Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction

Limin Wu [a,b,*], Dong-Jun Seo [a,c], Julie Demargne [a,c], James D. Brown [a,c], Shuzheng Cong [a,c], John Schaake [a,d]

[a] National Oceanic and Atmospheric Administration, National Weather Service, Office of Hydrologic Development, 1325 East-West Highway, Silver Spring, MD 20910, USA
[b] Wyle Information Systems, 1651 Old Meadow Road, McLean, VA 22102, USA
[c] University Corporation for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307, USA
[d] Consultant, 1A3 Spa Creek Landing, Annapolis, MD 21403, USA

ARTICLE INFO

SUMMARY

Reliable and skillful precipitation ensemble forecasts are necessary to produce reliable and skilful hydrologic ensemble forecasts. It is well known that, in general, raw precipitation ensemble forecasts from the numerical weather prediction (NWP) models are not very reliable and that, for short-range prediction, human forecasters add significant skill to the NWP-generated single-valued quantitative precipitation forecasts (QPF). In this paper, we describe and evaluate a statistical procedure for producing precipitation ensemble forecasts from single-valued QPFs. The procedure is based on the bivariate probability distribution between the observed precipitation and the single-valued QPF. The distribution is modeled as a mixed-type in which the relationship between the positive observed precipitation and positive forecast precipitation is assumed to be bivariate meta-Gaussian. We also describe and comparatively evaluate a generalized meta-Gaussian model in which the model parameter is optimized by minimizing the mean Continuous Ranked Probability Score. The performance of these procedures is assessed through dependent and cross validation using data for selected river basins in the service areas of the Arkansas-Red Basin, California-Nevada and Middle-Atlantic River Forecast Centers of the National Weather Service. The validation results show that, overall, the precipitation ensembles generated by the proposed procedures are reliable and capture the skill in the conditioning single-valued forecasts very well.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Reliable and skillful ensemble forecasts of precipitation and temperature are needed for reliable and skillful hydrologic ensemble forecasts (Demargne et al., 2007, 2010; Seo et al., 2006). Today, ensemble forecasts from numerical weather prediction (NWP) models are widely available from many sources. For example, the National Weather Service's (NWS) National Centers for Environmental Prediction (NCEP) issues short-, medium- and long-range ensemble forecasts from a suite of regional and global models, such as the Short-Range Ensemble Forecast (SREF) System, the Global Ensemble Forecast System (GEFS) and the Climate Forecast System (CFS). However, it is well known that, in general, the raw ensemble forecasts from NWP models are biased in the mean and spread. Also, even if they may not be biased at the model grid scale, they may be biased at the catchment scale, depending on the size of the basin. Correcting such biases, often referred to as post-

processing or statistical calibration, is an active area of research (Eckel and Walters, 1998; Gneiting et al., 2005, 2007; Hamill and Whitaker, 2006, Hamill et al., 2008). Bias correction for precipitation ensemble forecasts, however, has proven very challenging due to the large space–time variability of precipitation. It is expected that significant additional efforts will be needed to produce operational ensemble forecasts that are sufficiently reliable for hydrological applications, particularly for large precipitation amounts.

In the single-valued forecast process, human forecasters play an important role in improving the quality of hydrometeorological forecasts. For example, in the NWS, there are at least three places where human forecasters may add value to single-valued forecasts: the NCEP's Hydrometeorological Prediction Center (HPC), the Weather Forecast Offices (WFO) and the River Forecast Centers (RFC). At short ranges, such forecaster-modified forecasts are often considerably better than the raw NWP forecasts (Charba et al., 2003). Given the above state of operational forecasting of hydrometeorological variables, there is a need for techniques that can generate, from the forecaster-modified single-valued forecasts, ensemble forecasts that are reliable at the river basin scale for hydrologic applications.

---

* Corresponding author. Address: NWS/OHD,1325 East-West Hwy., 8th Floor, Silver Spring, MD 20910, USA. Tel.: +1 301 713 0640x123; fax: +1 301 713 0963.
 E-mail address: limin.wu@noaa.gov (L. Wu).

Recently, Schaake et al. (2007) developed such a technique; it generates ensemble forecasts of precipitation from single-valued quantitative precipitation forecasts (QPF) based on the lead-time-specific probability distribution of observed precipitation conditional on the single-valued QPF (hereafter referred to as QPF for brevity). The random numbers sampled from the conditional probability distributions for all lead times are then "shuffled" (Clark et al., 2004) to generate ensemble members that possess the spatiotemporal structure in the historically-observed precipitation in the ranked correlation sense. Referred to as the Ensemble Pre-Processor (EPP) in NWS, this technique, which is described in Schaake et al. (2007), has been used experimentally at a number of RFCs for several years, from which substantial experience has been gained, and areas of improvement have been identified. The primary purpose of this paper is to describe the improved procedures for statistical generation of precipitation ensemble forecasts from QPFs, and to present validation results from the perspective of operational hydrologic ensemble forecasting at the RFCs. Statistical techniques usually require large amounts of data for parameter estimation or calibration. The larger the calibration dataset, the smaller the sampling uncertainty will be in the modeled joint distribution of the single-valued forecast and verifying observation, all other factors being equal. Hamill et al. (2008) reports that a small training dataset was adequate for calibration of short-lead temperature forecasts but that, for precipitation, calibration using a 20-year dataset of weekly reforecasts greatly improved forecast skill when compared with calibration using a 30-day dataset. The historical archive at the RFCs of the forecaster-modified single-valued forecasts typically goes back only a few to several years. Therefore, we are particularly interested in assessing the data requirements for parameter estimation of the procedures described in this paper. This assessment is conducted through cross validation experiments. Dependent validation experiments are also carried out to assess the goodness of the statistical models used in the improved procedures under negligible sampling uncertainty.

The new contributions made by this work comprise: (1) an improved procedure based on a mixed-type bivariate meta-Gaussian distribution model for statistical generation of precipitation ensemble forecasts from QPF, and its evaluation; (2) a parameter optimization procedure for a generalized meta-Gaussian
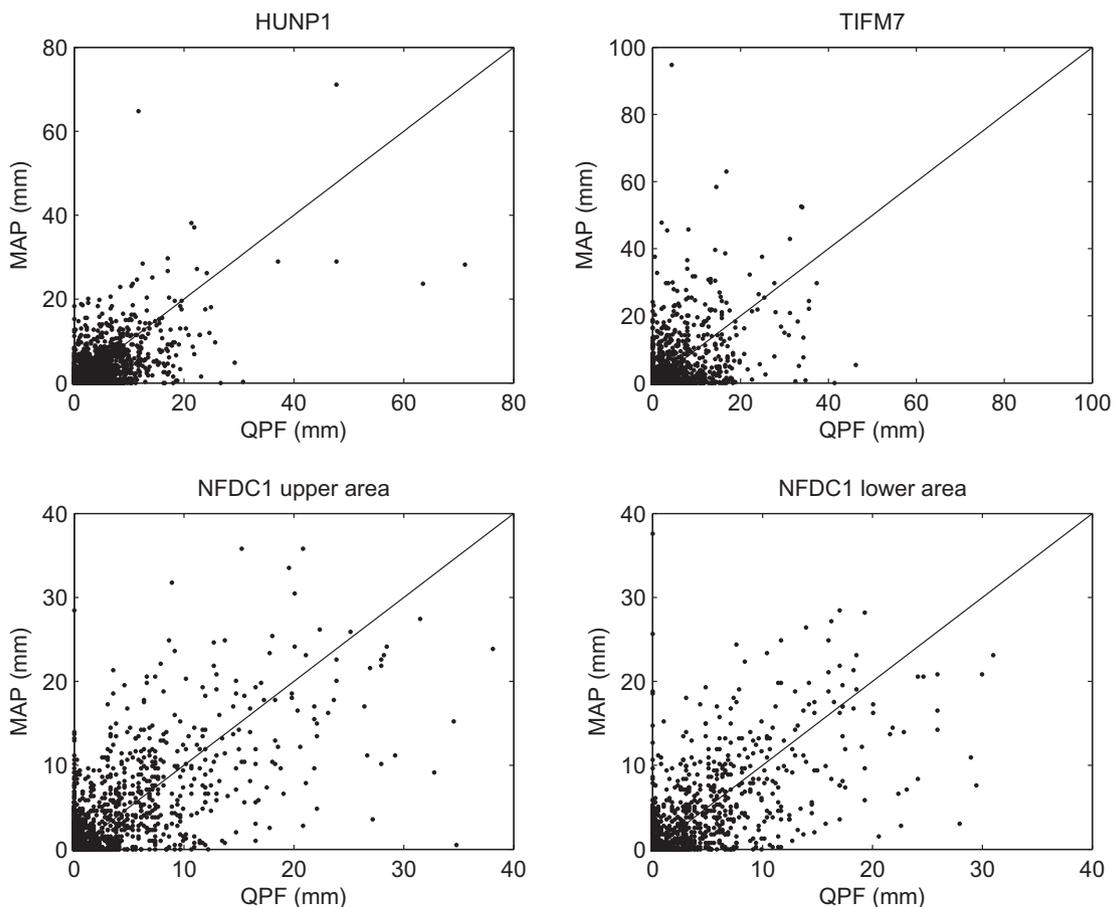
Table 1
Differences among the methods.

	Method 1	Method 2	Method 3
Mixed-type bivariate structure	No	Yes	Yes
Generalized meta-Gaussian model	No	No	Yes
Parameter optimization	No	No	Yes

Table 2
Data archive for the study river basins.

Basin ID	MAP	QPF (Days 1 and 2)
HUNP1	1 Oct 1948–12 Jul 2008	1 August 2002–12 Jul 2008
TIFM7	1 Jan 1961–30 Jun 2008	1 Jul 2003–30 Jun 2008
NFDC1	1 Oct 1960–30 Sep 2004	1 Oct 2000–30 Sep 2004

Fig. 1. (a) Scatter plots of 6-h MAP and QPF for Day 1 for HUNP1 (1 August 2002–12 July 2008), TIFM7 (1 July 2003–20 June 2008), and NFDC1 (1 October 2000–30 September 2004). (b) Same as (a) but on base 10 log–log scales.
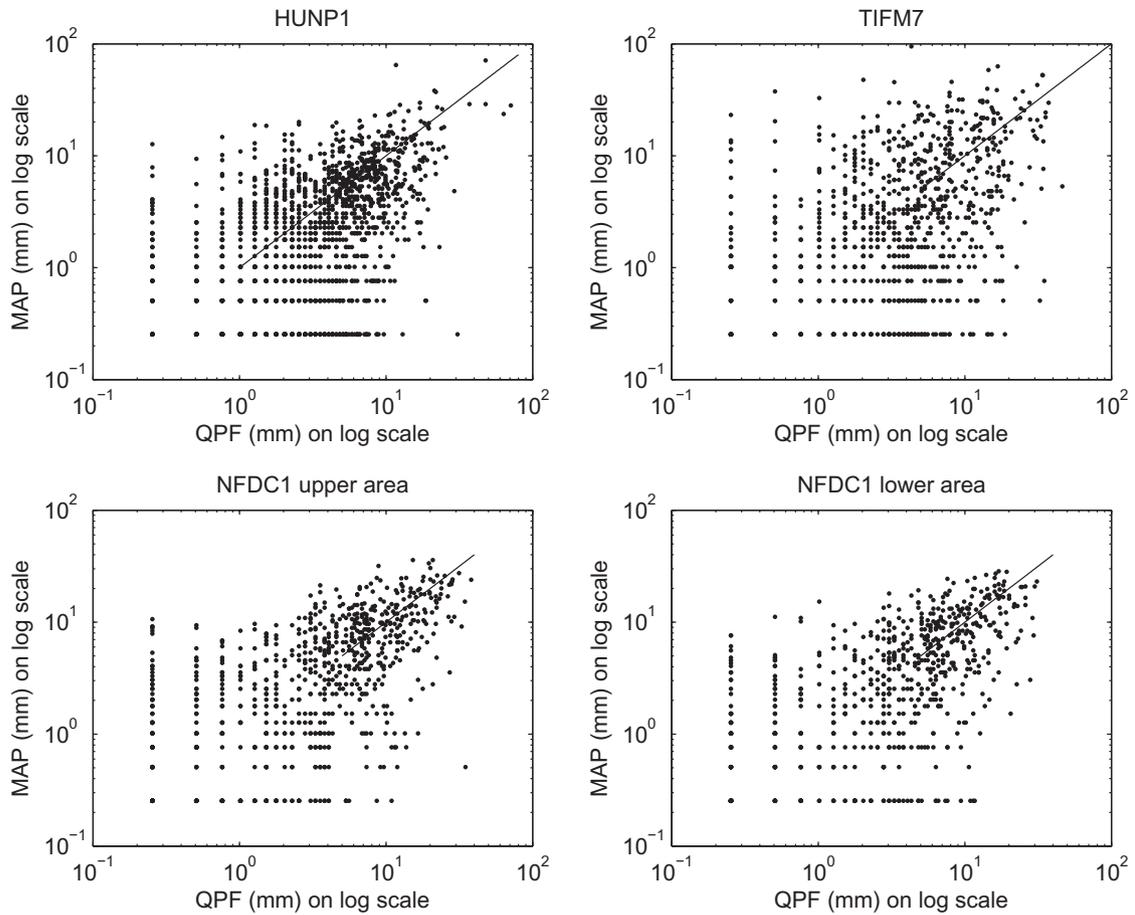
Fig. 1 (*continued*)

distribution model, and its evaluation; (3) new expressions for the conditional distributions associated with the mixed-type bivariate distribution of Herr and Krzysztofowicz (2005); and (4) an alternative formulation for the precipitation intermittency approximation of Schaake et al. (2007) in the form of the meta-Gaussian model.

