD. BSS for GEFS is the lowest and is frequently below zero, particularly at longer lead times. Evidently, GEFS PQPF

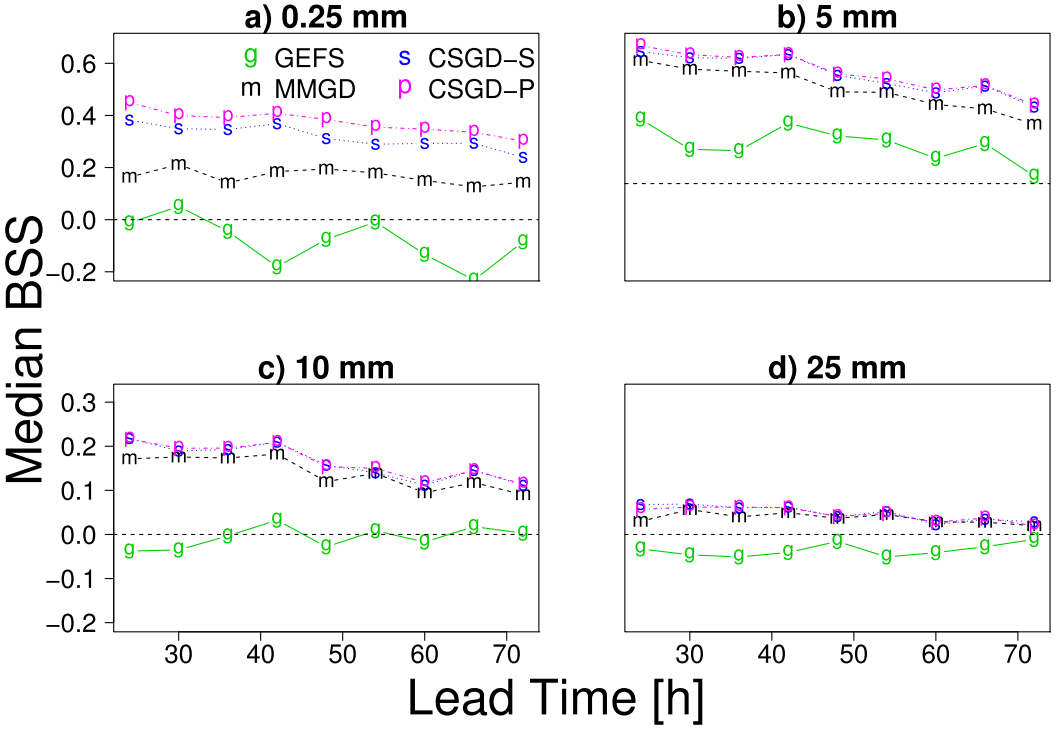

FIG. 5. Median BSS of the GEFS ensemble, MMGD, and CSGD PQPFs against lead times over the four precipitation thresholds (per 6 h): (a) 0.25, (b) 5, (c) 10, and (d) 25 mm. The dotted line at zero represents the climatology: points above (below) indicate better (worse) performance than climatology.

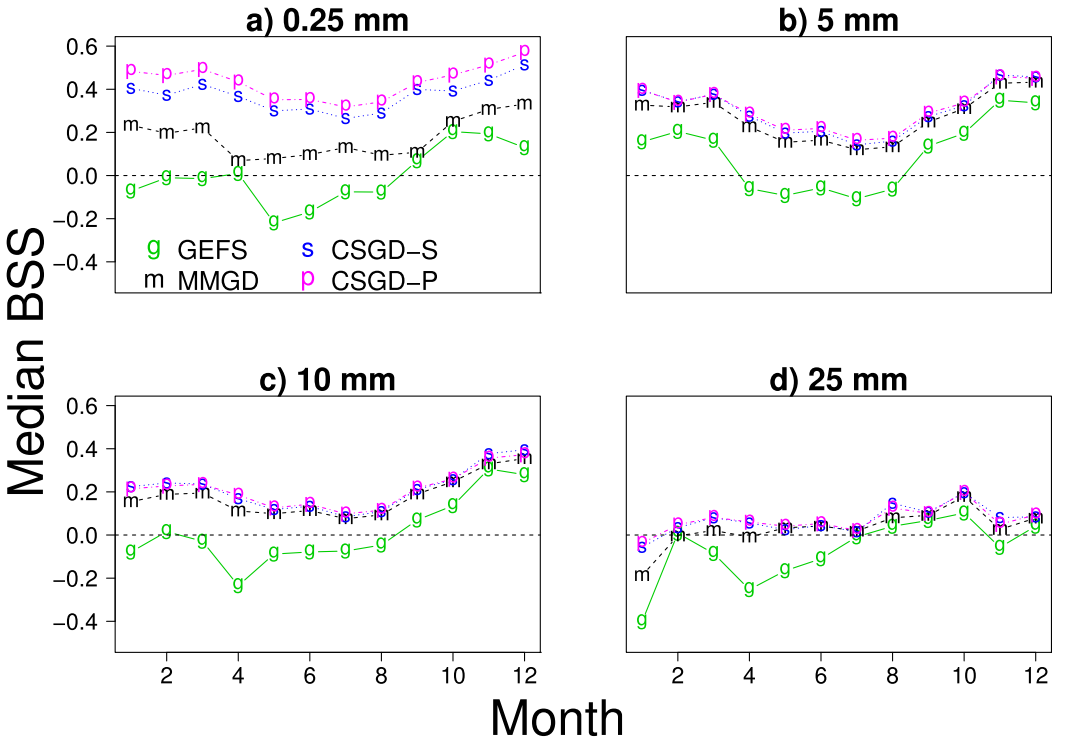

FIG. 6. As in Fig. 5, but for median BSS at 24-h lead time against months.

has marginally higher skill than climatology in differentiating wet versus dry conditions. Each of the four products shows a decline in BSS with lead time, reflecting the degradation of skill at long lead times. At the 5-mm threshold (Fig. 5b), CSGD schemes still outperform MMGD, but the difference in performance is much reduced. Also worth noting is that the BSS for the GEFS raw ensemble is much higher than that at the 0.25-mm threshold and stays positive across the lead times, suggesting that the raw GEFS forecast is more skillful than climatology at this threshold. At 10- and 25-mm thresholds (Figs. 5c and 5d, respectively), three features are evident: 1) BSS values for the two CSGD schemes are comparable while both are slightly higher than that for MMGD; 2) the difference among the three schemes becomes narrower at these thresholds; and 3) the BSS for GEFS is close to zero or negative, suggesting that the skill of GEFS is comparable or slightly inferior to climatology in identifying heavy precipitation.

