Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions*

MICHAEL SCHEUERER

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

THOMAS M. HAMILL

Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 18 February 2015, in final form 14 August 2015)

ABSTRACT

A parametric statistical postprocessing method is presented that transforms raw (and frequently biased) ensemble forecasts from the Global Ensemble Forecast System (GEFS) into reliable predictive probability distributions for precipitation accumulations. Exploratory analysis based on 12 years of reforecast data and $^1/_8°$ climatology-calibrated precipitation analyses shows that censored, shifted gamma distributions can well approximate the conditional distribution of observed precipitation accumulations given the ensemble forecasts. A nonhomogeneous regression model is set up to link the parameters of this distribution to ensemble statistics that summarize the mean and spread of predicted precipitation amounts within a certain neighborhood of the location of interest, and in addition the predicted mean of precipitable water. The proposed method is demonstrated with precipitation reforecasts over the conterminous United States using common metrics such as Brier skill scores and reliability diagrams. It yields probabilistic forecasts that are reliable, highly skillful, and sharper than the previously demonstrated analog procedure. In situations with limited predictability, increasing the size of the neighborhood within which ensemble forecasts are considered as predictors can further improve forecast skill. It is found, however, that even a parametric postprocessing approach crucially relies on the availability of a sufficiently large training dataset.

1. Introduction

Ensemble predictions are now routinely generated at operational weather prediction centers worldwide (Molteni et al. 1996; Toth and Kalnay 1993, 1997; Houtekamer and Derome 1995; Charron et al. 2010). Despite many improvements to them over the last two decades, precipitation forecasts from the ensembles are still typically unreliable, be it from insufficient model resolution, less-than-optimal initial conditions, suboptimal treatment of model uncertainty, and/or sampling error. For this reason, statistical postprocessing of the output of an ensemble prediction system is commonly an integral part of the forecast process, since it can improve the reliability and skill of probabilistic guidance (e.g., Wilks and Hamill 2007; Hamill et al. 2008, and references therein). By comparing past forecasts with their verifying observations, systematic biases and inadequate representation of forecast uncertainty can be identified, and the current forecast can be adjusted such as to minimize these systematic errors. When the forecasts are provided on a grid that is too coarse to resolve small-scale effects that affect the weather variable under consideration, many postprocessing methods also implicitly perform a statistical downscaling.

The statistical postprocessing of precipitation accumulations is far more challenging than the postprocessing of weather variables like surface temperature or wind speed for several reasons:

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/MWR-D-15-0061.s1.

Corresponding author address: Dr. Michael Scheuerer, Physical Sciences Division, NOAA/ESRL, 325 Broadway, R/PSD1, Boulder, CO 80305-3337.
E-mail: michael.scheuerer@noaa.gov

DOI: 10.1175/MWR-D-15-0061.1

1) Their mixed discrete/continuous nature (positive probability of being exactly zero, continuous value range for positive precipitation amounts) makes it difficult to find an adequate parametric distribution model.
2) Forecast uncertainty typically increases with the magnitude of expected precipitation amounts; this must be taken into account when setting up a model for the conditional distribution of observed precipitation amounts given the ensemble forecasts.
3) High precipitation amounts occur very infrequently; a customized treatment of these cases may, therefore, require a vast amount of training data.

The advantages and disadvantages of the different postprocessing approaches proposed in the literature are typically related to those three challenges. Nonparametric approaches like the analog method (Hamill and Whitaker 2006; Hamill et al. 2015) completely avoid the first two issues, but may be disproportionately affected by the third one since their treatment of high precipitation amounts neglects the information with training samples with lower precipitation amounts. Parametric methods, on the other hand, can extrapolate the relations found between observations and forecasts of low and moderate magnitudes to higher magnitudes. In doing so, they may reduce the demand for training data, but the quality of the corresponding predictions strongly depends on the adequacy of the parametric assumptions that have to be made. Examples of parametric approaches that have been developed for quantitative precipitation forecasts include Bayesian model averaging (BMA; Sloughter et al. 2007), extended logistic regression (ExLR; Wilks 2009; Ben Bouallègue 2013; Messner and Mayr 2014), and ensemble model output statistics (EMOS; Scheuerer 2014). All of them make somewhat ad hoc assumptions about the parametric form of the predictive distributions: Sloughter's BMA method models precipitation occurrence/nonoccurrence separately and assumes gamma distributions for positive precipitation amounts; ExLR implies the assumption of censored logistic distributions; Scheuerer's EMOS method assumes censored generalized extreme value distributions (GEVs). To deal with the issue of heteroscedasticity mentioned above, BMA and ExLR commonly apply power transformations to both forecasts and observations, with powers chosen such as to make the forecast error terms more homoscedastic. Scheuerer's EMOS method utilizes two different ensemble statistics that serve as predictors for the scale parameter of the censored GEV distributions.