This paper is organized as follows. Section 2 describes the problem and the solution approaches. Section 3 describes the study river basins and the data used. Section 4 describes the estimation procedure for the statistical parameters. Section 5 presents the dependent and cross validation results. Section 6 summarizes the conclusions and future research recommendations.

2. Description of the procedures

For a deterministic hydrologic forecast system, it is known that the uncertainty in the QPF is a primary source of forecast error. In this work, we develop statistical procedures to capture this uncertainty. The resulting precipitation ensemble forecasts are used to drive the hydrologic model to produce hydrologic ensemble forecasts. In single-valued hydrologic forecasting with lumped hydrologic models at the RFCs, 6-h QPFs are input to the hydrologic models to produce streamflow forecasts at basin outlets. The input QPF values are the Mean Areal Precipitation (MAP) amounts predicted to fall within the basin, which is referred to as the Forecast Mean Areal Precipitation (FMAP). If the historical relationship between the forecast precipitation and the corresponding observed precipitation is known, stationary in time, and representative of the future, the uncertainty in the QPF may be modeled via the historical relationship between the forecasts and the verifying obser-

vations under similar conditions. The idea then is to generate, for each 6-h period of lead time, an ensemble precipitation forecast from the conditional distribution of observed precipitation given the 6-h QPF. Because of intermittency, we model precipitation as a binary-continuous variable. The joint distribution of the forecast and observed precipitation amounts is then of mixed-type (Herr and Krzysztofowicz, 2005). The continuous–continuous part can be modeled as bivariate meta-Gaussian (Kelly and Krzysztofowicz, 1997), which supplies a simple mathematical structure and analytical expressions for the above conditional distribution. Below, we describe the development in Herr and Krzysztofowicz (2005) that is necessary to obtain the expressions for the conditional distribution.

Let X and Y denote the 6-h QPF and the corresponding observed MAP, respectively. Define the joint probability distribution function of X and Y by F:

$$F(x,y) \equiv P(X \leqslant x, Y \leqslant y). \qquad (1)$$

The above can be decomposed as:

$$F(x,y) = p_{00} + p_{10}G_X(x) + p_{01}G_Y(y) + p_{11}D(x,y), \qquad (2)$$

where

$$p_{00} \equiv P(X = 0, Y = 0), \qquad (3)$$

$$p_{10} \equiv P(X > 0, Y = 0), \qquad (4)$$

$$p_{01} \equiv P(X = 0, Y > 0), \qquad (5)$$

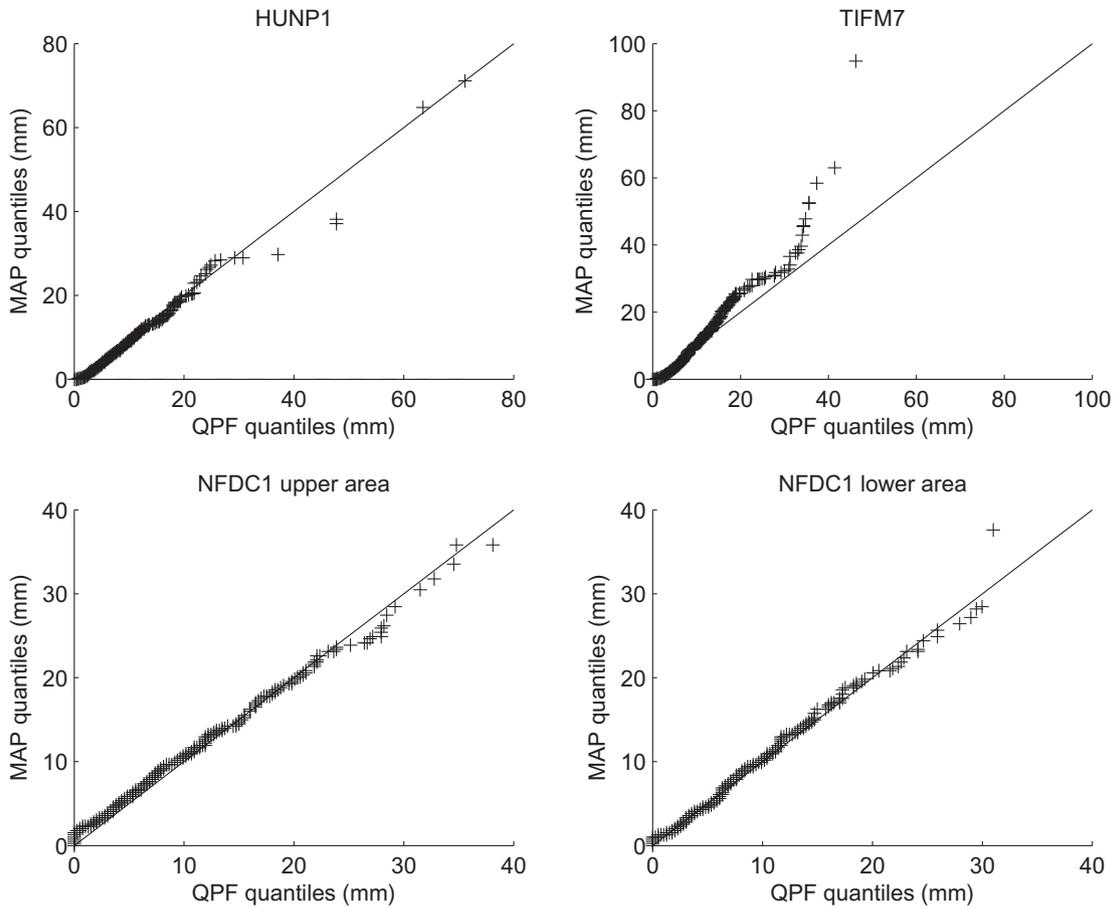$$p_{11} \equiv P(X > 0, Y > 0), \qquad (6)$$

Fig. 2. Same as Fig. 1 but quantile–quantile plots.

$$G_X(x) \equiv P(X \leqslant x | X > 0, Y = 0), \tag{7}$$

$$G_Y(y) \equiv P(Y \leqslant y | X = 0, Y > 0), \tag{8}$$

$$D(x,y) \equiv P(X \leqslant x, Y \leqslant y | X > 0, Y > 0). \tag{9}$$

Note that the conditional marginal cumulative distribution functions (CDF) of $(X \mid X > 0, Y > 0)$ and $(Y \mid X > 0, Y > 0)$ are continuous and so is their joint CDF, $D(x, y)$. We may then model $D(x, y)$ as bivariate meta-Gaussian as described in Appendix A.

To obtain expressions for the conditional CDF, $P(Y \leqslant y \mid X = x)$, we first define:

$$F_{Y|X}(y|x) \equiv P(Y \leqslant y | X = x). \tag{10}$$

We further define:

$$D_X(x) \equiv D(x, \infty) = P(X \leqslant x | X > 0, Y > 0), \tag{11}$$

$$D_{Y|X}(y|x) \equiv P(Y \leqslant y | X = x, X > 0, Y > 0), \tag{12}$$

$$F_{X|X>0}(x) \equiv P(X \leqslant x | X > 0). \tag{13}$$

Then we have for Eq. (10) for $x = 0$:

$$F_{Y|X}(y|x) = a + (1-a)G_Y(y), \tag{14}$$

where $a = p_{00}/(p_{00} + p_{01})$. If $p_{00} + p_{01} = 0$, then $P(X = 0, Y \geqslant 0) = 0$, which means we cannot define conditional probability of Y given $X = 0$ for this case. For $x > 0$, we have:

$$F_{Y|X}(y|x) = c(x) + (1 - c(x))D_{Y|X}(y|x), \tag{15}$$

where

$$c(x) = \frac{p_{10}g_X(x)}{p_{10}g_X(x) + p_{11}d_X(x)} \tag{16}$$

In Eq. (16), $g_X(x)$ and $d_X(x)$ denote the probability density functions (PDF) corresponding to $G_X(x)$ in Eq. (7) and $D_X(x)$ in Eq. (11), respectively. If $p_{10} = 0$, Eq. (15) becomes $F_{Y|X}(y|x) = D_{Y|X}(y|x)$. If $p_{10} = 0$ and $p_{11} = 0$, $F(x, y)$ degenerates into a univariate CDF without arriving at Eq. (16). Note that $F_{Y|X}(y|x)$ in Eqs. (14) and (15), which correspond to Eqs. (10a) and (10b) of Herr and Krzysztofowicz (2005), respectively, is a composite of a discrete and a continuous distribution. Eqs. (14)–(16) constitute the mixed-type meta-Gaussian model proposed in this work and are used to generate the precipitation ensembles. Section 4 describes how the distributions and parameters in these equations are estimated.

Schaake et al. (2007) use an alternative procedure, which essentially models each marginal CDF as a combination of two continuous distributions, the first approximating the probability mass at zero and the second describing the probability distribution of positive precipitation. For description of the alternative procedure in the form of the meta-Gaussian model, the reader is referred to Appendix C. This procedure does allow the normal quantile transform (NQT, Kelly and Krzysztofowicz, 1997) of observed and forecast precipitation. However, the correlation coefficient between the two transformed variates now represents a weighted combination of the skill in the QPF not only for predicting the amount of precipitation but also for predicting the occurrence of precipitation, where the relative weight depends on the magnitude of the probability of precipitation (PoP). As such, estimation of the above
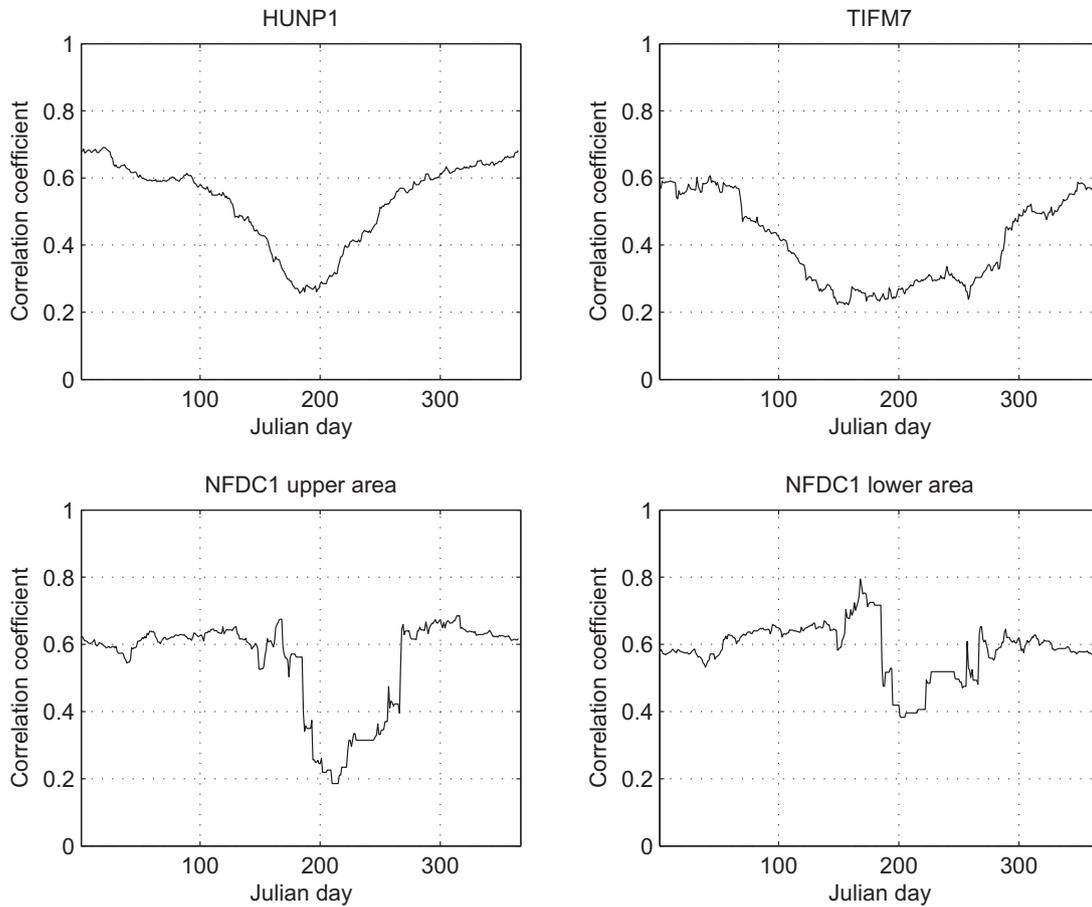
Fig. 3. Sample correlation coefficient of NQT-transformed 6-h MAP and Day-1 QPF for HUNP1, TIFM7, and NFDC1 upper and lower areas.

correlation coefficient as well as its use in the ensemble generation is tricky. In Schaake et al. (2007), this correlation coefficient is specified by a weighted average of the Pearson product-moment correlation coefficient of the untransformed variates (including zeros) and that of the transformed variates that best fit p_{00}, p_{10}, p_{01}, and p_{11} in Eqs. (3)–(6).