The seasonal variations of the BSS at the four aforementioned thresholds are shown in Figs. 6a–d. At the 0.25-mm threshold (Fig. 6a), the two CSGD schemes broadly outperform MMGD for much of the year. The outperformance is most striking over the summer when the skill tends to be the lowest. Note that GEFS underperforms climatology between May and August, and this is likely the cause of its overall negative BSS shown in Fig. 5a. Each of the postprocessed PQPFs manages to outperform GEFS and climatology. Between the two CSGD schemes, CSGD-P clearly outperforms CSGD-S throughout the year, though only by a small margin. At higher thresholds (Figs. 6b–d), CSGD schemes still outperform MMGD over the summer, but the gap between CSGD and MMGD PQPFs becomes increasingly narrow, as does the gap between the postprocessed PQPFs and the raw ensemble.

To identify the key causes for the differential performance between CSGD and MMGD schemes in the BSS, we perform a decomposition of the Brier score per Murphy (1973), by stratifying the forecasts by the probability of falling into 1 of 10 evenly distributed bins between 0 and 1. The resulting aggregate reliability and resolution are shown in Fig. 7, whereas uncertainty is omitted as it is independent of the forecast and is therefore identical among the three PQPFs. For the aggregate reliability (Figs. 7a and 7b) in general, CSGD-P PQPF has a conspicuously higher reliability and resolution than CSGD-S and MMGD at the lower thresholds (0.25 and 1 mm; Figs. 7a and 7b). The difference in performance, again, narrows at higher thresholds. Another notable feature is that the reliability of CSGD-S PQPF is much lower than that of CSGD-P, and it is comparable to that of MMGD. As for the resolution, it is evident in Figs. 7c and 7d that both CSGD schemes are superior to MMGD

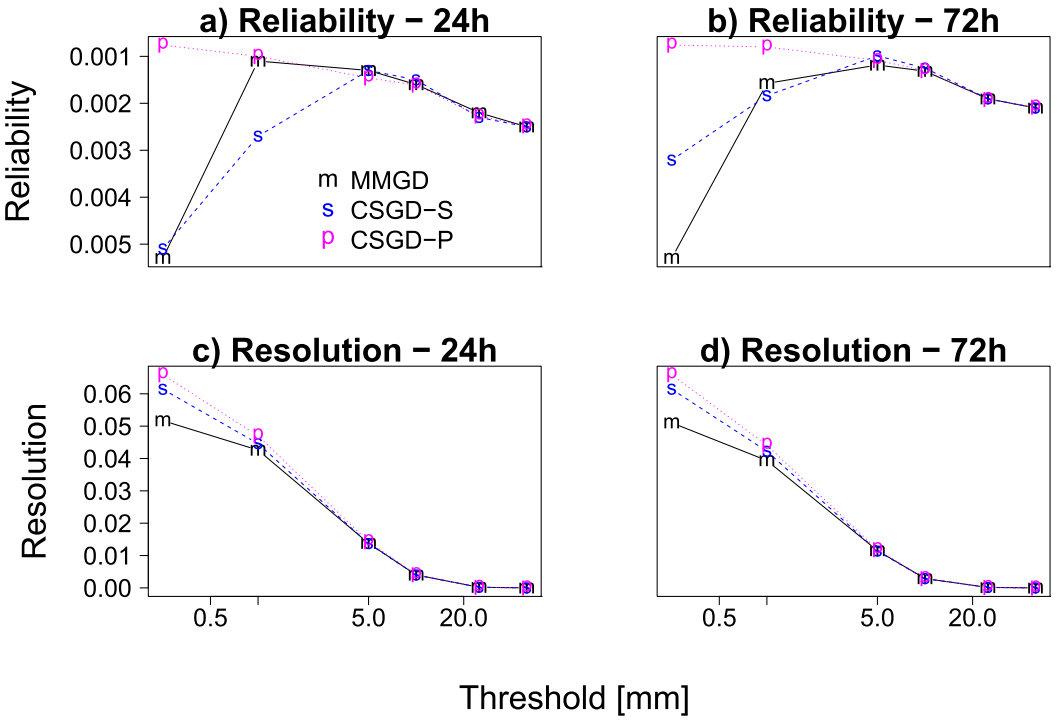

FIG. 7. Brier score decomposed: differences in aggregate reliability and resolution among MMGD, CSGD-S, and CSGD-P at a range of precipitation thresholds. Shown are (a) reliability at 24-h lead time, (b) reliability at 72-h lead time, (c) resolution at 24-h lead time, and (d) resolution at 72-h lead time. Note that the reliability measure is plotted upside-down in (a) and (b) as lower values indicate better performance.

at the two lowest thresholds, while the difference between the two CSGD schemes is relatively minor. This suggests that, at least for the light precipitation thresholds, the use of ensemble spread and POP as predictors mainly helps to improve aggregate reliability, whereas it is the structural difference of the two postprocessing mechanisms, along with the use of optimization in CSGD, that determine the difference in resolution. As resolution is of higher magnitude, it appears that the latter differences play a more critical role in shaping the superior performance of CSGD versus MMGD.

To further assess the reliability of the three PQPFs over a specific forecast probability range, we plot the reliability diagrams at the 24-h lead time (Fig. 8). At the lowest threshold (0.25 mm; Fig. 8a), it is clear that CSGD-P outperforms CSGD-S and MMGD from the low to middle probability range, where its observed frequency is the closest to the forecast probability. The observed frequency of CSGD-S tends to be higher than the forecast probability, pointing to an underforecast in these categories. The MMGD-based results are in the opposite direction, pointing to an underforecast. In the higher probability range, the differences among the three PQPFs tend to diminish. At the threshold of 5 mm (Fig. 8b), reliability diagrams of the three products are quite similar, though there is a sign of an underforecast by MMGD and CSGD-P in the middle range, and by CSGD-S in the higher range. This feature persists to some extent even at higher thresholds (Figs. 8c and 8d). A more striking feature in these two panels is the severe overforecast of MMGD at the higher probability categories [P = (0.8, 1.0) at 25 mm]. A close examination reveals that all the forecasts in this category per MMGD PQPF reside in the 6-h interval that ends at 0000 UTC 1 October 2010, over which the GEFS severely overforecasts precipitation amounts over a swath of the area. This will be further explored in the subsequent case study.