In this paper we will leverage NOAA's second-generation GEFS reforecast dataset (Hamill et al. 2013) to systematically develop a parametric model for the conditional distribution of observed precipitation amounts given the ensemble forecasts. This will eventually lead to an approach similar to the one proposed by Scheuerer (2014), but based on censored, shifted gamma distributions (CSGD), and a more sophisticated heteroscedastic regression model that accounts for some further peculiarities of precipitation. In section 2 we briefly describe the forecast and observation data used in this study, and we introduce our CSGD model in section 3. Section 4 describes the actual postprocessing approach, which proceeds in three steps: first, the ensemble forecasts are adjusted such as to match the observation climatology, and are condensed into four ensemble statistics. Second, a CSGD model for the unconditional (climatological) distribution of the observations is fitted. Third, a nonhomogeneous regression model is set up that links the ensemble statistics to the CSGD parameters, and results in a conditional distribution model for the observations given the ensemble forecasts. This model is relatively complex, but a comparison with nonparametrically estimated conditional distributions of observed precipitation amounts shows that a certain degree of flexibility (and thus complexity) is necessary to address the peculiarities of precipitation. The benefit of developing a sophisticated parametric approach will become clear in section 5, where probabilistic forecasts generated by our method are verified and compared against those obtained with a state-of-the-art analog approach. The latter is nonparametric, thus even more flexible than the CSGD approach, and easier to implement. In situations where training data are sparse (e.g., rare events), however, the predictive performance of the CSGD method is favorable. Further experiments are presented that study how the different components of the CSGD contribute to the overall performance, and how reducing the amount of training data affects the quality of the fitted regression model. Section 6 provides a summary and points out challenges with parametric postprocessing approaches that require further investigation.

2. Data

The postprocessing method developed here is applied to 12-hourly accumulated precipitation forecasts during the period from January 2002 to December 2013 for lead times up to +6 days. All the forecast data were obtained from the second-generation GEFS reforecast dataset; the same data were used in a recent paper by Hamill et al. (2015), which discusses variants of the analog method for statistical postprocessing of ensemble precipitation forecasts. For precipitation, individual

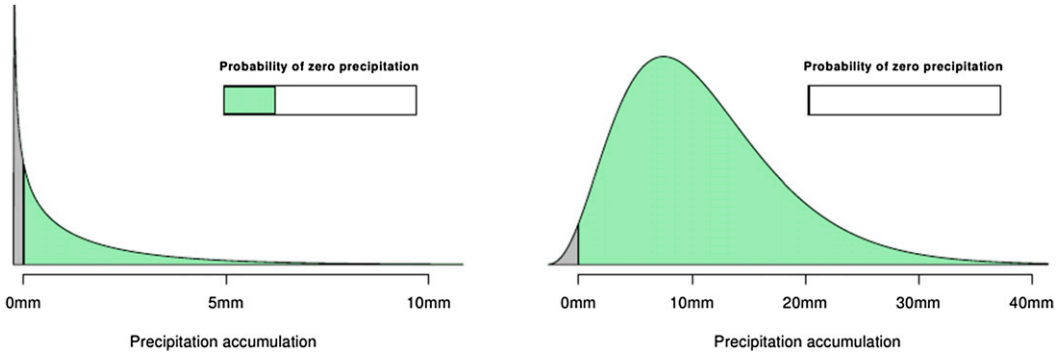

FIG. 1. Examples of censored, shifted gamma distributions. The fractions of the probability density function that fall below zero (shown in the gray shading) translate into a positive probability of being exactly zero.

forecasts by the 11-member GEFS reforecast ensemble were retrieved, and forecast data were extracted on GEFS's native Gaussian grid at ~½° resolution in an area surrounding the contiguous United States. Total-column ensemble-mean precipitable water is used as an additional predictor in our regression model, and the corresponding forecasts were interpolated to the same grid before further processing. Again as in Hamill et al. (2015), postprocessing and verification is performed against precipitation analyses from the climatology-calibrated precipitation analysis (CCPA) dataset of Hou et al. (2014), which were obtained on a ~⅛° grid inside the contiguous United States. The downscaling from the ~½° to the ~⅛° resolution will implicitly be part of the postprocessing procedure.

3. The censored, shifted gamma distribution

To set up a parametric postprocessing method, a suitable class of probability distributions must be identified. As precipitation occurrence/nonoccurrence and amount are modeled jointly, a convenient way to do so is using a continuous distribution that permits negative values, and left-censoring it at zero (i.e., replacing all negative values by zero). The censoring turns the probability for negative values of the uncensored distribution into a probability of observing a value equal to zero, thus ensuring requirement 1 described in section 1.