The distribution $D(x, y)$ of $(X \leqslant x, Y \leqslant y \mid X > 0, Y > 0)$ in (9) is modeled as bivariate meta-Gaussian. In modeling $D(x, y)$, we transform its marginal distributions associated with X and Y to standard normal distributions Z and W, respectively, via normal quantile transform. The joint distribution of Z and W is assumed bivariate standard normal. Under this assumption, the meta-Gaussian model of $D(x, y)$ will be identical to $D(x, y)$ itself (Appendix A). In hydrologic applications, this assumption may not be satisfied. If this assumption does not hold, then the conditional mean and variance of the modeled distribution may not match those of the original distribution $D(x, y)$ well, resulting in a lack of reliability in the ensembles generated from the model. Below, we develop a generalization of the meta-Gaussian model in an attempt to address this problem. This generalized model provides a parameter that can be used to adjust the model's dependence structure so that the ensembles generated are optimal under the criterion of choice. Such an approach has been used successfully by Seo et al. (2006).

Consider the following linear model:

$$U = bZ + \Theta_b, \qquad (17)$$

where $b > 0$, $\Theta_b \sim N(0, Var[W - bZ])$, W and Z are standard normal variates with correlation coefficient ρ, and Θ_b is assumed independent of Z. We have $Var[\Theta_b] = 1 + b^2 - 2b\rho$, $U \sim N(0, Var[\Theta_b] + b^2)$ and $\gamma = b/(Var[U])^{1/2}$, where γ denotes the correlation coefficient

between Z and U. The joint distribution of (Z, U) is bivariate normal. When $b = \rho$, (Z, U) is bivariate standard normal. The conditional mean of U given Z is $E[U \mid Z = z] = bz$. The conditional variance of U given Z is $Var[U \mid Z = z] = 1 + b^2 - 2b\rho$ and we have $Var[U \mid Z = z; b] \geqslant Var[U \mid Z = z; b = \rho]$. In this work, ρ is estimated as the sample correlation coefficient between the NQT-transformed variates Z and W. Appendix B provides an expression for the conditional distribution that can be used to produce ensembles. Section 4 describes how b is estimated.

The three procedures described in this section are all related to the meta-Gaussian model, yet they differ in significant ways. The procedure described in Schaake et al. (2007), referred to as Method 1 hereafter, can be understood as application of meta-Gaussian model to precipitation amounts with implicit treatment of precipitation intermittency. The mixed-type bivariate structure, on the other hand, treats precipitation intermittency explicitly. The continuous–continuous component of this structure can be modeled as the meta-Gaussian distribution (Method 2) or as an extension of the meta-Gaussian distribution with an adjustable parameter (Method 3). The differences among these three procedures are contrasted in Table 1.

Any one of the three procedures described above may be used to generate ensemble traces of FMAP for the basin of interest. To generate ensemble traces over the entire forecast horizon, the procedure is repeated for each time step within the forecast horizon using the parameters specific for the time step. For any given forecast time step, the ensemble generation process consists of two steps. First, a forecast probability distribution as the conditional distribution given the single-valued forecast is estimated. Next, the Schaake Shuffle (Clark et al., 2004) is applied. The Schaake

Shuffle can be understood as a procedure that connects the ensemble members at consecutive time steps within the forecast horizon such that the ensemble traces exhibit, in the rank correlation sense, the historically-observed temporal variability. Here, we give a description of the procedure. For a given forecast time step, we first construct an ensemble of MAP values collected across the historical years available in the data archive and ordered by the historical years. Then, a sample of the size of that of the MAP ensemble is drawn from the forecast distribution. The sample points may come out in any order. The Schaake Shuffle arranges the sample points in such a way that the nth largest sample point in the forecast ensemble has the same position as the nth largest member of the MAP ensemble so that the forecast ensemble acquires the prescribed order for its members. The Schaake Shuffle, therefore in this respect, is an ordering process. The Schaake Shuffle can also be understood as an adjusting process with respect to the MAP ensemble. The ratio of each member of the MAP ensemble to the same-rank member of the forecast sample can be calculated. The forecast ensemble can be understood as obtained from the MAP ensemble through the multiplicative operation of mapping the corresponding members using the ratio (special treatment for the members of 0-value is needed). In short, the Schaake Shuffle amounts to adjusting the observed historical ensemble traces so that, for each time step in the forecast horizon, the empirical CDF of the adjusted historical ensemble members matches the conditional CDF estimated from one of the three procedures. The end result is a set of forecast ensemble traces with similar temporal variability as the historical MAP ensemble traces. Similarly, the Schaake Shuffle also applies to multiple basins provided the same historical years are used to construct the ensembles across the basins. The Schaake Shuffle, therefore, replicates in the forecast ensemble traces the spatiotemporal rank correlation of the observed historical ensemble traces.

3. Study basins and data used

The procedures described above are evaluated using real-world data for selected river basins in the service areas of the Arkansas-Red Basin (AB-), California-Nevada (CN-) and Middle Atlantic (MA-) River Forecast Centers (RFC). The QPFs are produced by the RFCs based on the guidance products from NCEP HPC. The 6-h observed MAP values come from rain gauge-only or multisensor (Seo and Breidenbach, 2002) analysis.

The following three basins are used in the study, the Huntingdon basin (basin ID: HUNP1; drainage area: 2113 km^2) of the Juniata River in Pennsylvania, the Elk River basin at Tiff City (basin ID: TIFM7; drainage area: 2258 km^2) in Arkansas and Missouri, and the North Fork basin of the American River (basin ID: NFDC1; drainage area: 890 km^2) in California. NFDC1 consists of an upper- and a lower-elevation zone. A historical archive of 6-h MAP and the corresponding 6-h QPF for a forecast horizon of 2 days (forecast period 0–24 h is referred to as Day 1 and 24–48 h Day 2 hereafter) are available for the three basins (see Table 2). Note that the observed data span a much longer period than the QPF data. This is necessary for the Schaake Shuffle (Clark et al., 2004) used at the end of the ensemble generation process (see Section 5.1).

Figs. 1 and 2 show the scatter and quantile–quantile plots, respectively, of Day-1 6-h QPF for all 6-h sub-periods and the corresponding MAP for the study basins. Note in Fig. 1a that the QPF for NFDC1 shows the largest skill owing to the larger predictability of orographic precipitation whereas the QPF for TIFM7 shows the smallest due to significantly smaller predictability of convective precipitation. The skill contrasts clearly in Fig. 1b, where base 10 log–log scales are used. Note in Fig. 2 that the QPFs for NFDC1 and HUNP1 are remarkably unbiased, a reflection of the added skill provided by the human forecasters.

4. Parameter estimation

To estimate the parameters for the mixed-type meta-Gaussian model, historical data are needed, comprising the QPF and corresponding MAP for each of the 6-h time steps. Three aspects were considered in parameter estimation: sub-daily-scale variations in QPF skill and diurnal variations in variability of precipitation, seasonal variations in QPF skill and variability of precipitation, and the length of the historical records. To account for seasonal variations, a moving window was used to subset, for a given 6-h time step, the QPF and MAP data from all historical years that fall within the time window centered at each Julian day of the year. Often, seasonal changes in the precipitation regime occur over a relatively short period of time. Due to limited sample size, however, it was not possible to select a sufficiently small window. In this work, we used a compromise choice of 91 days for the window size. The skill in the 6-h QPF varies from one 6-h period to another. Also, there often exist strong diurnal variations in variability and predictability of precipitation and so it is desirable to obtain the statistics at a 6-h scale. In this work, however, the sample size was too small to estimate the parameters for each 6-h time step. Consequently, for all basins we pooled all 6-h data within each 24-h period, thereby ignoring sub-daily-scale variations in QPF skill and diurnal variations in precipitation variability.

The marginal distributions associated with $d(x, y)$ in Eq. (9), namely, the distribution of $(X \leqslant x \,|\, X > 0, Y > 0)$ and that of $(Y \leqslant y \,|\, X > 0, Y > 0)$, may be estimated using parametric or nonparametric techniques. We found that the 2-parameter Weibull distribution fits the data well for the AB- and MARFC basins. For the CNRFC basin NFDC1, on the other hand, lack of fit was evident. Other parametric distributions were also tried, including the 2-parameter Gamma, 3-parameter Pearson Type III, and 4-parameter kappa distributions using the L-moment method (Hosking, 2005). None of these distributions, however, provided a good fit for the MAP data of NFDC1. We then tried nonparametric distributions using Gaussian kernel smoothing with plug-in optimal bandwidth selection. Azzalini (1981) and Reiss (1981) demonstrate that nonparametric distribution estimation with kernel smoothing can outperform the classic empirical distribution estimation with asymptotic efficiency gains. Choosing a bandwidth parameter is required in kernel smoothing. In this work, we use the plug-in bandwidth estimation technique proposed by Hansen (2004), which minimizes an estimate of the asymptotic mean integrated squared error. This nonparametric technique worked well for the NFDC1 MAP data as well as for the AB- and MARFC basins. A less than satisfactory aspect of the nonparametric technique, however, is that the estimated distribution may exhibit a step function-like behavior in the data-sparse upper tail. The Gaussian kernel smoothing technique was used to produce the validation results presented in Section 5.

The density functions in Eq. (16) were estimated by parametric distribution modeling. Estimating $g_X(x)$ (see Eq. (7)) in Eq.(16), however, is challenging for two reasons: this density function may be highly skewed and the number of data points for $(X \,|\, X > 0, Y = 0)$ may be very small. We tried various distribution models available in Hosking (2005). We found that the 3-parameter Pearson type III distribution works best in most cases. In some cases, the 4-parameter kappa distribution works best. Nonparametric distributions with the Gaussian kernel were also tried, but found to have excessive fluctuations. We also tried an empirical method (Cong et al., 2006) for estimating $c(x)$ in Eq. (16), but found the results to be sensitive to the sampling procedure of this method. Appendix D provides three alternative expressions for $c(x)$ in Eq. (16), which may allow for larger samples for estimating $c(x)$ under certain situations. In producing the validation results pre-

sented in Section 5, we used the 3-parameter Pearson type III distribution for the density functions in Eq. (16).

Fig. 3 shows the sample correlation coefficient between the 6-h QPF of Day-1 and the corresponding MAP in the bivariate transformed space for HUNP1, TIFM7, and both the upper and lower zones of NFDC1. The figure shows the seasonality and regional variations in the predictive skill in the QPF for these basins and, for NFDC1, variations in elevation as well. The time window used for pooling the observed and forecast pairs of precipitation data was 91 days. With such a large window size, the correlation coefficient over seasonal transition periods may be overly smoothed as the seasonal transition in precipitation regime often takes place in a very short period of time. It is important to point out that, because the correlation coefficient used in the procedure is conditioned only on the time of the year, the procedure may incorrectly prescribe the correlation coefficient for "out-of-season" storms, possibly resulting in over- or under-spread in the conditional CDF. Such storm type- or regime-dependent modeling of the statistics is a future endeavor.

For the generalized linear model in Eq. (17), the parameter b needs to be estimated. In this work, we experimented with two objective functions in estimating an optimal b value. The primary one is the mean Continuous Ranked Probability Score (CRPS) (Hersbach, 2000), which is one of the most widely used performance measures in ensemble forecast verification. The mean CRPS measures the overall quality of probabilistic forecasts as the expected squared error of the forecast probabilities for all possible events (Jolliffe and Stephenson, 2003). As noted in the introduction, one of the overriding criteria for the proposed procedures is that the ensembles fully capture the skill in the single-valued QPF. In this regard, the mean CRPS is a particularly attractive criterion; it is equivalent to the mean absolute error for single-valued forecasts and hence allows quantitative comparison between ensemble and single-valued forecasts. The other objective function we used is the Root Mean Square Error (RMSE) of the observed and the corresponding ensemble 60th percentile. The RMSE may be used when large errors of the observed and forecast values are particularly undesirable.

5. Validation

In this section, we present the results from the validation experiments, which included estimating the model parameters, producing hindcasts and carrying out verification. To produce hindcasts, the QPFs from the historical record were used. The hindcasting experiments were performed with the following settings. The size of the moving window for the QPF and MAP data in parameter estimation was 91 days (see Section 4). Because of the limited amount of archived data for the study basins (see Table 2), sub-daily-scale and diurnal variations in the QPF and MAP were ignored. The statistical parameters were hence estimated from the data pooled over all four 6-h sub-periods within each 24-h period. The marginal distributions were estimated for all basins using the nonparametric technique with Gaussian kernel smoothing and plug-in optimal bandwidth selection (see Section 4). The density functions in Eq. (16) were modeled by the 3-parameter Pearson type III distribution as indicated in Section 4.

The validation experiments were designed to answer the following questions: (Q1) How good is the meta-Gaussian model for generating precipitation ensembles for different regions, seasons, and amounts? (Q2) Is the proposed procedure for the mixed-type meta-Gaussian model capable of generating reliable and skillful precipitation ensembles under minimum sampling uncertainty (i.e. in dependent validation) and in the presence of sampling uncertainty (i.e. in independent validation)? (Q3) How much data does one need for parameter estimation to produce reliable precip-

itation ensembles that capture the skill in the single-valued forecast? (Q4) How much do regional variations in the QPF skill affect the quality of the ensembles generated by the proposed procedure? (Q5) How much does the generalized model with parameter optimization improve the quality of precipitation ensembles? Below, Sections 5.1–5.3 address Q1, Q2 through Q4, and Q5, respectively.