The differences among the three schemes are summarized by CRPSS. Figures 9a and 9b show the averaged CRPSS as a function of lead time and month, respectively. Again, it is clear that both CSGD schemes outperform MMGD across lead times. Between CSGD-P and CSGD-S, the former performs consistently better. The seasonal variation of CRPSS is a mirror image of that of RMSE (Fig. 3), with a trough over the summer where RMSE peaks. The performance difference between the CSGD schemes and MMGD is consistent across seasons. As pointed out by Hersbach (2000), among others, CRPS can be interpreted as an integral of the Brier score over threshold values. The difference in

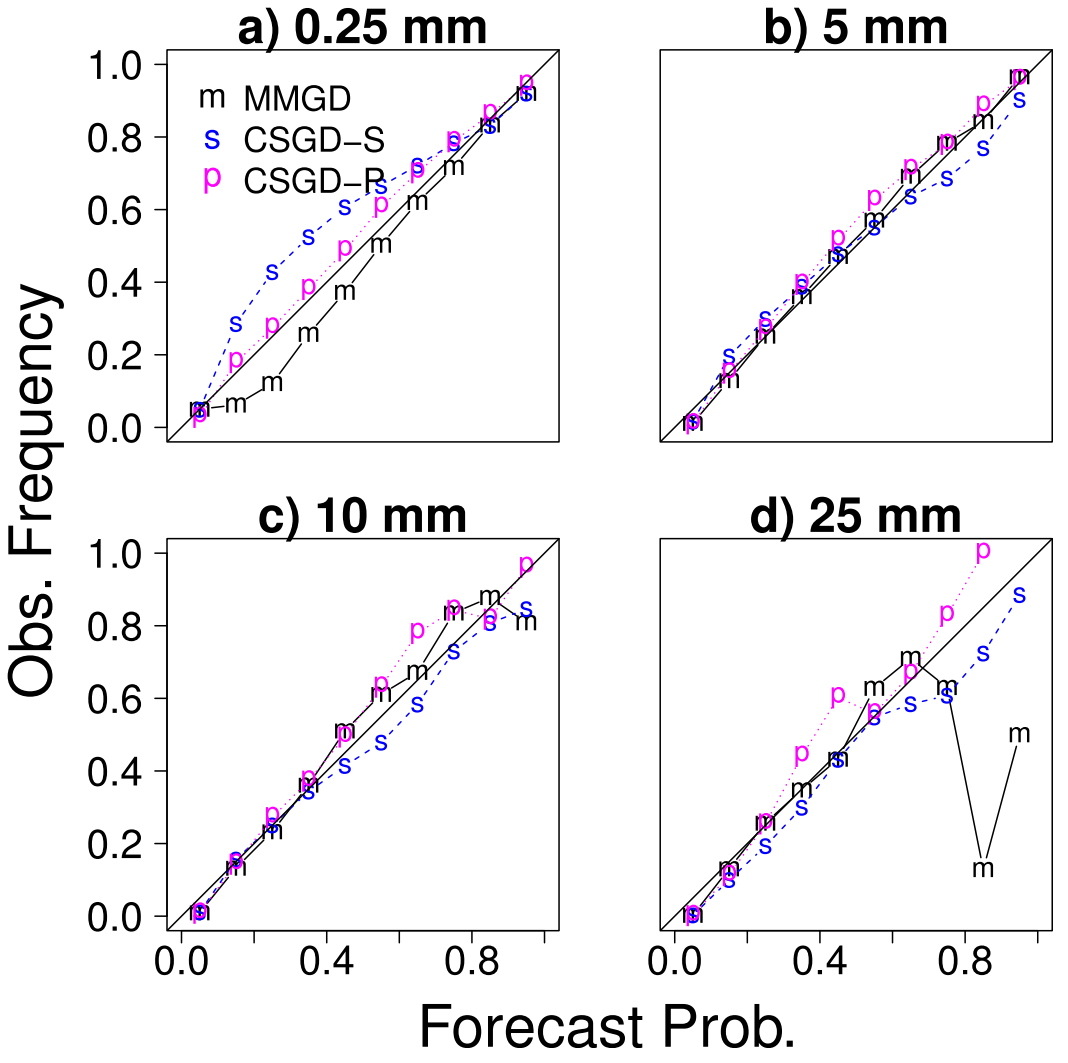

FIG. 8. Reliability diagram of the GEFS raw ensemble, MMGD, and CSGD PQPFs at 24-h lead time for a forecast threshold of (a) 0.25, (b) 5, (c) 10, and (d) 25 mm.

CRPSS could be attributed to the difference in performance in identifying light rainfall, as shown in the comparison of BSS (Figs. 5 and 6). Since the instances of light precipitation are more numerous, the superior performance of CSGD over these instances tends to dominate the CRPS. The depressed CRPSS over summer is consistent with the observation of Scheuerer and Hamill (2015), and it reflects the difficulty of GEFS in capturing summertime convective precipitation.

The performance of the three schemes at different elevation zones is shown in Fig. 10. Here, the CRPS and CRPSS from each scheme are plotted against the elevation zone number for winter (December–February) and summer (June–August). The observed precipitation amount in general increases with elevation for both seasons, but the trends in CRPS and CRPSS differ between the two seasons. Over the winter (Fig. 10a), CRPS for all three schemes exhibits a trough at zone 4 (400–500-m range), beyond which it tends to increase along with the precipitation amount. By contrast, for the summer, CRPS declines sharply with increasing elevation while the precipitation amount increases (Fig. 10b). For the winter, both CSGD schemes outperform MMGD across elevation zones (Figs. 10a and 10c), though the difference tends to be narrow between zones 4 and 9. Also of note is that the CRPSS for MMGD tends to follow the trend of precipitation (increasing with elevation beyond zone 4), whereas CRPSS for CSGD schemes is inversely related to precipitation. Over the summer (Fig. 10b), CRPSS for MMGD tends to increase with precipitation amount until zone 9, whereas CRPSS based on CSGD schemes declines with precipitation. Among the three schemes, CSGD-P remains the best performer whereas CSGD-S underperforms MMGD at higher elevations (zone 7 and beyond).