Exploratory data analysis reveals another challenging requirement for conditional distributions of precipitation accumulations: when the predictor variable (e.g., the ensemble-mean precipitation forecast) is small, then a strongly right-skewed distribution is called for, but the required skewness becomes smaller and smaller as the predictor variable's magnitude increases. To some extent, this behavior can be addressed by using gamma distributions, which are characterized by a shape parameter k and a scale parameter θ. Those two parameters are related to the mean μ and the standard deviation σ of the gamma distribution via

$$k = \frac{\mu^2}{\sigma^2}, \quad \theta = \frac{\sigma^2}{\mu} \tag{1}$$

(Wilks 2011, section 4.4.3). Since the predictive standard deviation increases more slowly than the predictive mean as the predictor variables increase, the shape parameter k increases, and as k increases, skewness decreases.

A disadvantage of the gamma distribution is that its value range is nonnegative. To make the above censoring idea feasible, we therefore introduce an additional parameter $\delta > 0$. This shifts the cumulative distribution function (CDF) of the gamma distribution somewhat to the left. That is, if F_k denotes the CDF of a gamma distribution with unit scale and shape parameter k, then the CDF $\tilde{F}_{k,\theta,\delta}$ of our CSGD model is defined by

$$\tilde{F}_{k,\theta,\delta}(y) = \begin{cases} F_k\left(\dfrac{y-\delta}{\theta}\right) & \text{for} \quad y \geq 0 \\ 0 & \text{for} \quad y < 0 \end{cases}. \tag{2}$$

Using the relations in (1), this distribution can also be parameterized by μ, σ, and δ: μ reflects the expected magnitude of precipitation, σ parameterizes prediction uncertainty, and δ reduces the magnitude of precipitation somewhat and controls the probability of zero precipitation. An illustration of the CSGD is given in Fig. 1. Note that σ affects both the continuous part of the distribution and the point mass at zero, which we feel is consistent with its interpretation as an uncertainty parameter: if the expected amount of precipitation is 1 mm, high forecast uncertainty implies that there is still a certain chance of observing much more precipitation, but also a significant chance of observing no precipitation at all. Increasing σ while keeping μ and

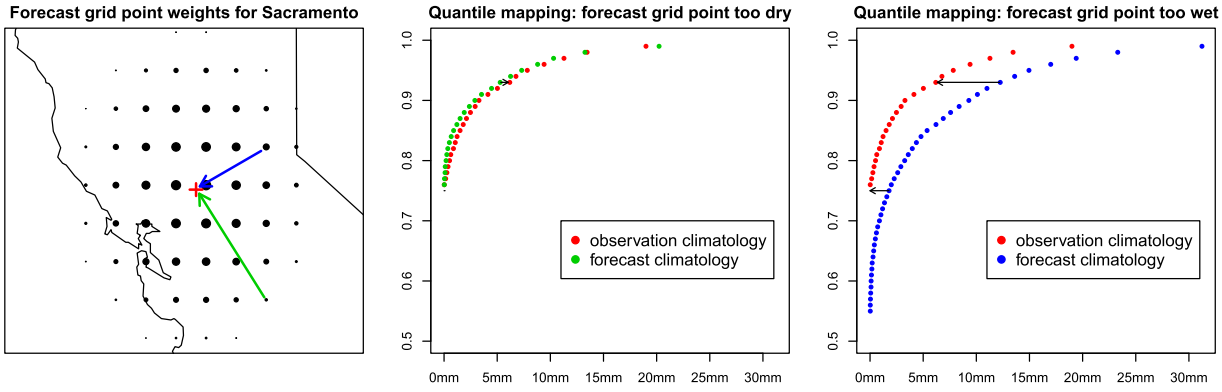

FIG. 2. (left) Illustration of the neighborhood weighting scheme and the climatology adjustment for an analysis grid point ("+") near Sacramento, CA, and $r = 2°$. Forecast grid points are denoted by "•," and their area is proportional to the weight w_{sx}. (middle),(right) Illustration of, for two of these forecast grid points, how the corresponding forecasts are adjusted by quantile mapping.

δ fixed shifts more mass below the censoring threshold, and thus accounts for both implications of increased uncertainty. A two-stage approach that models precipitation occurrence and amount separately offers more flexibility, but does not have a single parameter that can be interpreted as uncertainty in this way.