5.1. Validation of the meta-Gaussian model

If the bivariate normality condition holds, $D(x, y)$ in Eq. (2) is equal to its meta-Gaussian distribution. Herr and Krzysztofowicz (2005) studied the bivariate normality condition for 24-h precipitation amounts at a number of rain gauge locations in two river basins in the Appalachian Mountains. Their findings are mixed in that in most of the cases the normality hypothesis was not rejected for one basin but rejected for the other basin. Thus, a natural question arises as to whether one can still obtain satisfactory results from the meta-Gaussian model when the bivariate normality is not met. This question may be answered by evaluating the results of the meta-Gaussian modeling. To that end, rather than testing bivariate normality, here we assess the goodness of fit of the meta-Gaussian model of $D(x, y)$ against the empirical joint distributions of $(X \leqslant x \mid X > 0, Y > 0)$ and $(Y \leqslant y \mid X > 0, Y > 0)$. A good fit between the modeled and the empirical joint distributions is a necessary condition for reliable ensembles. Two-dimensional Kolmogorov–Smirnov tests (Peacock, 1983; Fasano and Franceschini, 1987) are available for evaluating goodness of fit between two bivariate distributions. Here, as a simpler alternative, we evaluate goodness of fit between the modeled and the empirical distributions by comparing their conditional distributions given QPF. This approach allows us to see directly how the collective distributional quality of the ensembles may vary as the conditioning QPF varies.

To evaluate the goodness of fit, we carried out the following hindcasting experiment. The data were pooled into four seasons: winter (DJF), spring (MAM), summer (JJA), and fall (SON). A sample was drawn from the modeled distribution for each season in the hindcast period for comparison with the empirical joint distribution for that season. Due to the limited sample size, we partitioned the range of QPF into two sub-ranges separated at the median to obtain conditional distributions. Fig. 4 shows the results for HUNP1 for Day 1 obtained from dependent validation (see Section 5.2), which are based on 6 years of QPF and MAP data (see Table 2). The ensemble size, which matches the number of historical years in the long-term MAP record used in the Schaake Shuffle, is 51 members. The figure may be summarized as follows. For winter, the empirical conditional distribution matches that of the modeled distribution very well for QPF greater than or equal to the median, but lack of fit is evident for QPF less than the median. For summer, the fit is good for both cases. The results for spring and fall (not shown) are generally similar to those for winter but inferior for QPF greater than or equal to the median. The results for TIFM7 and NFDC1 for Day 1 are similar to those for HUNP1 and are not shown. It is interesting to note that the fit appears to be good when QPF is greater than or equal to the median for winter for all three basins whereas no clear patterns are apparent for the other seasons. Reduced nonstationarity in skill during winter, as suggested by the relatively uniform correlation coefficient (see Fig. 3), may be a contributing factor to the above observation.

5.2. Validation of ensembles from the mixed-type meta-Gaussian Model (Method 2)

To assess the quality of precipitation ensemble hindcasts generated by the proposed procedure, we examined reliability diagrams, relative operating characteristic curves (ROC) and mean CRPS.
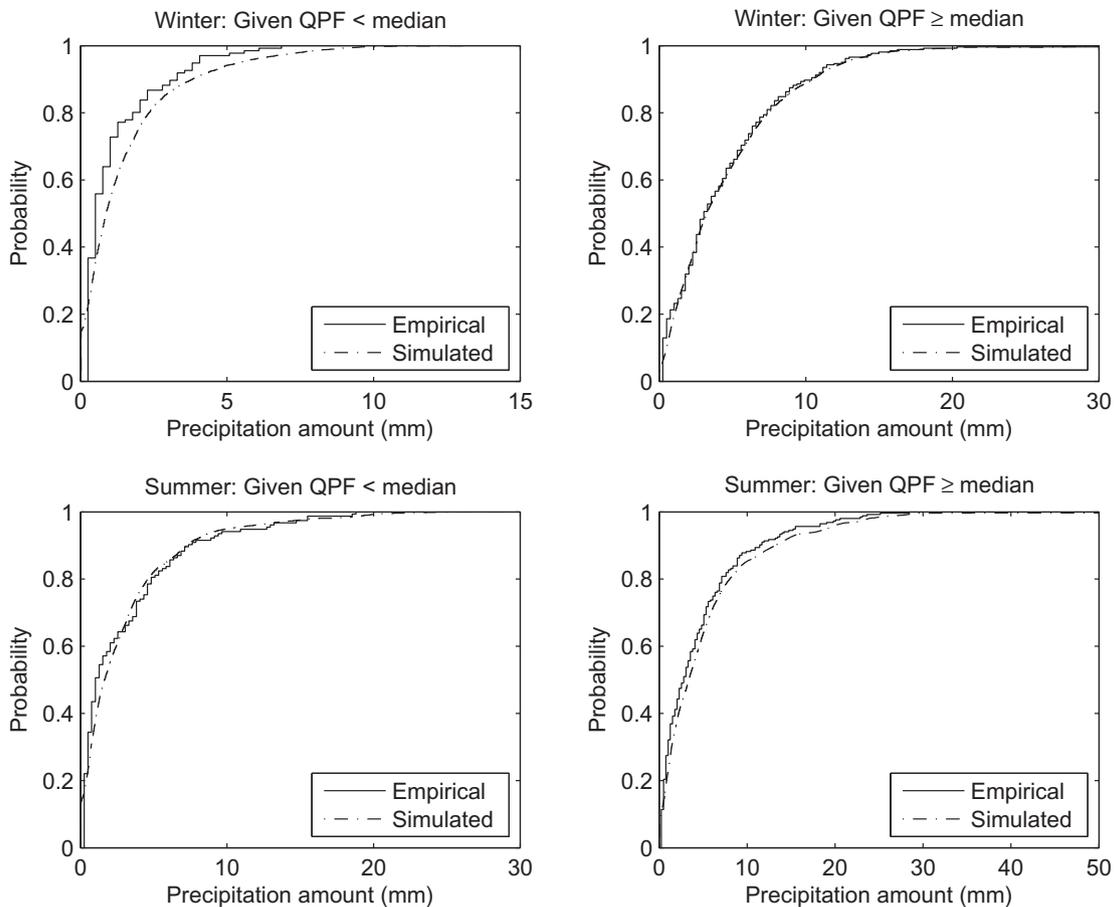
Fig. 4. Empirical conditional CDFs of positive 6-h MAP given positive 6-h QPF (Day 1) against those of positive 6-h MAP simulated from the mixed-type meta-Gaussian distribution given the 6-h QPF. The results are for HUNP1 for the winter and summer months.

The latter two are particularly useful as they allow quantitative quality assessment of ensemble forecasts using the single-valued forecasts. The ROC curve connects, at varying levels of predicted probability, the paired false alarm (FAR) and hit rates (HR) associated with the ensemble forecasts, which may be checked against those of the single-valued forecasts. The mean CRPS of ensemble forecasts may be checked against the mean absolute error of single-valued forecasts, as the mean CRPS reduces to the mean absolute error for single-valued forecasts.

We present the results from dependent validation (DV) and leave-one-year-out cross validation (CV). In DV, the same data set used in parameter estimation was also used in hindcasting. Hence, DV is a check on the goodness of the fit of the model to the observed conditions without regard to that of the future conditions. Both DV and CV were carried out using the entire QPF data and the corresponding MAP data described in Section 4.

Figs. 5–8 show the reliability diagrams for ensemble hindcasts of 6-h precipitation for HUNP1 and TIFM7. In each figure, there are four reliability diagrams, corresponding to four different thresholds of 0, 2.54, 3.18 and 6.35 mm (0, 1/10, 1/8 and 1/4 in). We could not use larger thresholds because of inadequate sample size. As such, we are not able to make a strong statement about the reliability of 6-h precipitation ensembles generated by the proposed technique for thresholds exceeding 6.35 mm. In these figures, the DV results are presented together with the CV results for ease of comparison. Also shown are the 95% confidence intervals estimated via the percentile method of bootstrapping (Bröcker and Smith, 2007; Chernick, 2008). Note that the confidence intervals at the lowest forecast probabilities may be too small to be

visually discernible. In Fig. 6, for the 6.35 mm threshold, two points appear without accompanying confidence intervals due to lack of sample size (and hence no sampling variability in bootstrapping). The reliability diagram for the threshold of 0 mm indicates how reliable the ensembles are at predicting the probability of precipitation (PoP). These figures indicate reliable ensemble prediction of PoP for Days 1 and 2 in both DV and CV. For other thresholds, the precipitation ensembles are generally reliable, within the sampling uncertainty bounds. Note that the difference between the DV and CV results is relatively small. It suggests that only several years' worth of data may be required for parameter estimation to realize in real-time applications the level of performance seen in DV.

Figs. 9–12 show the ROC diagrams for ensemble hindcasts for HUNP1 and TIFM7. In these figures, we define the event as 6-h precipitation exceeding the threshold. In each figure, there are four ROC diagrams corresponding to four different thresholds of 0, 2.54, 6.35 and 12.7 mm. The DV and CV results are shown together for ease of comparison. The area under the ROC curve (AUC) is also given as obtained by the *perfcurve* function in the MATLAB Statistics Toolbox (MATLAB Statistics Toolbox™ 7 User's Guide). For the positive thresholds, the ensembles associated with zero QPF and zero MAP were excluded from the calculation so that we may isolate the discriminatory skill in non-dry conditions. The HUNP1 ensembles show better discrimination than the TIFM7, in agreement with the relative magnitude in correlation for the two basins seen in Fig. 3. Note that the DV and CV results are very similar for ROC as well, which suggests that a historical archive of about 6 years may be enough to realize in the prediction mode the level of discrimination attainable in DV.
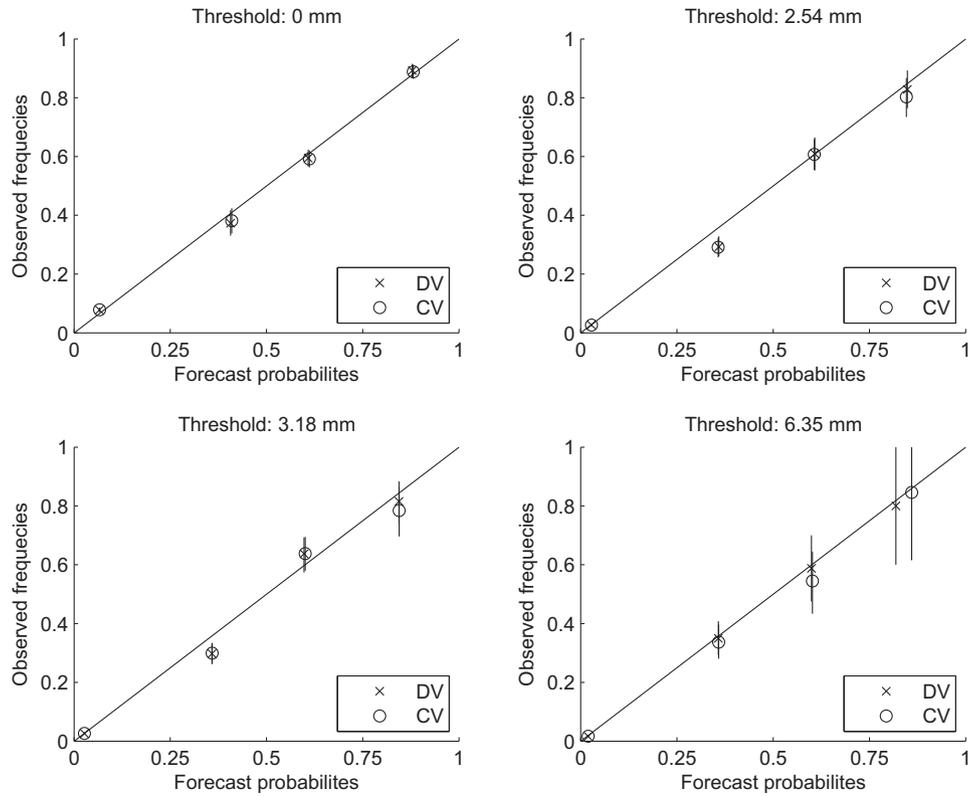
Fig. 5. Reliability diagrams for ensemble hindcasts from Method 2 of 6-h precipitation for all four 6-h periods within Day 1 for HUNP1. The vertical bars denote 95% confidence interval. Note that confidence intervals at the lowest forecast probability are indiscernible.
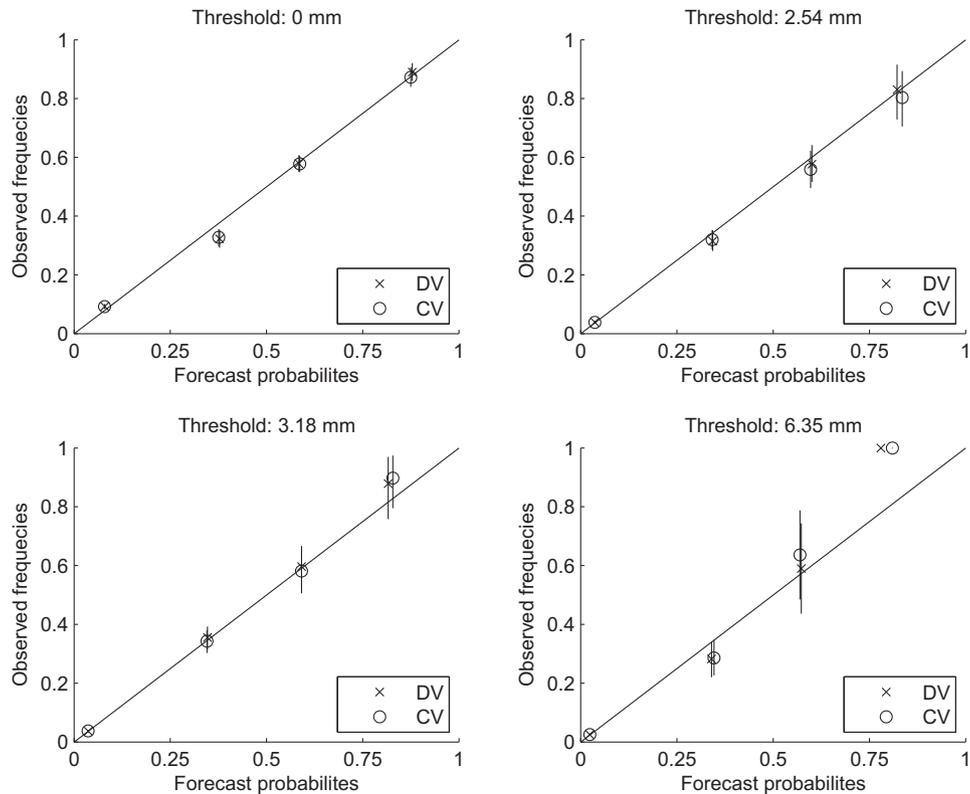