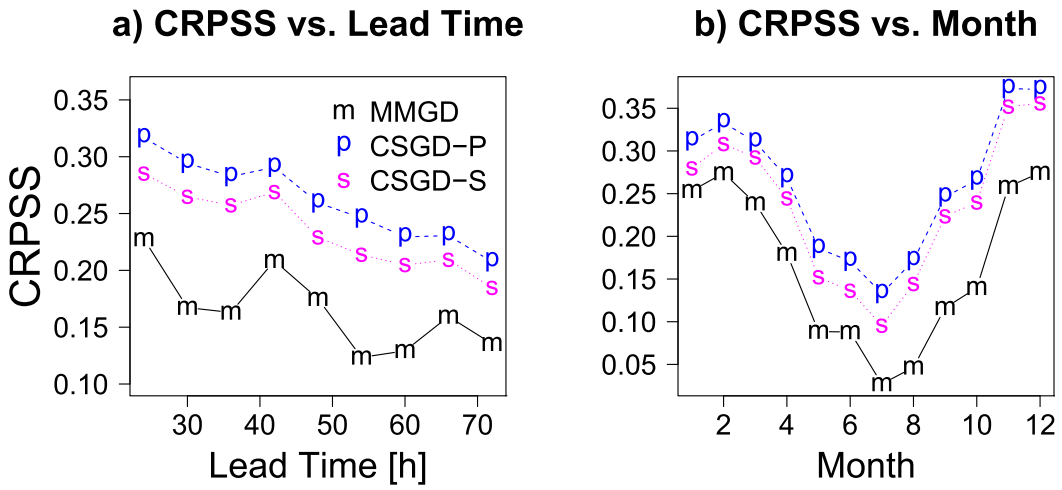

FIG. 9. (a) Monthly-averaged CRPSS of PQPFs (based on MMGD, CSGD-S, and CSGD-P) vs lead time and (b) CRPSS for PQPFs at 24-h lead time vs month.

Further analyses of spatial precipitation patterns (not shown) over the 2010–14 period point to a wintertime precipitation maximum over the west slope of the Appalachians. Note that the windward maximum is also shown in the PRISM-based, long-term climatology (Fig. 1), a clear indication of orographic influence. GEFS performs poorly over this wintertime maximum, and this poor performance helps explain the increasing CRPS (declining skill) for the winter. For the summer, though precipitation in general increases with elevation,

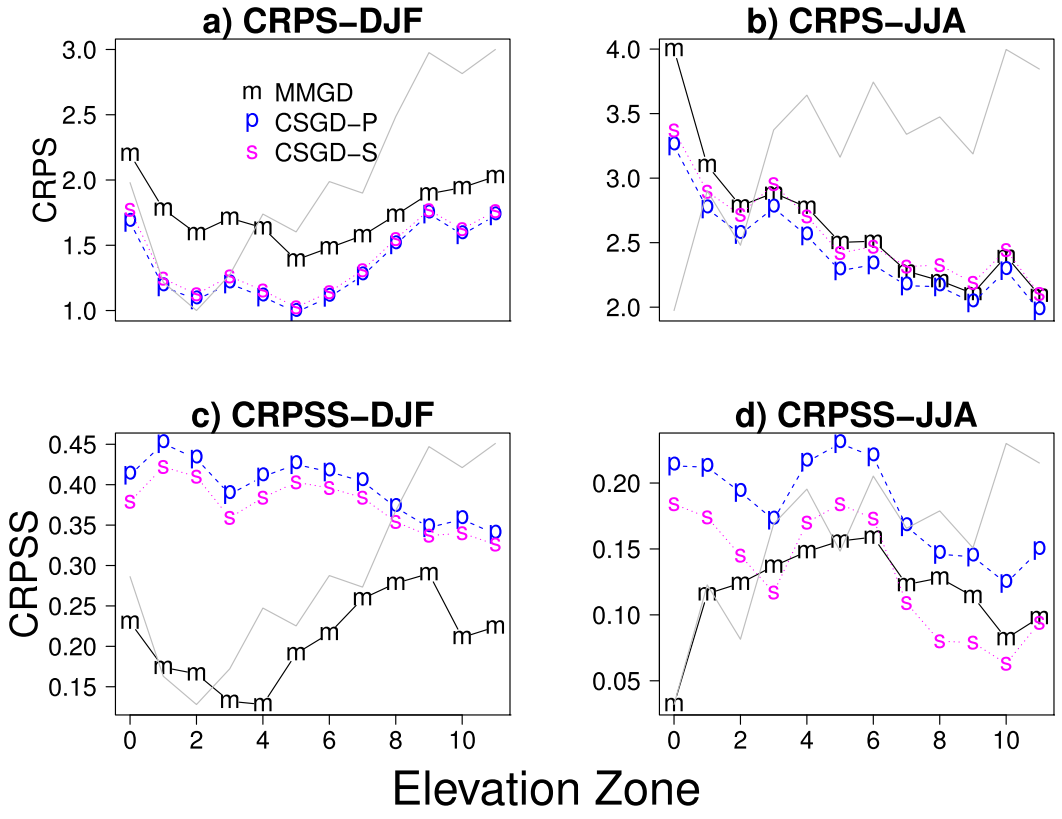

FIG. 10. CRPS and CRPSS for PQPF at 24-h lead time over different elevation zones for winter (DJF) and summer (JJA). Shown are (a) CRPS for winter, (b) CRPS for summer, (c) CRPSS for winter, and (d) CRPSS for summer. The thin gray line in each panel marks the averaged observed precipitation at each elevation zone.

the maxima are located off the coast where elevation is low. As GEFS exhibits low skill over the coastal maxima, this difference manifests as a monotonic decline in the CRPS of the three schemes. Therefore, topography exerts an influence on the skill of the model and that of the postprocessed ensemble, primarily by controlling the spatial distribution of heavy precipitation. However, it remains unclear why the CRPSS from both CSGD schemes declines with elevation, whereas the results for MMGD are mixed.

c. *Case study*

The skills of MMGD and CSGD in processing heavy precipitation are illustrated through the forecast and observed rainfall fields over a major historical storm event. This event took place between late September and early October of 2010 over the mid-Atlantic region, covering a wide swath of area extending from North Carolina to New York. The maximum storm total accumulation exceeds 400 mm, and widespread flooding conditions were reported.

Our attention is given to the 6-h period ending at 0000 UTC 1 October 2010. This is the period when an abrupt drop in reliability is seen in the MMGD PQPF at the 24-h lead time and the 25-mm threshold (Fig. 8d). Figure 11 shows the observed fields and corresponding forecast ensembles for PQPF means (Figs. 11a–d) and the spatial fields of CRPS for MMGD and CSGD PQPFs (Figs. 11e,f). It appears that the rainfall was clustered over the northeastern and southwestern portions of the domain, with accumulation amounts at the centers exceeding 80 mm. These centers are divided by a zone of relatively light accumulation over the central portion of the rainfall belt (Fig. 11a). As shown in Fig. 11b, the GEFS ensemble mean at the 24-h lead time accurately depicts the magnitude and geographic extent of the rainfall belt, but fails to resolve the fine structure in space: it depicts a single, ellipsoid-shaped storm cell rather than the disjoint structure as seen in the observation. The pixels where MMGD PQPF reports unrealistically high probability of exceeding 25 mm (marked by yellow crosses in Fig. 11b) mostly reside inside the GEFS storm center. Both MMGD and CSGD help reduce the forecast mean (Figs. 11c,d), while neither successfully resolves the spatial details. Between the two PQPFs, the one derived from CSGD shows lower CRPS values near the forecast storm center (see pixels marked by yellow crosses in Figs. 11e,f). This indicates that CSGD is more effective in correcting the severe overforecast of GEFS over the light precipitation zone shown in Fig. 11a. For the northwest and southeast corners of the domain where heavy precipitation indeed occurred, CRPS for CSGD PQPF remains slightly lower.