4. Postprocessing method

Having selected a family of probability distributions, we propose a procedure to link the three parameters of this distribution to the ensemble forecasts. This is done in three steps. First, quantile mapping is performed to adjust the ensemble precipitation forecasts such as to match the observation climatology. The adjusted forecasts are then reduced to two statistics that measure mean and spread of predicted precipitation accumulations. A further statistic is calculated that measures the mean precipitable water. In the second step, we fit a CSGD model as in (2) to the observed daily climatology of 12-h precipitation accumulations at each grid point (separately for each month). This CSGD model is the basis for a heteroscedastic regression model that is set up in the third step, and links the ensemble statistics from step 1 to the CSGD parameters, thus defining a predictive distribution for the observed precipitation accumulations, given the ensemble forecasts. We now consider each step in more detail.

a. Quantile mapping and ensemble statistics

As a first step in our postprocessing scheme, we attempt to correct systematic errors in ensemble forecast climatology. For example, the underlying numerical weather prediction may produce too many days with light precipitation and underforecast heavy precipitation events. Alternatively, these errors can arise due to coarser spatial resolution of the forecast grid compared to the grid on which analyzed precipitation is available. This first step can therefore also be viewed as a preliminary downscaling procedure.

Let s be a location associated with some analysis grid point. Prediction errors of the ensemble forecasts may result from inaccurately predicted magnitudes of a precipitation event as described above, but may also be caused by displacement errors. For example, a front or thunderstorm may have been predicted by the numerical weather prediction (NWP) model, but its position might be shifted away somewhat from its true position. The ensemble size of operational ensemble forecast systems is usually too small to represent this position uncertainty, and we, therefore, follow Scheuerer (2014) and consider ensemble forecasts at all forecast grid points within a certain neighborhood $N(s)$ of s as potential predictors for the analyzed precipitation amount at s. Forecast f_{xj} of ensemble member j at forecast grid point x is thus used multiple times to calculate ensemble and spatial means and spreads for all analysis grid point neighborhoods $N(s_1)$, $N(s_2)$, ... containing x. Each time, the forecasts within $N(s)$ are adjusted such that their climatology matches the respective observation climatology as illustrated in Fig. 2. This is achieved via quantile mapping: for each forecast f_{xj} we determine to which quantile $q_{f,x}(p)$, $p \in [0, 1]$ of the forecast climatology it corresponds, and then map it to the corresponding quantile $q_{o,s}(p)$ of the observation climatology. The quantiles are estimated from the training sample; for the GEFS ensemble considered here, all members are exchangeable, can thus be assumed to have the same forecast climatology, and can be pooled for the purpose of estimating the forecast quantiles. Estimating higher quantiles still comes with substantial sampling variability, and to make our quantile

mapping procedure more robust, we, therefore, resort to a linear approximation of the mapping function for quantiles above the 90% quantile (details of this procedure are given in appendix A in the online supplemental material).

To use the adjusted ensemble forecasts within a regression framework, they are next condensed into statistics that summarize the most important information. As discussed above, we propose that all forecast grid points in $N(s)$—which we take as a neighborhood around s with radius r—should be considered in determining these statistics, but we still expect forecasts at grid points closer to s to be more informative about the precipitation at s. Following Scheuerer (2014), we, therefore, weigh the forecast grid points according to their distance to s and let

$$w_{sx} \sim \max\left\{1 - \left[\frac{\mathrm{dist}(x,s)}{r}\right]^2, 0\right\}$$

with a constant of proportionality chosen such that the weights sum up to one (see left panel of Fig. 2 for an illustration of this weighting scheme). Assuming that we have an m-member ensemble with adjusted precipitation forecasts $\tilde{f}_{x1}, \ldots, \tilde{f}_{xm}$ and forecasts $\chi_{x1}, \ldots, \chi_{xm}$ of precipitable water, we consider the following ensemble statistics:

$$\mathrm{POP}_{f,s} := \frac{1}{m}\sum_{j=1}^{m}\sum_{x\in N(s)} w_{sx}\mathbf{1}_{\{\tilde{f}_{xj}>0\}}, \tag{3}$$

$$\overline{f}_s := \frac{1}{m}\sum_{j=1}^{m}\sum_{x\in N(s)} w_{sx}\tilde{f}_{xj}, \tag{4}$$

$$\overline{\chi}_s := \frac{1}{m}\sum_{k=1}^{m}\sum_{x\in N(s)} w_{sx}\chi_{xk}, \quad \text{and} \tag{5}$$

$$\mathrm{MD}_{f,s} := \frac{1}{m^2}\sum_{j,j'=1}^{m}\sum_{x,x'\in N(s)} w_{sx}w_{sx'}|\tilde{f}_{xj} - \tilde{f}_{x'j'}|. \tag{6}$$