Fig. 6. Same as Fig. 5 but for Day 2. For the threshold of 6.35 mm, two points appear without accompanying confidence intervals due to no sampling variability in bootstrapping.
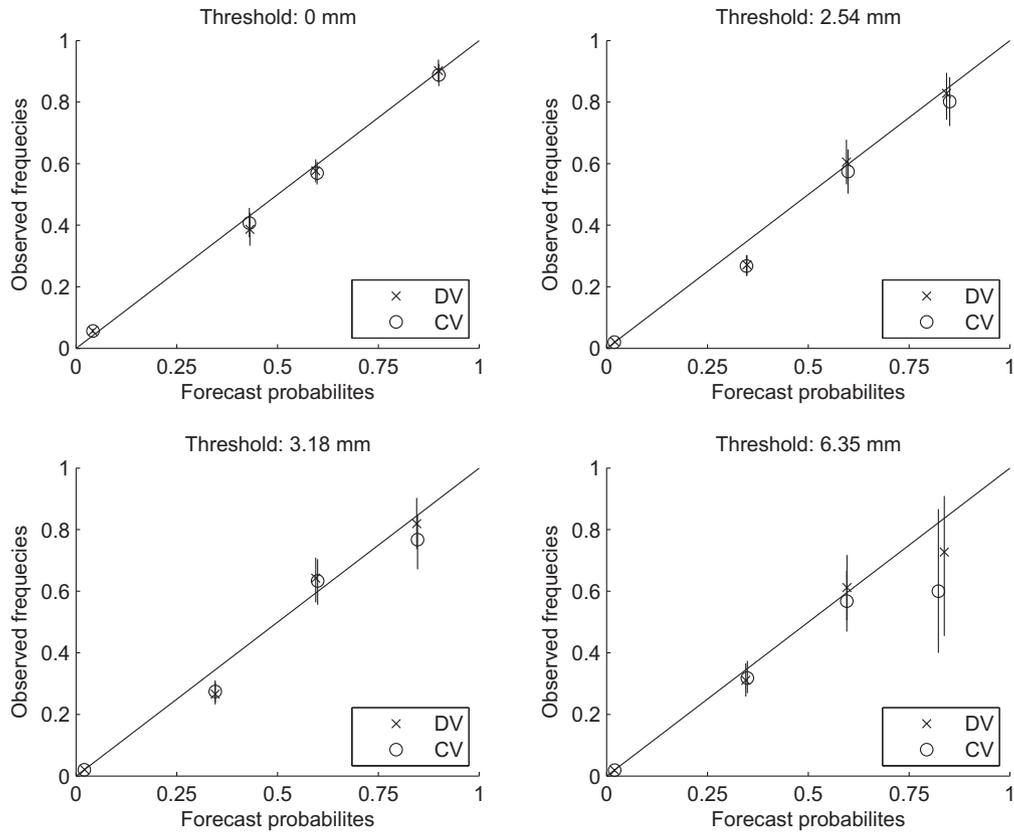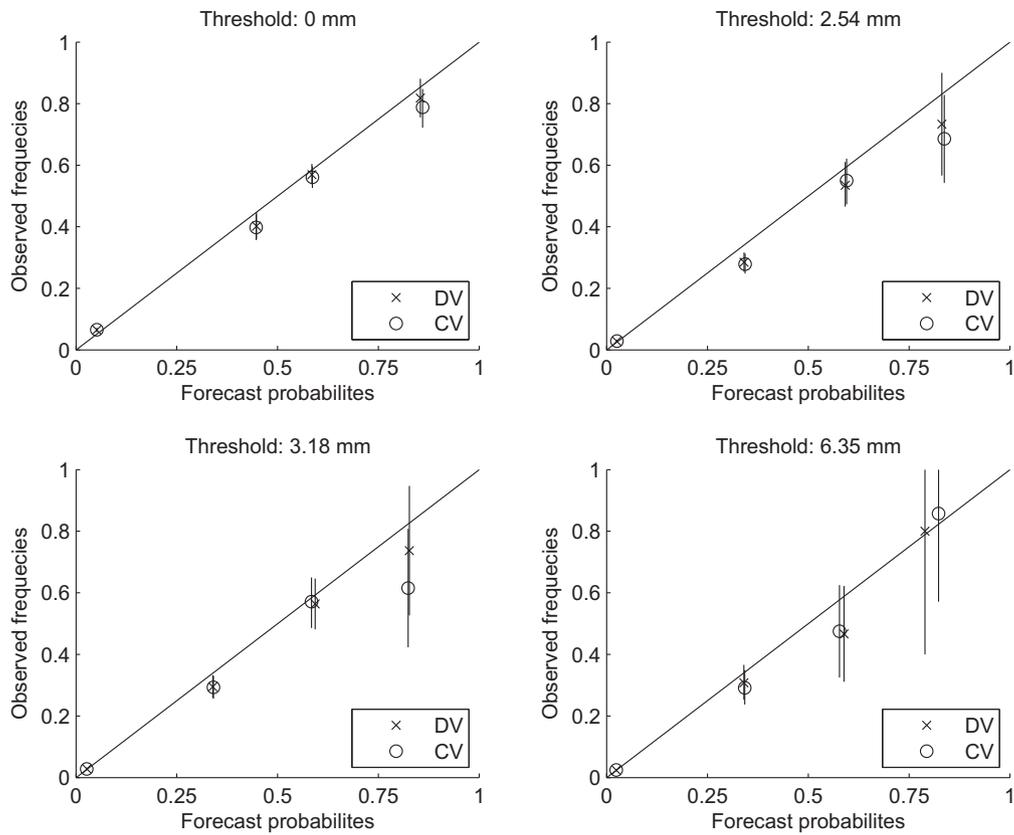
Fig. 7. Same as Fig. 5 but for TIFM7.
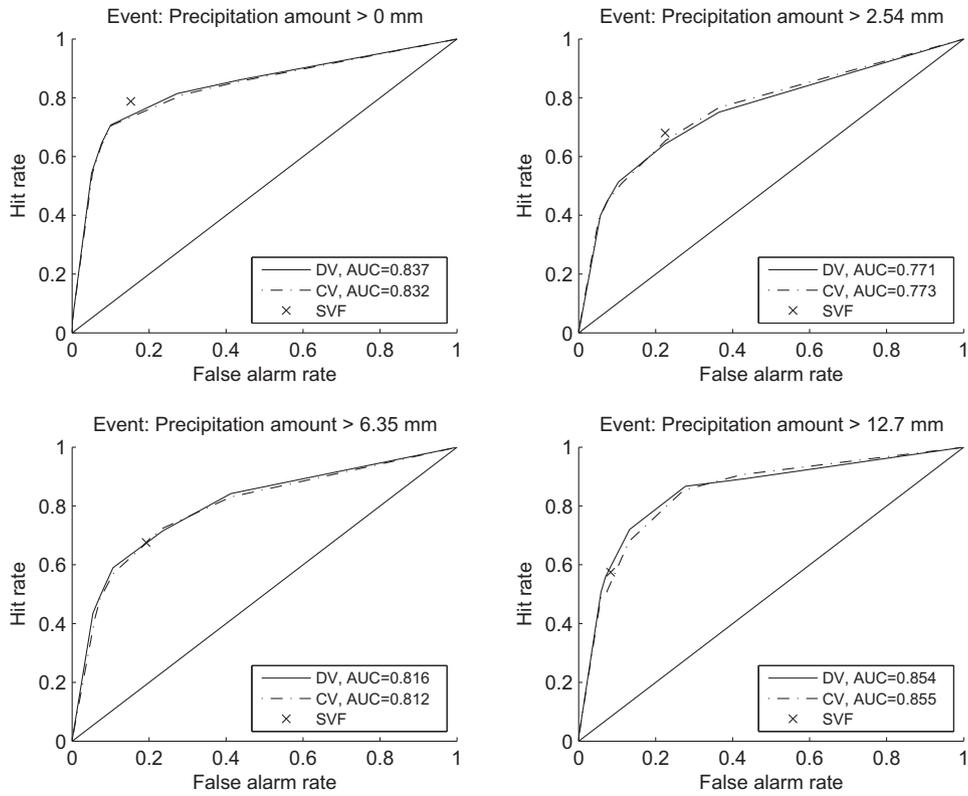

Fig. 8. Same as Fig. 7 but for Day 2.

Fig. 9. ROC diagrams for ensemble hindcasts from Method 2 and single-valued QPFs of 6-h precipitation for all four 6-h periods in Day 1 for dependent validation (DV), cross validation (CV), and single-valued QPF (SVF). The results are for HUNP1. For the event threshold 0, all ensemble forecasts were used. For positive event thresholds, ensemble hindcasts associated with QPF = 0 and MAP = 0 were excluded.
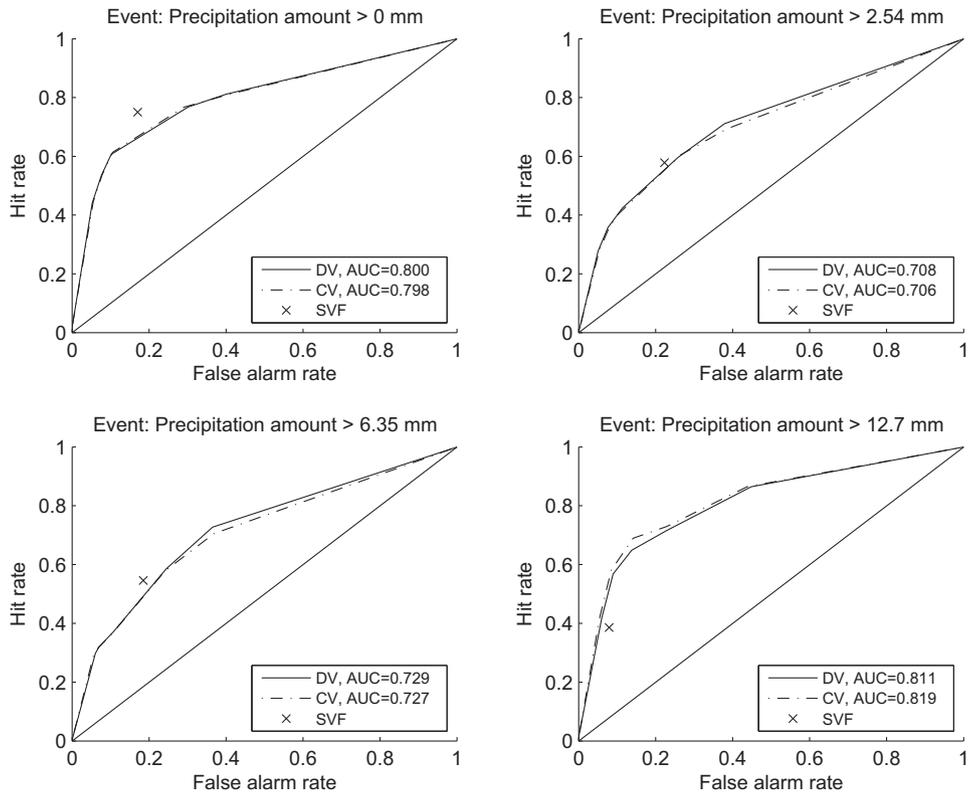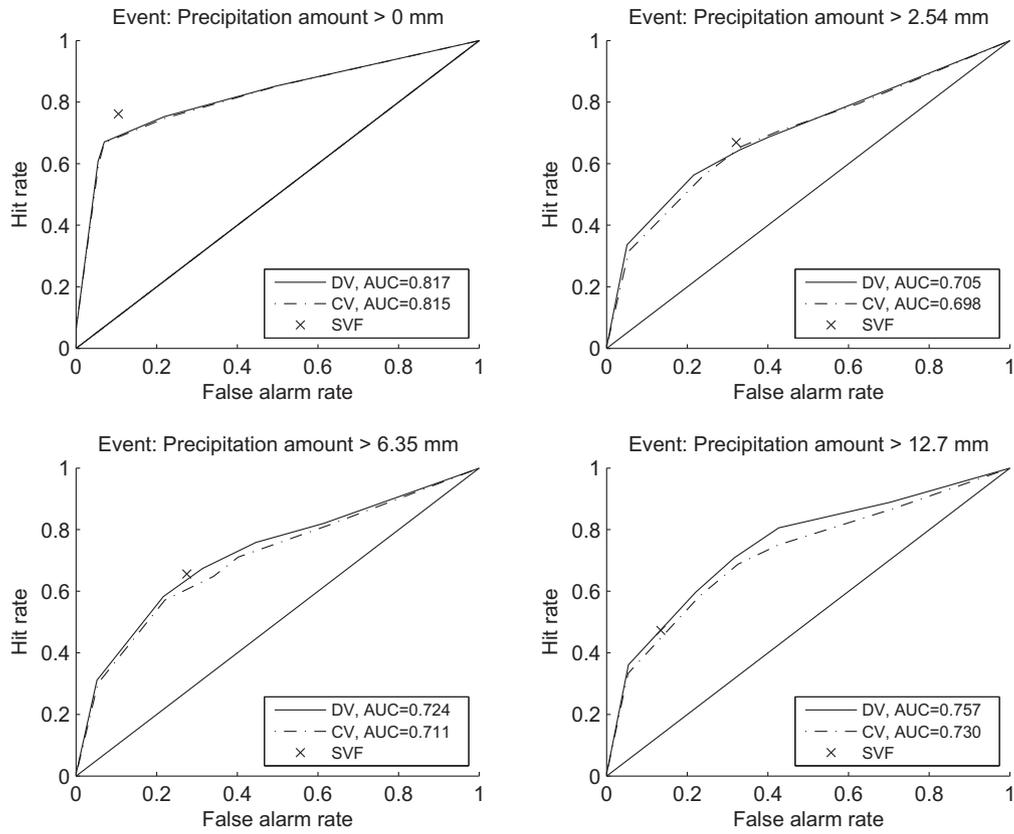