Figure 12 compares the probability distribution of postprocessed PQPFs based on MMGD and CSGD over the aforementioned group of pixels where false alarms are reported for MMGD (Fig. 11a), and over a selected pixel where the mean from MMGD and CSGD PQPFs is nearly identical and yet the CRPS contrasts sharply. Shown in Fig. 12a is the scatterplot of probability of the forecast precipitation amount exceeding 25 mm [$P(Z > 25\,\text{mm})$] over a 6-h interval based on MMGD and CSGD. The probability $P(Z > 25\,\text{mm})$ based on MMGD is greater than 0.8 over all these pixels, as they were chosen as such. By contrast, $P(Z > 25\,\text{mm})$ for CSGD PQPF is uniformly lower and is below 0.8 for each pixel. The presence of these false alarms leads to a depressed reliability of MMGD PQPF over the probability range of between 0.8 and 0.9. In CSGD PQPF, in none of these instances did the probability of exceedance go beyond 0.8; the few instances in this probability bracket all correspond to observed precipitation exceeding 25 mm, resulting in a high reliability of CSGD PQPF (Fig. 8d). Figure 12b further illustrates the difference in the MMGD- and CSGD-based distribution of precipitation amounts for one pixel where the PQPF mean is comparable (around 30 mm), yet the $P(Z > 25\,\text{mm})$ is quite different (0.46 for CSGD versus 0.87 for MMGD). As shown in Fig. 12b, the PDF of CSGD PQPF at this pixel features a sharper peak. This results in a higher CDF at the 25-mm threshold. As a consequence, the exceedance probability, which is a complement of CDF, is much lower for CSGD.

Figure 13 illustrates the impacts of conducting CSGD-style spatial smoothing and quantile adjustment of the raw ensemble forecast on the performance of MMGD over the 6-h interval on 1 October 2010. For this particular case, these postprocessing measures lead to a net reduction in RMSE (from 12 to 11 mm, not shown). The MMGD-based PQPF mean declines for a majority of pixels where rainfall amounts are large, though the results are mixed at lower amounts (Fig. 13a). In Fig. 13b, it is clear that CRPS declines after the adjustments over the center of the domain where GEFS overforecasts the precipitation amount (Fig. 11b), whereas it increases somewhat over the periphery of the forecast storm cell. Evidently, the introduction of CSGD-style postprocessing leads to a suppression of large forecast amounts through the MMGD approach. For this particular event, overforecasting by GEFS is a dominant feature, and this suppression alleviates the overforecasting and thereby helps improve the PQPF skill by reducing CRPS. It must be noted that this outcome is not necessarily applicable to other events or to the 5-yr validation period as a whole. Indeed, our analysis (not shown) indicates that applying MMGD without the MMGD-style preprocessing results

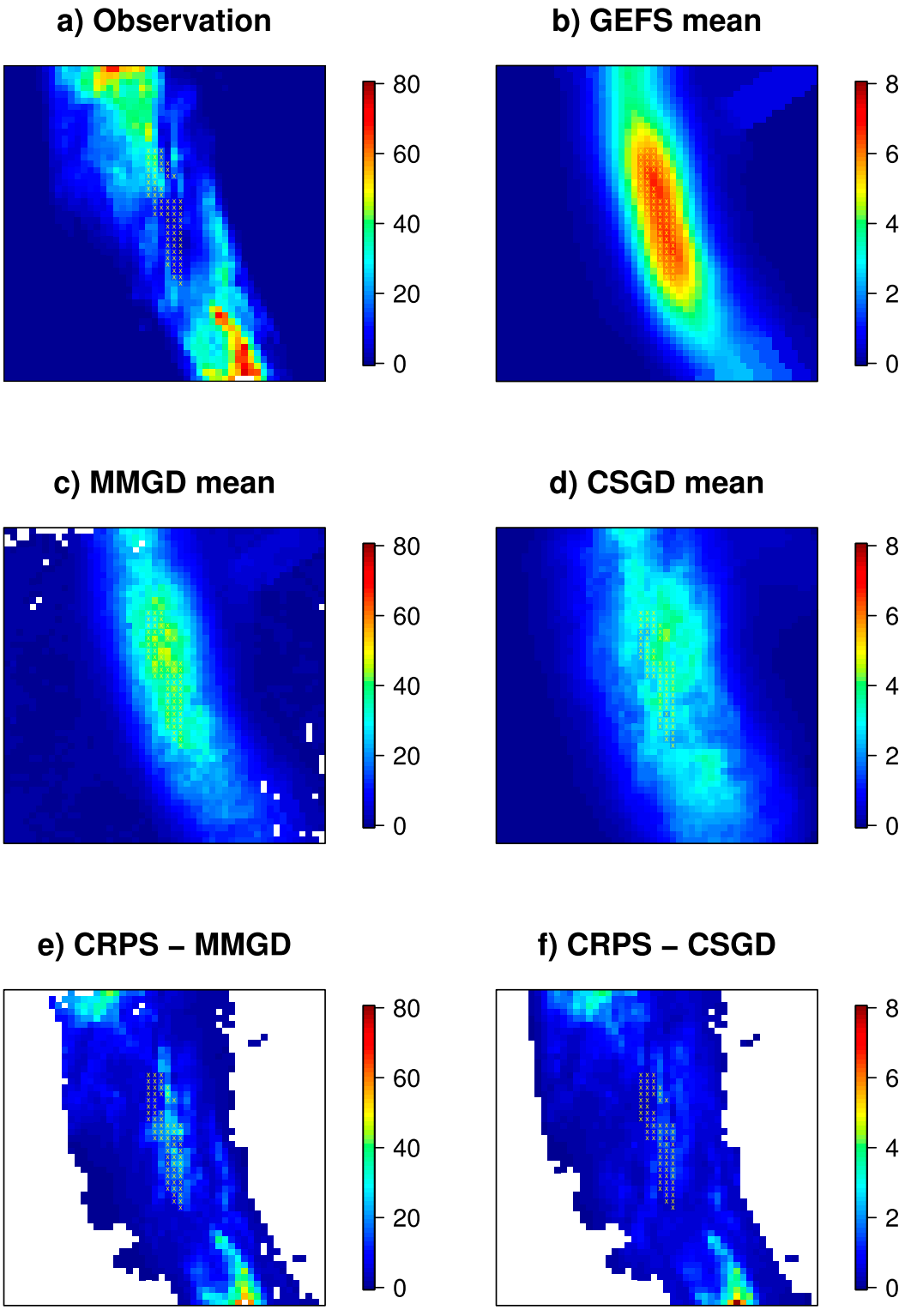

FIG. 11. Observed and forecast rainfall fields (mm) for the 6-h interval ending at 0000 UTC 1 Oct 2010: (a) observed (radar–gauge QPE), (b) ensemble mean from GEFS, (c) mean of MMGD PQPF, and (d) mean of CSGD PQPF. (e),(f) The CRPS for this interval for MMGD and CSGD PQPFs. Yellow crosses mark the pixels where the probability of exceeding 25 mm is between 0.8 and 0.9 per MMGD PQPF.