The first statistic describes the probability of precipitation derived from the (augmented and weighted) ensemble. The second and third are the weighted means of predicted adjusted precipitation accumulations and precipitable water over all ensemble members and all forecast grid points in $N(s)$. The fourth statistic measures the dispersion of the predicted precipitation accumulations both between ensemble members and between grid points in $N(s)$. Unlike Scheuerer (2014), we do not use separate measures of dispersion for those two sources of variability in order to keep the number of parameters in our heteroscedastic regression model (defined below) as few as possible. We finally note that the adjustment of forecasts in $N(s)$ to the observation climatology at s via quantile mapping achieves two goals: first, it produces an implicit downscaling to the precipitation at elements on a finer grid; and second, it results in a homogenization of the forecasts within $N(s)$, so that the aggregation of forecasts within a large neighborhood to ensemble statistics is reasonable also in, for example, mountainous regions with substantially varying climatologies.

b. Unconditional precipitation accumulations

Although our main interest is in modeling the conditional distribution of observed precipitation accumulations given the ensemble forecasts, we first consider their unconditional (i.e., climatological) distributions. Studying those is much easier and yet quite instructive, as the conditional distributions should converge toward the unconditional distribution as forecast skill decreases. Moreover, they will allow us to parameterize the conditional distributions such as to make them more comparable across grid points with different climatologies. To fit the parametric CDF $\tilde{F}_{\mu,\sigma,\delta}$ to the empirical CDF $\hat{F}_n$ of the observations $y_1, \ldots, y_n$ at this grid point, we minimize the integrated quadratic distance

$$d_{\mathrm{IQ}}(\tilde{F}_{\mu,\sigma,\delta}, \hat{F}) = \int_0^\infty [\tilde{F}_{\mu,\sigma,\delta}(t) - \hat{F}_n(t)]^2\, dt \tag{7}$$

in μ, σ, and δ. According to Thorarinsdottir et al. (2013), this is equivalent to minimizing the mean continuous ranked probability score (CRPS):

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{crps}(\tilde{F}_{\mu,\sigma,\delta}, y_i), \tag{8}$$

where

$$\mathrm{crps}(F,y) = \int_{-\infty}^{\infty}[F(t) - H(t-y)]^2\, dt, \tag{9}$$

and $H(\cdot)$ is the Heaviside step function (i.e., it is equal to 1 if $t \geq 0$ and zero otherwise). After reparameterizing, the integral on the right-hand side can be expressed in closed form as

$$\begin{aligned}\mathrm{crps}(\tilde{F}_{k,\theta,\delta}, y) = {} & \theta\tilde{y}[2F_k(\tilde{y}) - 1] - \theta\tilde{c}F_k(\tilde{c})^2 + \theta k[1 + 2F_k(\tilde{c})F_{k+1}(\tilde{c}) - F_k(\tilde{c})^2 \\ & - 2F_{k+1}(\tilde{y})] - \frac{\theta k}{\pi}B\left(\frac{1}{2}, k+\frac{1}{2}\right)[1 - F_{2k}(2\tilde{c})],\end{aligned} \tag{10}$$

where $\tilde{c} := -\delta/\theta$, $\tilde{y} := [(y-\delta)/\theta]$, and $B(\cdot,\cdot)$ is the beta function (a derivation of this formula is given in appendix B in the online supplemental material). The availability of a closed form expression makes model fitting through numerical CRPS minimization computationally efficient. When performing this minimization, the constraint $\delta \geq -\mu$ is imposed in addition to the constraints μ, $\sigma > 0$ and $\delta \leq 0$ that are required for the distribution model to be well defined. The reason for this will become clearer later in this section, when we set up the regression model for the conditional distribution of the observation given the forecasts.

For solving the constrained optimization problems numerically, we use the FORTRAN 77 implementation of the Linearly Constrained Optimization Algorithm (LINCOA) by M. J. D. Powell [details of this algorithm have not been published yet, but the usual way of choosing a new vector of variables is described in Powell (2014)]. A starting value for the optimization is obtained through the following rationale: if we had $\sigma = \mu$, the underlying gamma distribution would have a shape parameter $k = 1$, which corresponds to the special case of an exponential distribution. For this distribution, the mean over all nonzero precipitation amounts is an estimate of μ (and σ), for any probability of precipitation π_{pop}, and δ can subsequently be estimated as $\delta = \mu \log(\pi_{\text{pop}})$. For the 12-h accumulations considered here, the best-fitting k is typically smaller than 1, with μ being overestimated by the assumption of an exponential distribution. Moreover, the first-guess estimates proposed above might violate the constraint $\delta \geq -\mu$. We, therefore, improve our first guess by fixing σ, gradually decreasing μ, and recalculating $\delta = -(\mu/k)F_k^{-1}(1-\pi_{\text{pop}})$ until $\delta > -\mu/2$. The resulting values of μ, σ, and δ are then used as starting values for the numerical minimization of (8). If $\pi_{\text{pop}} < 0.02$, we expect the number of days with nonzero precipitation to be too small to warrant stable estimates, and we, therefore, take the starting values as the final estimates. For extremely dry grid points with $\pi_{\text{pop}} < 0.005$, even the simple preliminary estimates might be unreliable, and we use ad hoc values $\mu = 0.0005, \sigma = 0.0182$, and $\delta = -0.000\,49$ to set up a parametric distribution model for the analyzed climatology. This choice complies with the constraint $\delta \geq -\mu$ and corresponds to a CSGD distribution with a probability of precipitation of slightly less than 0.005.