Fig. 10. Same as Fig. 9 but for Day 2.
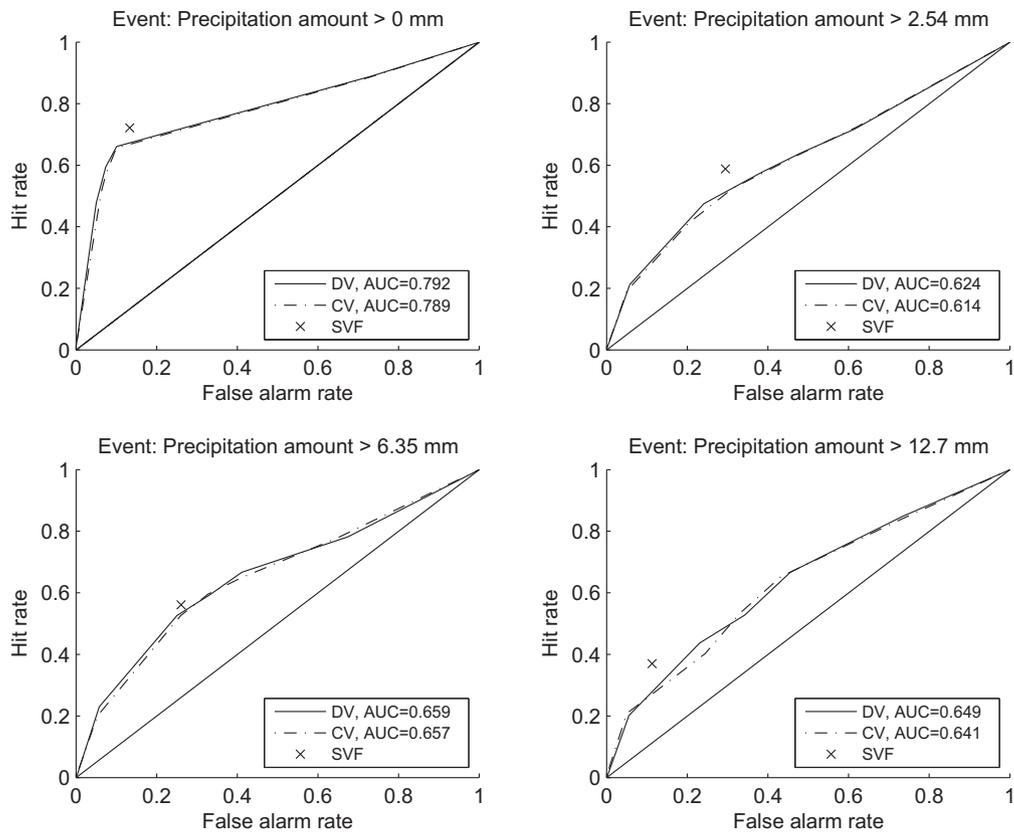
Fig. 11. Same as Fig. 9 but for TIFM7.

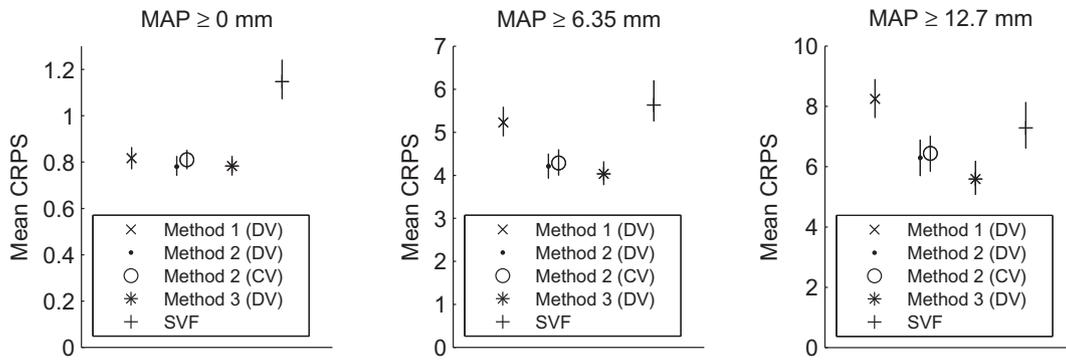**Fig. 12.** Same as Fig. 11 but for Day 2.

Fig. 13. Mean CRPS for ensemble hindcasts of 6-h precipitation for all four 6-h periods in Day 1 from October through May. The results are for NFDC1 with upper and lower areas combined. The vertical bars denote the 95% confidence intervals. The mean CRPS values are conditioned on MAP $\geqslant 0$, 6.35 and 12.7 (mm).

Figs. 9–12 also show the HR and FAR of the single-value QPF used to condition the modeled bivariate distribution. By comparing the HR and FAR of the single-valued forecast with the ROC curves of the ensemble forecast, we may assess how well the ensemble forecast may capture the discriminatory skill in the QPF. A ROC curve that lies on or above and to the left of the (HR, FAR) of the single-valued QPF is an indication that the ensemble forecast has approximately the same level of discrimination as the single-valued forecast for that particular definition of the event. These figures show that the proposed procedure is generally successful in capturing the discriminatory skill in the conditioning QPF for the positive event thresholds. For the no-precipitation threshold, however, the ensemble forecasts have smaller discriminatory skill than the single-valued forecasts. It suggests that modeling of precipitation intermittency may need further improvement. For TIFM7 for Day 2, the ensembles underperform the QPFs noticeably. It reflects the difficulty of generating ensembles using the meta-Gaussian model when the single-valued forecast has very limited skill.

5.3. Comparison of the methods

In this subsection, we compare Schaake et al. (2007) (Method 1), the mixed-type meta-Gaussian model (Method 2), and the mixed-type bivariate distribution with the generalized linear model (Method 3) described in Section 2. These methods differ mainly in two aspects: precipitation intermittency modeling and dependence structure modeling. Differences in the first aspect can be described as follows. Method 1 accounts for precipitation intermittency implicitly by modeling each of the marginal distributions as a convex combination of continuous distributions, resulting in continuous conditional distributions for sampling ensemble members. The positive ensemble members less than a threshold value are set to zero. Both Methods 2 and 3 model precipitation intermittency explicitly, giving discrete–continuous conditional distributions for sampling ensemble members. Differences in the second aspect are the following. For Method 1, the dependence structure of the transformed space is still bivariate standard normal, but only partially specified. Thus, the correlation coefficient of the dependence structure is estimated by a weighted average of the Pearson product-moment correlation coefficient of the untransformed variates (including zeros) and that of the transformed variates that best fit p_{00}, p_{10}, p_{01}, and p_{11} in Eqs. (3)–(6). Method 2 simply uses Pearson's correlation coefficient for the dependence structure of the continuous–continuous part of the bivariate distribution. Method 3 uses an extended linear model for the transformed space, providing a parameter for adjusting the dependence structure to yield optimal ensembles under the criterion of choice.

We use CNRFC's NFDC1 in the comparison. The time period covered is the wet season of October through May in the entire period of record (see Table 2). The comparison is made using the mean CRPS, the reliability diagrams, and the ROC curves. These results are based on the same data set. As such, while there exist in some cases significant sampling uncertainties, as shown by the relatively large confidence intervals, one may still draw firm conclusions from the mean CRPS on the relative performance among the three methods, at least for those events captured in the data set. Note that the period of record is several years long, and includes a reasonably diverse set of events. Figs. 13 and 14 show the mean CRPS for the three methods for 6-h precipitation ensemble forecasts for Days 1 and 2, respectively. There are three plots in each figure, corresponding to the mean CRPS conditional on MAP greater than or equal to 0, 6.35 and 12.7 mm. Note the large reduction in the mean CRPS by Methods 2 and 3 over Meth-
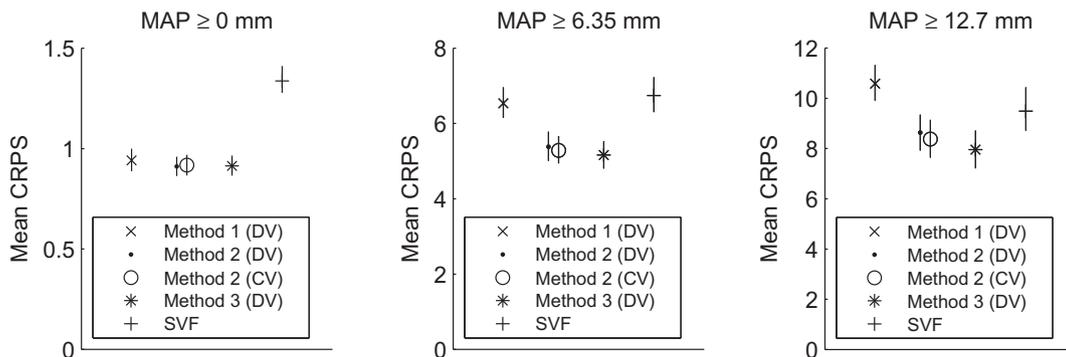

Fig. 14. Same as Fig. 13 but for Day 2.

od 1. Method 3 moderately improves over Method 2 for the conditioning thresholds of MAP greater than or equal 6.35 and 12.7 mm, with respective reductions of 4.1% and 10% for Day 1, and 3.8% and 7.1% for Day 2. The results of Method 3 shown in Figs. 13 and 14 were obtained using the mean CRPS as the objective function and as such provide performance bounds with respect to the mean CRPS for the methods examined in this work. As noted earlier, the mean CRPS reduces to the mean absolute error (MAE) for single-valued forecasts, which allows a comparative check on the overall quality of the ensemble forecasts against that
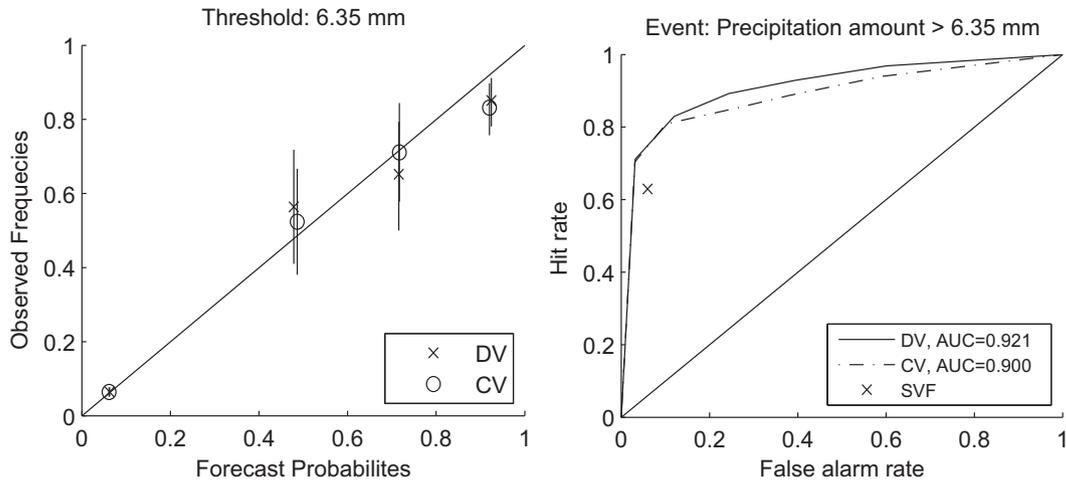