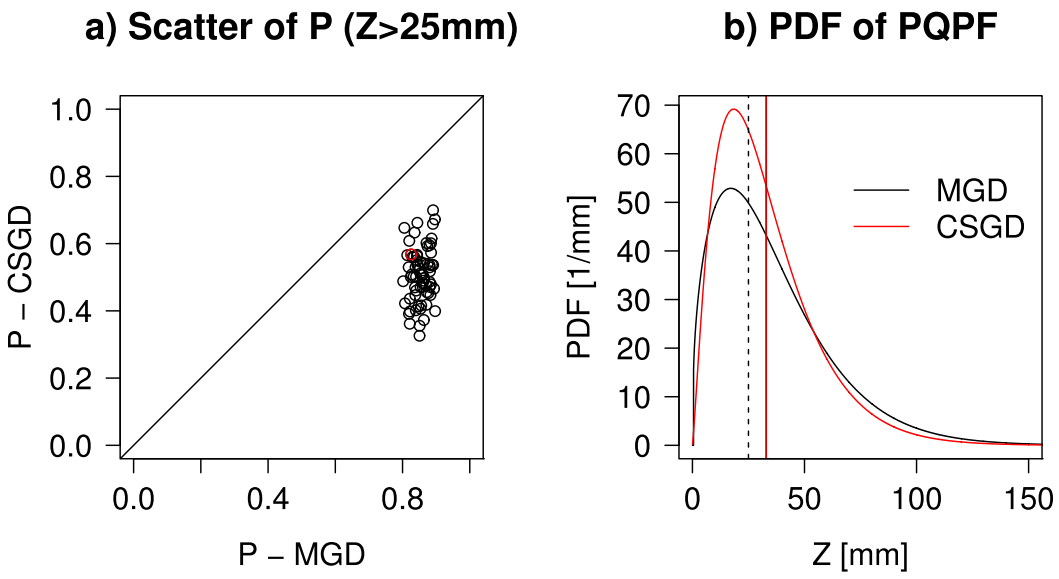

FIG. 12. (a) Scatterplot of probability of exceeding 25 mm based on MMGD and CSGD-P, where the red circle marks the pixel with a comparable conditional mean based on the two approaches but a large difference in exceedance probability, and (b) PDF for the aforementioned pixel, for the 6-h interval ending at 0000 UTC 1 Oct 2010. Vertical lines in (b) mark the 25 mm (dotted line) and PQPF means (solid lines).

in a much lower bias and a slightly lower RMSE. A possible explanation is that the use of a spatially smoothed forecast and observation creates a larger number of instances where both the forecast and observation are beyond zero, and this artificially introduces a wet bias in the MMGD-based PQPF. Meanwhile, spatial smoothing has the effect of reducing the heavier forecast amounts, which may have resulted in a more conservative MMGD-based PQPF for heavier events.

5. Summary and conclusions

This study assesses the skills of PQPFs obtained via two postprocessing mechanisms: MMGD, a mature, operational paradigm that is a part of HEFS, and CSGD, a more recent, regression-based method. While MMGD uses the exclusively raw ensemble mean as the conditional variable, CSGD has the ability to integrate additional predictors such as the ensemble spread and probability of precipitation. To discern the impacts of incorporating the latter predictors, both the full (CSGD-P) and a simplified version of CSGD (CSGD-S) were implemented, with the latter employing only the ensemble mean as does MMGD. PQPFs from MMGD and the two versions of CSGD were compared over the mid-Atlantic region of the United States on PQPF of 6-h accumulation for lead times between 24 and 72 h. Adjustments that are a part of training for CSGD, including the spatial smoothing of the forecast and observations and quantile corrections, are adopted in the MMGD calibration to highlight the differences related to structural factors.

Our assessment using multiple metrics indicates that CSGD in general outperforms MMGD, in terms of both the accuracy of the ensemble mean and the sharpness of the distribution. Interestingly, the simplified version of CSGD (CSGD-S), which uses only the GEFS ensemble mean as the predictor, while performing slightly worse than the full CSGD (CSGD-P), manages to outperform MMGD by a small margin. While the MMGD and both CSGD schemes tend to reduce the tail of the values, the latter are able to correct the aggregate bias to near neutral, whereas MMGD renders only minor changes to the bias. The CSGD schemes also show better ability in reducing the RMSE, especially over the summer season when RMSE is high. These features persist across the range of lead times investigated. Comparisons of the schemes over elevation zones suggest that the differential performance of the three schemes is to a large extent determined by the locations of the precipitation maxima, which are influenced by terrain and wind direction. For example, a wintertime maximum is present west of the Appalachians, pointing to the orographic influence of heavy storms. Both CSGD schemes outperform MMGD over this maximum, but this differential performance is more an indication of the higher skill of CSGD for heavy events and does not necessarily imply its relative robustness for orographic storms. It is also found that incorporating ensemble spread and POP yields much larger improvements in the skill of