Figures 3 and 4 show examples of fitted CSGDs at a very wet location (West Palm Beach, Florida) and a very dry location (Phoenix, Arizona), respectively. The empirical and the fitted, parametric CDFs are virtually indistinguishable. The approximate character of the parametric distribution becomes more obvious when we plot its quantiles against the sorted observations. In those Q–Q plots we observe quite strong departures from the diagonal, especially in the upper tail. However, this is also where we expect significant sampling variability. To understand to what extent the departures might just be random, we add pointwise 95% Monte Carlo intervals by simulating 10 000 samples of the same size as the original observations according to the fitted distribution model, sorting them, and reporting the 2.5% and 97.5% quantile of the first elements, second elements, and so forth. The black dots in the Q–Q plots in Figs. 3 and 4 (and in all other examples that we studied) are mostly inside the 95% Monte Carlo intervals, suggesting that the distribution family proposed here is adequate for modeling unconditional distributions of precipitation accumulations.

c. Regression equations

The final step is now to set up and fit a regression model for the conditional distribution of observed precipitation accumulations given the forecasts. To this end, the ensemble statistics for location s defined above must be linked to the parameters μ_s, σ_s, and δ_s of our CSGD model in (1) and (2). Denote by $\mu_{\text{cl},s}$, $\sigma_{\text{cl},s}$, and $\delta_{\text{cl},s}$ the parameters of the climatological CSGD at s, and by $\overline{f}_{\text{cl},s}$ and $\overline{\chi}_{\text{cl},s}$ the climatological means of $\overline{f}_s$ and $\overline{\chi}_s$, respectively, calculated as averages of these quantities over the current training sample. We fix $\delta_s = \delta_{\text{cl},s}$ and model the conditional CSGDs as deviations from the climatological CSGD. The most basic regression model would let μ_s increase linearly with $\overline{f}_s$ and account for the fact that uncertainty about precipitation amounts increases as their amplitude increases by letting σ_s increase proportional to the square root of μ_s:

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

$$\sigma_s = \alpha_{6,s}\sigma_{\text{cl},s}\sqrt{\frac{\mu_s}{\mu_{\text{cl},s}}}. \tag{12}$$

This model has only three regression coefficients and is thus comparable in terms of model complexity with extended logistic regression (Wilks 2009). Equation (11) can easily be extended to include additional predictors for μ_s:

$$\mu_s = \mu_{\text{cl},s}\left(\alpha_{2,s} + \alpha_{3,s}\text{POP}_{f,s} + \alpha_{4,s}\frac{\overline{f}_s}{\overline{f}_{\text{cl},s}} + \alpha_{5,s}\frac{\overline{\chi}_s}{\overline{\chi}_{\text{cl},s}}\right). \tag{13}$$