Fig. 15. Left: Reliability diagrams for ensemble hindcasts from Method 3 in dependent validation (DV) and in leave-one year-out cross validation (CV) of 6-h precipitation for all four 6-h periods in Day 1. The results are for NFDC1 from October through May. The vertical bars denote the 95% confidence intervals. Right: ROC curves for ensemble hindcasts from Method 3 in DV and CV of 6-h precipitation for all four 6-h periods in Day 1. Also shown are the false alarm rate and hit rate of the conditioning single-valued QPF (SVF). The results are for NFDC1 from October through May.
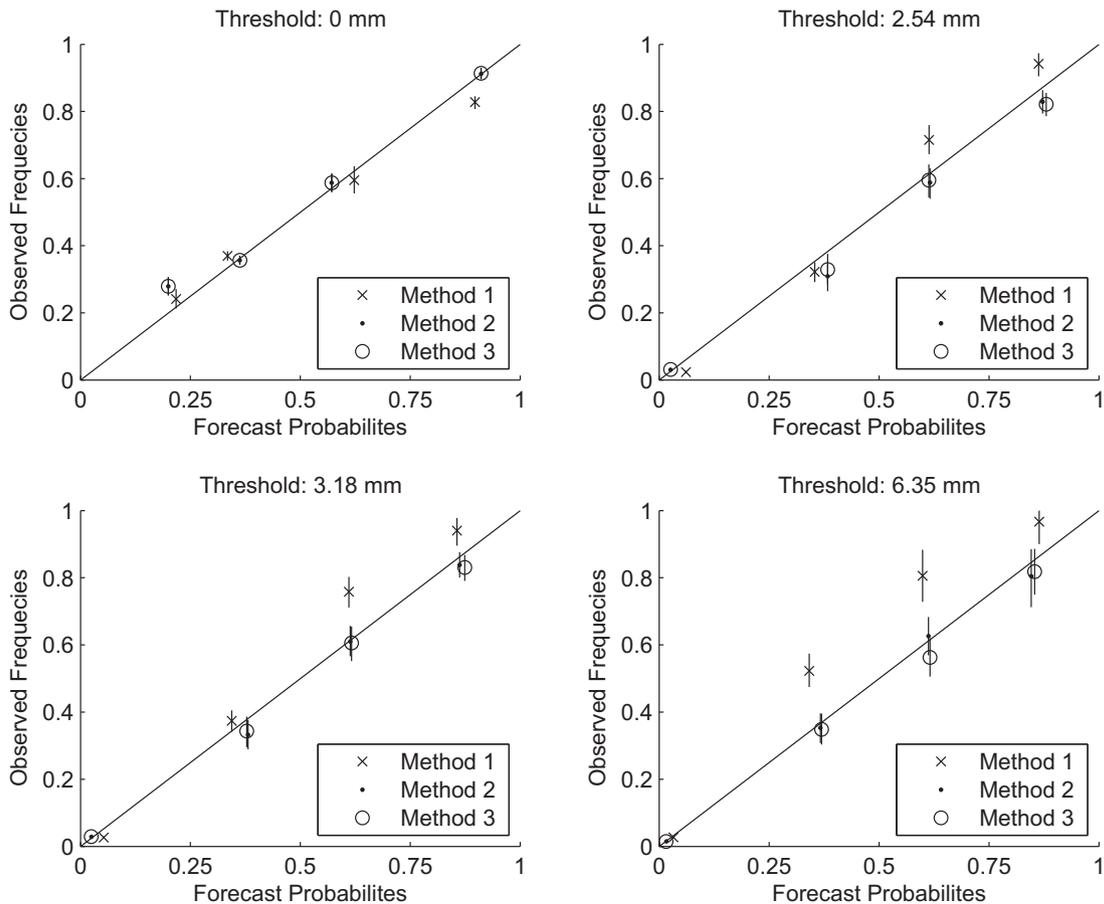

Fig. 16. Reliability diagrams for ensemble hindcasts in DV of 6-h precipitation for all four 6-h periods in Day 1. The results are for NFDC1 from October through May. The vertical bars denote the 95% confidence intervals.

of the single-valued forecasts. Some caution is necessary in such an assessment, however, in that ensemble forecasts have an inherent advantage over single-valued forecasts of similar skill owing to the smoothing effects. Accordingly, we would like to see the mean CRPS of the ensemble forecasts being clearly smaller than the MAE of the single-valued forecasts. Note in Figs. 13 and 14 that the above is indeed achieved by Methods 2 and 3. Fig. 15 shows the reliability and ROC diagrams of hindcast ensembles from Method 3 in DV and CV for the selected threshold and event definition. For the results given in Fig. 15, we used RMSE of the observed and the corresponding ensemble 60th percentile as the objective function (all other results from Method 3 presented in this subsection were produced with the mean CRPS as the objective function). Figs. 13–15 show that, for both Methods 2 and 3, the difference in performance between DV and CV is rather small, an indication that several years' worth of data may be enough for parameter estimation (see also Section 5.2). The results for Method 1 are similar and are not shown.

Figs. 16 and 17 show the reliability diagrams for the three methods. They are based on DV. It is seen that the ensembles from Methods 2 and 3 are considerably more reliable than those from Method 1. The reliability results of Method 1 indicate under-forecasting, which may be attributed to poor estimation of the correlation coefficient between the transformed variates and lack of fit of the parametric distribution model, which is Weibull, to the marginal probability distributions in the NQT. The reliability results of Method 2 and Method 3 are similar. Note that the corresponding DV results of Method 3 for threshold 6.35 mm shown in Figs. 15 and 16 are slightly different due to different objective functions used in the optimization, as indicated earlier in the discussion on

Fig. 15. The ROC results are similar among all three methods and are not shown. The above results indicate that, as expected, the reduction in the mean CRPS by Methods 2 and 3 over Method 1 come mostly from improved reliability.

6. Conclusions and future research recommendations

Two statistical procedures (Methods 2 and 3 described in Section 2) for generating precipitation ensemble forecasts from single-valued quantitative precipitation forecasts (QPF) were described and evaluated using observed Mean Areal Precipitation (MAP) and QPF data for three basins in the service areas of the Arkansas-Red Basin (AB-), California-Nevada (CN-), and Middle Atlantic (MA-) River Forecast Centers (RFC) of the National Weather Service (NWS). Both procedures are based on modeling the relationship between the MAP and QPF by a mixed-type bivariate distribution, in which the relationship between the positive MAP and positive QPF is modeled as meta-Gaussian and generalized meta-Gaussian.

The main conclusions are as follows:

- The goodness of the meta-Gaussian model, as assessed by comparing the modeled cumulative distribution functions (CDF) with the empirical, varies with the time of the year and the magnitude of the conditioning single-valued QPF. For the study basins and for Method 2, the goodness of fit is the best for larger-than-median QPFs in the wet season, for which the single-valued QPF has the largest skill.
- The validation results show that, overall, the precipitation ensembles generated by the proposed procedures (Methods 2
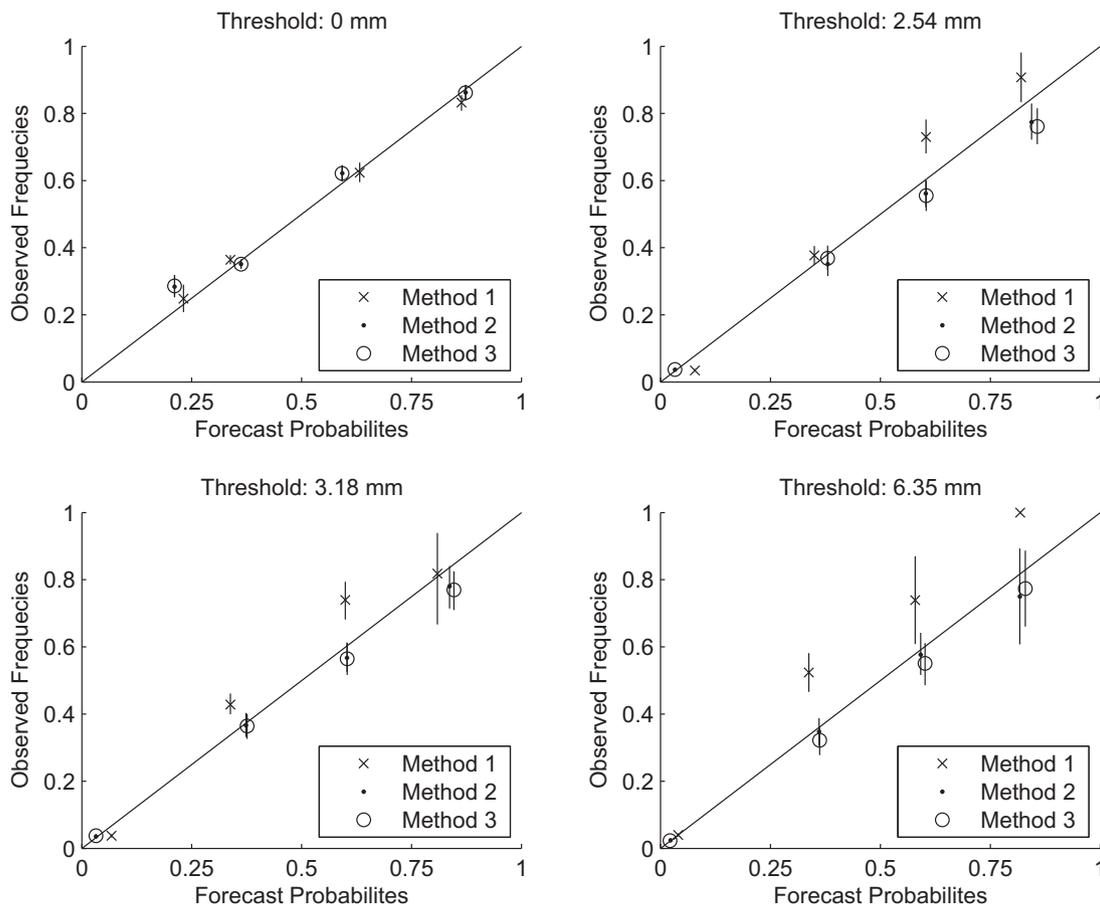

Fig. 17. Same as Fig. 16 but for Day 2.

and 3) are reliable. The proposed procedures are generally successful in capturing the discriminatory skill in the single-valued QPF when precipitation is predicted, but less so in capturing the skill in discriminating precipitation vs. no precipitation and when the single-valued forecast has very limited skill (i.e. the Day-2 forecast in the southern plains). Additional work is necessary to address these.

- The similarity of the dependent and cross validation results suggests that a dataset of several years in length is sufficient to estimate the model parameters with acceptable sampling uncertainty. This is an extremely important consideration given the reality that the period of record for single-valued QPFs is rather short at most RFCs.
- Comparative evaluation indicates that the proposed procedures (Methods 2 and 3) provide considerable improvement over the initial version (Method 1). Large reductions in the mean CRPS for Methods 2 and 3 over Method 1 are achieved (Figs. 13 and 14); in particular, the reductions for Method 3 over Method 1 for Day 1 are 3.4%, 22%, and 30%, respectively, for the three conditioning thresholds of MAP greater than or equal to 0, 6.35 and 12.7 mm. The evaluation also indicates that the improvement comes largely from improved reliability.
- In hydrologic ensemble forecasting, we want to provide forecasts for as long a prediction horizon as possible. The forecasts of the frozen GFS (Hamill et al., 2008) have lead-times up to 14 days, much longer than the RFC/HPC issued single-valued forecasts. The ensemble mean of the frozen GFS reforecast may be used with the procedures described in this work to produce forcing ensembles for hydrologic models. In this work, assessment of reliability was limited to small to medium amounts of precipitation due to the limited period of record for QPF. With data archive of more than 20 years available for the frozen GFS reforecast, we may be able to do the assessment for larger amounts for ensembles produced from the ensemble mean of the frozen GFS. We plan to use the GFS reforecast dataset (Hamill et al., 2008) to evaluate the quality of ensembles for larger amounts.
- To minimize the impact of temporal nonstationarity in MAP, QPF, and their correlation in model parameter estimation, better methods for temporally pooling data need to be developed. Wavelet analysis can potentially be used to characterize the temporal nonstationarity. The standardization and de-standardization procedures described in Krzysztofowicz and Evans (2008) might be used to treat nonstationarity in the first two moments of the data.
- Several methods have been developed in modeling precipitation amounts and spatial correlation of precipitation fields (Bardossy and Plate, 1992; Berrocal et al., 2008; Makhnin and McAllister, 2009). Bardossy and Plate (1992) used a power-transformed normally distributed variate truncated at zero to describe both intermittency and positive precipitation. Berrocal et al. (2008) used a two-stage approach, in which precipitation occurrence is considered first, positive precipitation accumulation is then modeled conditionally on precipitation occurrence using a Gamma distribution. These methods offer diverse approaches to precipitation field modeling. It would be very interesting to see how well they may perform in tackling the problem considered in this work.