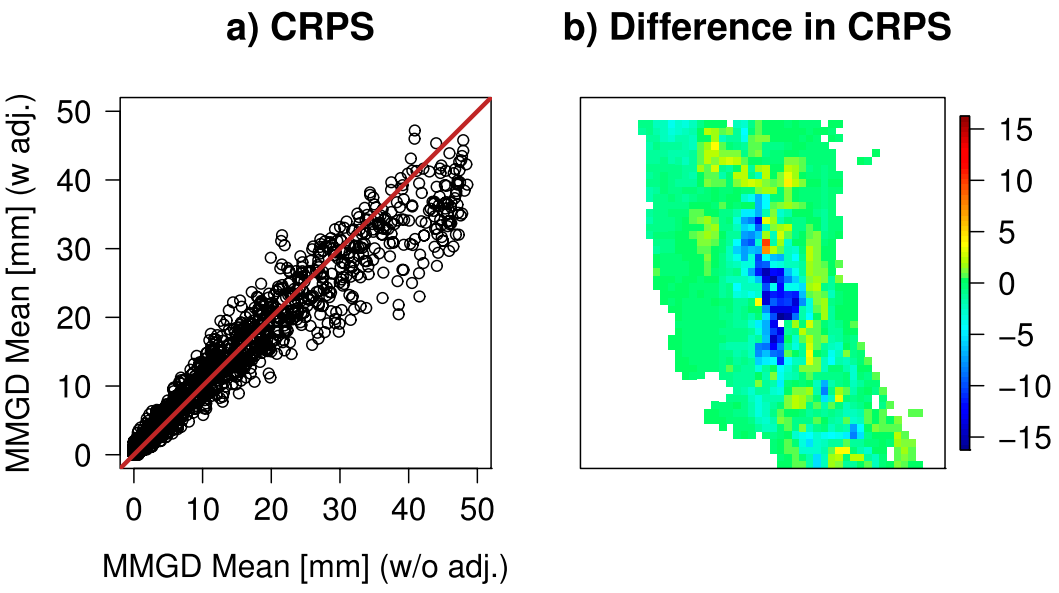

FIG. 13. (a) Scatterplot of CRPS for PQPFs based on MMGD without and with quantile adjustment and spatial smoothing for the 6-h interval ending at 0000 UTC 1 Oct 2010, and (b) spatial plot of the difference in CRPS for the PQPF based on MMGD with and without adjustment. Blue (red) colors indicate improvement (degradation) in skill.

CSGD over the summer than the winter: the CRPSS of CSGD-S drops below of that MMGD at higher elevations, whereas that of CSGD-P is consistently higher.

Further comparison of the three sets of PQPFs shows that CSGD-based PQPFs are in general more skillful based on the BSS, CRPSS, and reliability. Yet, it must be noted that the performance gap of CSGD and MMGD, as measured by the BSS, tends to contract at higher precipitation thresholds. Decomposition of the Brier score reveals a similar trend: the CSGD-P PQPF is overall more reliable and has a higher resolution than CSGD-S and MMGD at the lowest precipitation threshold (0.25 mm), but the differences among the PQPF sets gradually diminish at higher thresholds. The observations also suggest that integrating ensemble spread and POP primarily benefit *reliability* rather than *resolution* in distinguishing precipitating versus dry conditions and in identifying light precipitation events. Scrutiny of reliability also uncovers an issue of overforecasting by MMGD PQPF for a rare event on 1 October 2010, for which it indicates high probability of relatively heavy rain (beyond 25 mm) over locations where the actual rainfall amounts were much lower. For this event, GEFS was unable to capture the spatial structure of the storm, resulting in an overforecast of the amount over the light precipitation bands. Applying MMGD yields relatively minor corrections to the GEFS-based distribution. By comparison, applying the CSGD schemes yields a much sharper distribution with a lower mean and more conservative exceedance probability at higher thresholds.

The study suggests that the preprocessing measures, such as spatial smoothing and quantile correction of GEFS forecasts, are unlikely to be major contributors to the outperformance of CSGD. Partly responsible for this is that these measures lead to an increased number of instances of light precipitation amounts and reduced correlation. These effects may have distorted the forecast–observation relationship and made it less robust. Indeed, we found that the large positive bias in the MMGD PQPF arises mostly as a result of the reduced number of dry intervals that increases the probability of precipitation and distorts the marginal distribution of the forecast $F_X(x)$ [Eq. (4)]. Therefore, we posit that it is the CSGD's model structure, along with its use of optimization of CRPS, which is responsible for its outperformance of MMGD. This minimization of CRPS helps improve the accuracy of the mean as well as that of higher moments of the PQPF; it also to an extent compensates for any detrimental effects of preprocessing on the forecast of light precipitation amounts as seen in the MMGD outcome.

The overall outperformance confirms that CSGD is a viable, and a potentially more robust, alternative to MMGD for postprocessing medium-range ensemble precipitation forecasts both in the framework of HEFS and for the NWM's forcings engine. Typically, ensemble streamflow prediction is generated by feeding traces of ensemble forcings to a hydrologic model. Therefore, the sharpness and reliability of the precipitation forecast will to a large extent determine the sharpness and reliability of the ensemble streamflow forecast. Given

CSGD's outperformance for severe rainfall events, it is expected that its introduction to MEFP would yield tangible gain in the skill of predicting flooding events. In addition, the high-resolution NWM forecast relies on the interpolated raw ensemble precipitation forecast from GEFS as one of the key drivers. Either the CSGD or MMGD scheme can yield enhancement to the accuracy and sharpness of the PQPF. Implementation of these schemes in these systems would offer the opportunity to more comprehensively assess their relative impacts on hydrologic forecasts at different scales and in different hydroclimate regimes. Finally, while CSGD is shown to be effective in processing precipitation forecasts, a major challenge lies ahead in maintaining the physical relationship among multiple processed variables. Our current mechanism is the Schaake shuffle (Clark et al. 2004), while alternatives, including the Bayesian processor of ensemble approach (Krzysztofowicz and Evans 2008) and the empirical copula coupling Schefzik et al. (2013), will be investigated in the near future.

Acknowledgments. The work was performed as a part of the project for developing a forecast forcings engine at the National Weather Service's Office of Water Prediction and was not associated with a particular grant. This article benefited from discussions among team members Yuqiong Liu, Haksu Lee, Yuxiang He, and Thomas Adams. Tomislava Vukicevic at OWP and Zoltan Toth and Rob Cifelli at the NOAA/Earth System Research Laboratory provided critical feedback and suggestions, and their contributions are graciously acknowledged here.