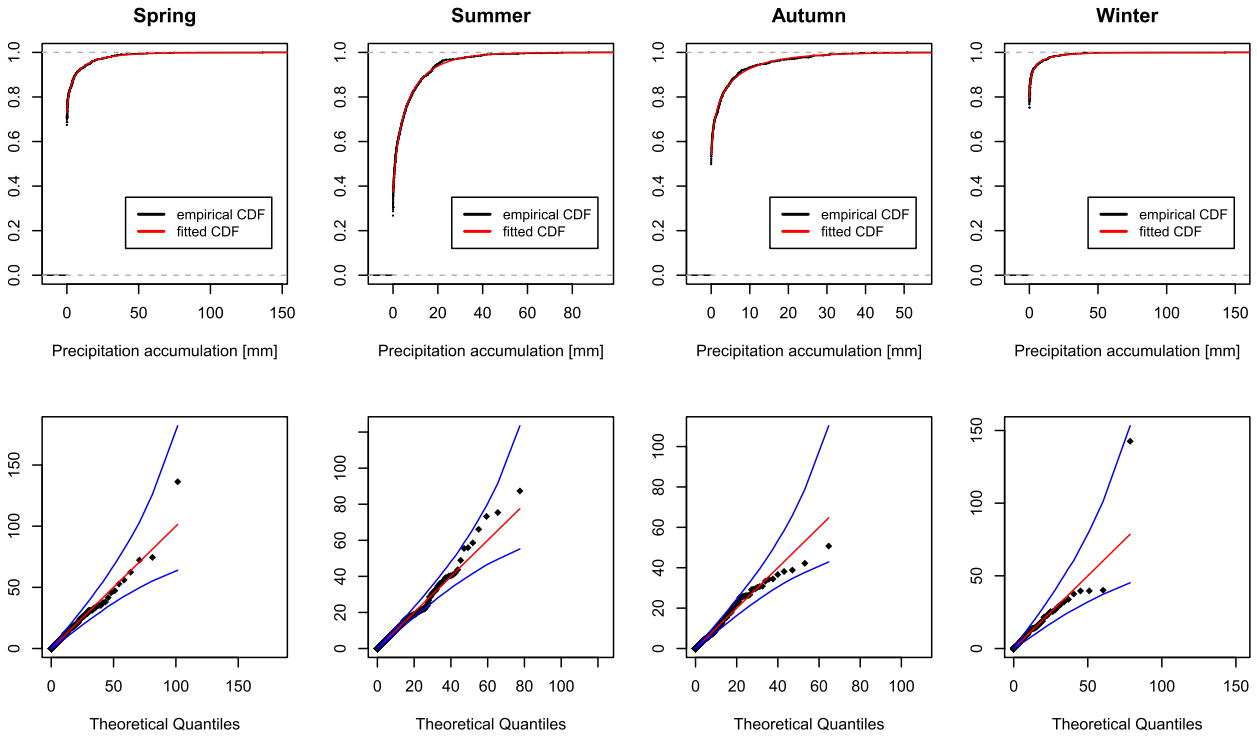

FIG. 3. (top) Empirical and fitted CDFs and (bottom) Q–Q plots of 12-hourly accumulated precipitation analyses in West Palm Beach, FL. The black dots in the bottom panels are the sorted observations, plotted against the corresponding theoretical quantiles from the fitted CSGD model. Ideally, they would lie on the diagonal (solid red line); due to sampling variability, however, any black dot lying within the pointwise 95% Monte Carlo intervals (solid blue lines) can still be considered consistent with the fitted model.

A predictor like $\mathrm{POP}_{f,s}$ can provide additional information about whether precipitation occurs or not (Sloughter et al. 2007; Bentzien and Friederichs 2012; Scheuerer 2014) while the consideration of precipitable water as an additional predictor can yield some improvement during the warm season where the forecast precipitation amount is more sensitive to the vagaries of the convective parameterization and its triggering scheme (Hamill and Whitaker 2006). Another generalization concerns the use of a suitable measure of ensemble spread as a predictor of the flow-dependent forecast uncertainty. Messner and Mayr (2014) have demonstrated that using such information can improve probabilistic wind speed predictions. For precipitation, Scheuerer (2014) has argued that location uncertainty should also be taken into account, and the statistic $\mathrm{MD}_{f,s}$ defined above summarizes the information in the ensemble forecasts about both of these sources of uncertainty. A final generalization relaxes the linearity assumption in (13) and the assumption in (12) that σ_s increases proportional to the square root of μ_s, and leads to the regression equations:

$$\mu_s = \frac{\mu_{\mathrm{cl},s}}{\alpha_{1,s}} \,\mathrm{log1p}\left[\mathrm{expm1}(\alpha_{1,s})\left(\alpha_{2,s} + \alpha_{3,s}\mathrm{POP}_{f,s} + \alpha_{4,s}\frac{\overline{f}_s}{\overline{f}_{\mathrm{cl},s}} + \alpha_{5,s}\frac{\overline{\chi}_s}{\overline{\chi}_{\mathrm{cl},s}}\right)\right] \quad \text{and} \tag{14}$$

$$\sigma_s = \alpha_{6,s}\sigma_{\mathrm{cl},s}\left(\frac{\mu_s}{\mu_{\mathrm{cl},s}}\right)^{\alpha_{7,s}} + \alpha_{8,s}\mathrm{MD}_{f,s}, \tag{15}$$

where $\mathrm{log1p}(x) = \log(1+x)$ and $\mathrm{expm1}(x) = \exp(x) - 1$. When $\alpha_{1,s}$ is close to zero, $\mathrm{expm1}(\alpha_{1,s}) \approx \alpha_{1,s}$ and $\mathrm{log1p}(\alpha_{1,s}z) \approx \alpha_{1,s}z$, and thus (14) reduces to the linear regression in (13). Some exploratory analysis (see also Fig. 6), however, suggests that a linear increase of μ_s with $\overline{f}_s$ is not always appropriate. In situations with reduced predictability (e.g., longer lead times, warm season), ensemble forecasts of high precipitation amounts are particularly unreliable and should be decreased proportionately more compared to forecasts of intermediate levels. This is the rationale behind the logarithm in (14). Increasing the parameter $\alpha_{1,s}$ reduces the growth of μ_s with increasing predictors and thus