Appendix A. The bivariate meta-Gaussian distribution model

Consider the joint cumulative distribution function (CDF) of two continuous variates X and Y:

$$F(x, y) \equiv P(X \leqslant x, Y \leqslant y). \tag{A1}$$

Denote the CDF of X by $F_X(x)$ and that of Y by $F_Y(y)$. Assume that $F_X(x)$ and $F_Y(y)$ are strictly increasing. Let Q denote the standard normal distribution function and Q^{-1} denote its inverse. Applying the normal quantile transform (NQT) to X and Y, respectively, we obtain two standard normal variates Z and W:

$$Z = Q^{-1}(F_X(X)), \tag{A2a}$$

$$W = Q^{-1}(F_Y(Y)). \tag{A2b}$$

Then $P(X \leqslant x, Y \leqslant y) = P(Z \leqslant z, W \leqslant w)$, where $z = Q^{-1}(F_X(x))$ and $w = Q^{-1}(F_Y(y))$. The joint distribution of Z and W is not necessarily bivariate standard normal. Let B denote the bivariate standard normal distribution function. Define:

$$H(x, y; \rho) \equiv B(Q^{-1}(F_X(x)), Q^{-1}(F_Y(y)); \rho), \tag{A3}$$

where ρ denotes the Pearson product-moment correlation coefficient between Z and W. If (Z, W) is standard normal, that is:

$$P(Z \leqslant z, W \leqslant w) = B(z, w; \rho), \tag{A4}$$

then $F(x, y) = H(x, y; \rho)$ and $H(x, y; \rho)$ is called bivariate meta-Gaussian distribution of X and Y (Kelly and Krzysztofowicz, 1997). If (Z, W) is not standard normal, what one can hope is that $F(x, y)$ is well approximated by $H(x, y; \rho)$:

$$F(x, y) \approx H(x, y; \rho). \tag{A5}$$

For the meta-Gaussian distribution of X and Y, the conditional distribution of Y given $X = x$ is given by (Kelly and Krzysztofowicz, 1997):

$$H_{Y|X}(y|x) = Q\left(\frac{Q^{-1}(F_Y(y)) - \rho Q^{-1}(F_X(x))}{(1 - \rho^2)^{1/2}}\right). \tag{A6}$$

For any p such that $0 < p < 1$, the p-probability conditional quantile of Y given $X = x$ is a value $y_{p|x}$ of Y such that $p = H_{Y|X}(y_{p|x}|x)$. From (A6), we have for $y_{p|x}$:

$$Y_{p|x} = F_Y^{-1}(Q(\rho Q^{-1}(F_X(x)) + (1 - \rho^2)^{1/2} Q^{-1}(p))). \tag{A7}$$

Appendix B. Extension of the bivariate meta-Gaussian model

As in Appendix A, let X and Y be continuous variates with strictly increasing CDF. Let Z and U be the variates defined in Eq. (17) in Section 2. The variates X and Z are related as in Eq. (A2a). Let $\tilde{Q}$ denote the CDF of U and $\tilde{Q}^{-1}$ the inverse of $\tilde{Q}$. A one-to-one correspondence between Y and U is offered by $\tilde{Q}(u) = F_Y(y)$. Denote the joint distribution function of Z and U by $\tilde{B}(z, u; \gamma)$. The joint distribution $\tilde{B}(z, u; \gamma)$ is bivariate normal and the conditional distribution of U given $Z = z$ is $N(bz, 1 + b^2 - 2b\rho)$, or expressed in the standard normal distribution function Q as:

$$P(U \leqslant u | Z = z) = Q((u - bz)/(1 + b^2 - 2b\rho)^{1/2}). \tag{B1}$$

Writing $\widetilde{B}(z, u; \gamma)$ as a function of x and y and denoting it as $\widetilde{H}(x, y)$, we have analogously to Eq. (A3):

$$\widetilde{H}(x, y) \equiv \widetilde{B}(Q^{-1}(F_X(x)), \tilde{Q}^{-1}(F_Y(y)); \gamma). \tag{B2}$$

We can show that $\widetilde{H}(x, y)$ is a bivariate CDF using, e.g., Theorem 6 in Rohatgi (1976, p.135). For this CDF, from its definition given by Eq. (B2), we can derive its marginal CDFs, which are $F_X(x)$ and $F_Y(y)$, and its conditional distribution of Y given $X = x$, which is:

$$\widetilde{H}_{Y|X}(y|x) = Q\left(\frac{\tilde{Q}^{-1}(F_Y(y)) - bQ^{-1}(F_X(x))}{(1 + b^2 - 2b\rho)^{1/2}}\right). \tag{B3}$$

An equation analogous to (A7) can be easily derived from (B3).

Appendix C. Implicit modeling of precipitation intermittency

Schaake et al. (2007) accounts for precipitation intermittency implicitly in precipitation ensemble generation. Here, we give an alternative formulation for this implicit approach in the framework of the bivariate meta-Gaussian model by writing each of the marginal distributions of $F(x, y)$ defined in Eq. (1) as a convex combination of continuous distributions (Law and Kelton, 2000, p. 449). In this formulation we model the precipitation amount as a continuous distribution with a stretch of its CDF unspecified for the precipitation amount in the interval of 0 to a threshold value. The probability mass corresponding to that stretch is the sum of the probability of zero precipitation and that of the precipitation amount that may be negligible in the modeling.

Now, let X denote forecast precipitation amount accumulated during a given time interval and Y denote the corresponding observed precipitation amount. Denote the joint distribution function of X and Y by Eq. (1). Let

$$b_{X1} \equiv P(X < x_t), \quad b_{X2} \equiv 1 - b_{X1},$$

$$b_{Y1} \equiv P(Y < y_t), \quad b_{Y2} \equiv 1 - b_{Y1},$$

$$F_{X1}(x) \equiv P(X \leqslant x | X < x_t),$$

$$F_{X2}(x) \equiv P(X \leqslant x | X \geqslant x_t),$$

$$F_{Y1}(y) \equiv P(Y \leqslant y | Y < y_t),$$

$$F_{Y2}(y) \equiv P(Y \leqslant y | Y \geqslant y_t),$$

where x_t and y_t are positive threshold values. The CDF of X and Y can be expressed, respectively, as:

$$F_X(x) = b_{X1}F_{X1}(x) + b_{X2}F_{X2}(x),$$

$$F_Y(y) = b_{Y1}F_{Y1}(y) + b_{Y2}F_{Y2}(y).$$

For the case of $y \geqslant y_t$, for $b_{Y2} \neq 0$, Eq. (A7) becomes:

$$Y_{p|x} = F_{Y2}^{-1}\left(\frac{Q(\rho Q^{-1}(b_{X1}F_{X1}(x)) + (1 - \rho^2)^{1/2}Q^{-1}(p)) - b_{Y1}}{b_{Y2}}\right),$$
$$\text{if } x < x_t, \tag{C1a}$$

$$Y_{p|x} = F_{Y2}^{-1}\left(\frac{Q(\rho Q^{-1}(b_{X1} + b_{X2}F_{X2}(x)) + (1 - \rho^2)^{1/2}Q^{-1}(p)) - b_{Y1}}{b_{Y2}}\right),$$
$$\text{if } x \geqslant x_t. \tag{C1b}$$

Note that the right-hand side of (C1b) depends on neither $F_{X1}(x)$ nor $F_{Y1}(y)$, which means that the modeling of $F_{X1}(x)$ and $F_{Y1}(y)$ does not affect $y_{p|x}$ in (C1b). If $F_{X1}(x)$ and $F_{Y1}(y)$ are not specified, the apparatus of estimating ρ, Pearson's correlation coefficient

between the transformed variates, through (A2a) and (A2b) is no longer available. One way of estimating ρ may be offered by expressions (A3) and (A5). For example, for a selected pair of values $x_a \geqslant x_t$ and $y_a \geqslant y_t$, we may find ρ numerically via:

$$\min_{\rho} |F(x_a, y_a) - B(Q^{-1}(b_{X1} + b_{X2}F_{X2}(xa)), \ Q^{-1}(b_{Y1} + b_{Y2}F_{Y2}(ya)); \rho)|. \tag{C2}$$

Numerical algorithms for computing the bivariate normal distribution function can be found in Divgi (1979). The formation of this minimization problem represents an attempt to approximate $F(x, y)$ using $H(x, y; \rho)$ given in (A3) by choosing a value of ρ.

For $x \geqslant x_t$, the conditional distribution $H_{Y|X}(y|x)$ can be sampled for ensemble generation as follows: First we note that $H_{Y|X}(y|x)$ is an increasing function of y for a given x (see (A6)). Obtain $p_t = H_{Y|X}(y_t|x)$ using (C1b). For a given probability value p_i, if $p_i \geqslant p_t$, use (C1b) to compute $y_{p|x}$ from p_i; if $p_i < p_t$, assign 0 to $y_{p|x}$. For the case of $x < x_t$, Eq. (14) may be used for ensemble generation.

Appendix D. Alternative expressions for $c(x)$

Below, we give alternative expressions for $c(x)$ in Eq. (16). Using the notation of Section 2, we write Eq. (11a) of Herr and Krzysztofowicz (2005) as:

$$p_{10}G_X(x) + p_{11}D_X(x) = (p_{00} + p_{11})F_{X|X>0}(x).$$

Here $G_X(x)$, $D_X(x)$, and $F_{X|X>0}(x)$ are defined in Eqs. (7), (11), and (13), respectively. Assuming their respective PDFs $g_X(x)$, $d_X(x)$, and $f_{X|X>0}(x)$ are continuous for $x > 0$, we have:

$$p_{10}g_X(x) + p_{11}d_X(x) = (p_{10} + p_{11})f_{X|X>0}(x).$$

It follows from the above equation and Eq. (16) that:

$$c(x) = \frac{p_{10}g_X(x)}{(p_{10} + p_{11})f_{X|X>0}(x)}, \tag{D1}$$

$$1 - c(x) = \frac{p_{11}d_X(x)}{p_{10}g_X(x) + p_{11}d_X(x)}, \tag{D2}$$

$$1 - c(x) = \frac{p_{11}d_X(x)}{(p_{10} + p_{11})f_{X|X>0}(x)}. \tag{D3}$$

In the above, we have derived three new equations that are mathematically equivalent to Eq. (16). Our numerical experiments indicate that Eqs. (16) and (D2) were more robust in numerical evaluation than Eqs. (D1) and (D3) when parametric models were used for the density functions involved. For example, in one numerical experiment, we observed that the right-hand side of (D1) exceeded 1, a theoretical upper bound for $c(x)$, for small x values. This pathological behavior can be attributed to the sensitivity of the ratio of two skewed density functions to the change in x.

References

Azzalini, A., 1981. A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika 68, 326–328.

Bardossy, A., Plate, E.J., 1992. Space–time model for daily rainfall using atmospheric circulation patterns. Water Resources Research 28, 1247–1259.

Berrocal, V.J., Raftery, A.E., Gneiting, T., 2008. Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. Annals of Applied Statistics 2, 1170–1193.

Bröcker, J., Smith, L.A., 2007. Increasing the reliability of reliability diagrams. Weather and Forecasting 22, 651–661.

Charba, J.P., Reynolds, D.W., McDonald, B.E., Carter, G.M., 2003. Comparative verification of recent quantitative precipitation forecasts in the national weather service. a simple approach for scoring forecast accuracy. Weather and Forecasting 18, 161–183.

Chernick, M.R., 2008. Bootstrap Methods: A Guide for Practitioners and Researchers, second ed. Wiley.

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

Cong, S., Herr, H.D., Wu, L., Demargne, J., Duan, Q., Seo, D.-J., 2006. Accounting for intermittency in generation of precipitation ensembles from quantitative precipitation forecast. 2006 Joint Assembly of AGU GS MAS MSA SEG UGM, 23–26 May 2006, H22B-06.

Demargne, J., Wu, L., Seo, D.-J., Schaake, J.C., 2007. Experimental hydrometeorological and hydrologic ensemble forecasts and their verification in the US National Weather Service. IAHS Publications Series (Red Books), vol. 313, pp. 177–187.

Demargne, J., Brown, J., Liu, Y., Seo, D.-J., Wu, L., Toth, Z., Zhu, Y., 2010. Diagnostic verification of hydrometeorological and hydrologic ensembles. Atmospheric Science Letters HEPEX special issue 11 (2), 114–122.

Divgi, D.R., 1979. Calculation of univariate and bivariate normal probability functions. The Annals of Statistics 7 (4), 903–910.

Eckel, F.A., Walters, M.K., 1998. Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. Weather and Forecasting 13, 1132–1147.

Fasano, G., Franceschini, A., 1987. A multidimensional version of the Kolmogorov–Smirnov test. Monthly Notices of the Royal Astronomical Society 225, 155–170.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society Series B: Statistical Methodology 69 (2), 243–268.

Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review 133, 1098–1118.

Hamill, T.M., Hagedorn, R., Whitaker, J.S., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecast. Part II: precipitation. Monthly Weather Review 136, 2620–2632.

Hamill, T.M., Whitaker, J.S., 2006. Probability quantitative precipitation forecasts based on reforecast analogs: theory and application. Monthly Weather Review 134, 3209–3229.

Hansen, B.E., 2004. Bandwidth Selection for Nonparametric Distribution Estimation. Working Paper.

Herr, H.D., Krzysztofowicz, R., 2005. Generic probability distribution of rainfall in space: the bivariate model. Journal of Hydrology 306, 234–263.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting 15, 559–570.

Hosking, J.R.M., 2005. Fortran Routines for Use with the Method of L-Moments. IBM Research Report, RC 20525 (90933).

Jolliffe, I.T., Stephenson, D.B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley and Sons, Chichester.

Kelly, K.S., Krzysztofowicz, R., 1997. A bivariate meta-Gaussian density for use in hydrology. Stochastic Hydrology and hydraulics 11, 17–31.

Krzysztofowicz, R., Evans, W.B., 2008. Probabilistic forecasts from the national digital forecast database. Weather and Forecasting 23, 270–289.

Law, A.M., Kelton, W.D., 2000. Simulation Modeling and Analysis, third ed. McGraw Hill.

Makhnin, O.V., McAllister, D.L., 2009. Stochastic precipitation generation based on a multivariate autoregression model. Journal of Hydrometeorology 10, 1397–1413.

Peacock, J.A., 1983. Two-dimensional goodness-of-fit testing in astronomy. Monthly Notices of the Royal Astronomical Society 202, 615–627.

Reiss, R.D., 1981. Nonparametric estimation of smooth distribution functions. Scandinavian Journal of Statistics 8, 116–119.

Rohatgi, V.K., 1976. An Introduction to Probability Theory and Mathematics Statistics. John Wiley & Sons.

Schaake, J.C., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., Seo, D.-J., 2007. Precipitation and temperature ensemble forecasts from single-value forecasts. Hydrology and Earth System Sciences Discussions 4, 655–717.

Seo, D.-J., Breidenbach, J.P., 2002. Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. Journal of Hydrometeorology 3, 93–111.

Seo, D.-J., Herr, H.D., Schaake, J.C., 2006. A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. Hydrology and Earth System Sciences Discussions 3, 1987–2035.