APPENDIX

Definitions of the BSS and CRPSS

The BSS is based on the Brier score, which measures the distance between the probability that a certain event would occur and the actual outcome. In precipitation forecasts, it is common to consider an event in which a given threshold τ, is exceeded for the forecast variable. A formal definition is given below:

$$\mathrm{BS}(\tau) = \frac{1}{N}\sum_{i=1}^{N}[p_i(\tau) - o_i(\tau)]^2, \tag{A1}$$

where N is the number of forecast instances, p_i is the probability of exceeding threshold τ, and o_i is the binary function that takes 1 when the observation exceeds threshold τ and 0 otherwise. To gauge the effectiveness of a forecast, we also employ the BSS, which is the normalized difference between the Brier score of a forecast BS_p and that of a reference BS_c:

$$\mathrm{BSS} = -\frac{\mathrm{BS}_p - \mathrm{BS}_c}{\mathrm{BS}_c} \tag{A2}$$

Murphy (1973) demonstrates that the Brier score can be decomposed into three elements: reliability, resolution, and uncertainty:

$$\begin{aligned}\mathrm{BS}(\tau) &= \mathrm{REL}(\tau) - \mathrm{RES}(\tau) + \mathrm{UNC}(\tau)\\ &= \frac{1}{N}\sum_{i=1}^{K} n_i[p_i(\tau) - (\overline{o}_i)(\tau)]^2\\ &\quad - \frac{1}{K}\sum_{i=1}^{K}[\overline{o}_i(\tau) - (\overline{o})(\tau)]^2 + [\overline{o}(\tau)]\{1 - [\overline{o}(\tau)]\},\end{aligned} \tag{A3}$$

where REL, RES, and UNC denote an aggregate measure of reliability, resolution, and uncertainty; K is the number of forecast categories; n_i is the number of forecasts in probability category i; $\overline{o}$ is the mean climatological probability and is independent of the forecast.

CRPS can be considered as an integration of the Brier score over a full range of thresholds (Hersbach 2000; Grimit et al. 2006). Its definition is given below:

$$\mathrm{CRPS} = \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}[F_j^f(x) - F_j^o(x)]^2 dx, \tag{A4}$$

where N is the number of forecast instances, $F_i^f(x)$ is the CDF of the forecast at the forecast instance i, and $F_i^o(x)$ is the observed distribution that takes the form of Heaviside function $H(x - o_i)$. Grimit et al. (2006) shows that the CRPS can be considered as a composite measure of mean absolute error and the sharpness of a distribution and will degrade to the former for a deterministic forecast. The CRPSS is a quantity constructed in a way similar to the BSS, that is,

$$\mathrm{CRPSS} = -\frac{\mathrm{CRPS}_p - \mathrm{CRPS}_r}{\mathrm{CRPS}_r}, \tag{A5}$$

where CRPS_p is the CRPS for the forecast whereas CRPS_r is based on the reference (climatology).

REFERENCES

Ajami, N. K., G. M. Hornberger, and D. L. Sunding, 2008: Sustainable water resource management under hydrological uncertainty. *Water Resour. Res.*, **44**, W11406, doi:10.1029/2007WR006736.

Barros, A., and R. Kuligowski, 1998: Orographic effects during a severe wintertime rainstorm in the Appalachian Mountains. *Mon. Wea. Rev.*, **126**, 2648–2672, doi:10.1175/1520-0493(1998)126<2648:OEDASW>2.0.CO;2.

Brown, J. D., 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, doi:10.1016/j.jhydrol.2014.05.028.

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518, doi:10.1175/1520-0493(1998)126%3C2503:IOESOE%3E2.0.CO;2.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, **4**, 401–412, doi:10.1175/1520-0434(1989)004<0401:SFBOTN>2.0.CO;2.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeor.*, **5**, 243–262, doi:10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Demargne, J., and Coauthors, 2014: The science of NOAA's operational Hydrologic Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, doi:10.1175/BAMS-D-12-00081.1.

Du, J., G. DiMego, M. S. Tracton, and B. Zhou, 2003: NCEP short-range ensemble forecasting (SREF) system: Multi-IC, multi-model and multi-physics approach. *Research Activities in Atmospheric and Oceanic Modeling*, J. Cote, Ed., CAS/JSC Working Group Numerical Experimentation Rep. 23, WMO/TD 1161, 5.09–5.10.

Georgakakos, A. P., H. Yao, M. Mullusky, and K. P. Georgakakos, 1998: Impacts of climate variability on the operational forecast and management of the upper Des Moines River basin. *Water Resour. Res.*, **34**, 799–821, doi:10.1029/97WR03135.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson, 2006: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Wea. Forecasting*, **132**, 1–17, doi:10.1256/qj.05.235.

Hamill, T. M., and J. S. Whitaker, 2006: Probability quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, doi:10.1175/MWR3237.1.

——, ——, M. Fiorino, and S. G. Benjamin, 2011: Global ensemble predictions of 2009's tropical cyclones initialized with an ensemble Kalman filter. *Wea. Forecasting*, **139**, 668–688, doi:10.1175/2010MWR3456.1.

——, G. Bates, J. Whitaker, D. Murray, M. Fiorino, T. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

Herr, H. D., and R. Krzysztofowicz, 2005: Generic probability distribution of rainfall in space: The bivariate model. *J. Hydrol.*, **306**, 234–263, doi: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, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hosking, J. R. M., 1990: L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. Roy. Stat. Soc.*, **52B**, 105–124, http://www.jstor.org/stable/2345653.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242, doi:10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2.

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

Kitzmiller, D., D. Miller, R. Fulton, and F. Ding, 2013: Radar and multisensor precipitation estimation techniques in National Weather Service hydrologic operations. *J. Hydrol. Eng.*, **18**, 133–142, doi:10.1061/(ASCE)HE.1943-5584.0000523.

Krzysztofowicz, R., and W. B. Evans, 2008: Probabilistic forecasts from the National Digital Forecast Database. *Wea. Forecasting*, **23**, 270–289, doi:10.1175/2007WAF2007029.1.

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

Pagano, T. C., and Coauthors, 2014: Challenges of operational river forecasting. *J. Hydrometeor.*, **15**, 1692–1707, doi:10.1175/JHM-D-13-0188.1.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/MWR2906.1.

Reed, S. M., and D. R. Maidment, 1999: Coordinate transformations for using NEXRAD data in GIS-based hydrologic modeling. *J. Hydrol. Eng.*, **4**, 174–182, doi:10.1061/(ASCE)1084-0699(1999)4:2(174).

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, doi:10.1214/13-STS443.

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

Seo, D.-J., A. Seed, and G. Delrieu, 2011: Radar and multisensor rainfall estimation for hydrologic applications. *Rainfall: State of the Science*, *Geophys. Monogr.*, Vol. 191, Amer. Geophys. Union, 79–104.

Smith, J. A., M. L. Baeck, A. Ntelekos, G. Villarini, and M. Steiner, 2011: Extreme rainfall and flooding from orographic thunderstorms in the central Appalachians. *Water Resour. Res.*, **47**, W04514, doi:10.1029/2010WR010190.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79, doi:10.1111/j.1600-0870.2007.00273.x.

Wu, L., 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, doi:10.1016/j.jhydrol.2011.01.013.

Zhang, Y., S. Reed, and D. Kitzmiller, 2011: Effects of retrospective gauge-based readjustment of multisensor precipitation estimates on hydrologic simulations. *J. Hydrometeor.*, **12**, 429–443, doi:10.1175/2010JHM1200.1.