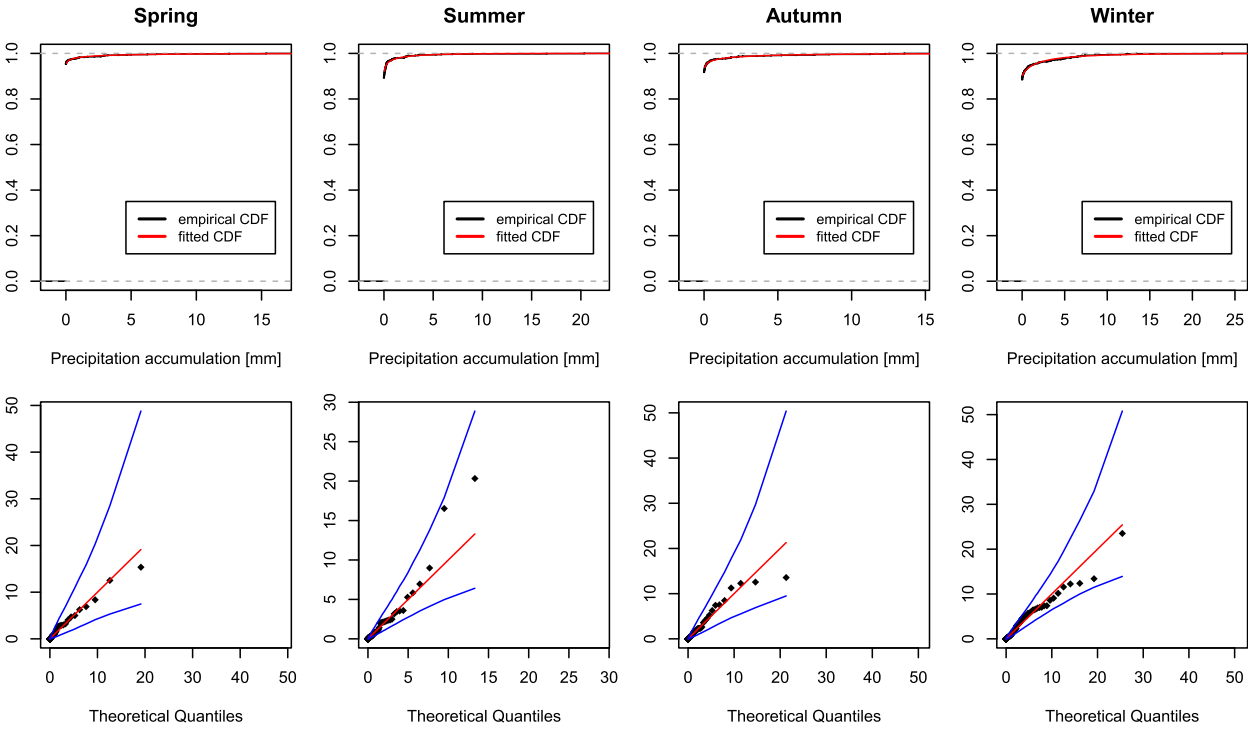

FIG. 4. As in Fig. 3, but for Phoenix, AZ.

accounts for the phenomenon just described. Equation (15) accounts for the heteroscedasticity in the uncertainty about precipitation accumulations in two different ways. The first term increases σ_s proportionally to a power of μ_s, while the second term is proportional to $\mathrm{MD}_{f,s}$ and thus accounts for flow-dependent uncertainty. Other parametric postprocessing methods for precipitation (e.g., ExLR, BMA) deal with heteroscedasticity by applying power transformations to both forecasts and observations with the goal of making their relation more homoscedastic. This might be preferable when only small training datasets are available because it results in a potentially less complex model for σ_s. It entails, however, the disadvantage of strongly distorting the scale of these variables. Consider, for example, the two hypothetical five-member ensembles (0.5, 1, 1.5, 2, 10) and (0, 2.5, 3.5, 4, 5), which have the same mean, but mean absolute differences of 4.0 and 2.3, respectively. The higher dispersion of the first ensemble results from one member predicting a substantially higher precipitation amount than the other members, which indicates a certain chance for heavier precipitation. If the mean absolute differences were calculated from cube root transformed forecasts, values of 0.595 and 0.730 would be obtained, suggesting more uncertainty in the second ensemble. This does not adequately reflect the situation in the original ensembles, and could thus reduce the value of flow-dependent uncertainty information in the ensemble. Modeling heteroscedasticity explicitly as in (15) avoids the need for data transformations but entails a more complex regression model. Including $\mu_{\mathrm{cl},s}$ and $\sigma_{\mathrm{cl},s}$ in the two regression equations does not change the actual model but is useful because it normalizes the regression parameters $\alpha_{1,s}, \ldots, \alpha_{8,s}$ and makes them more comparable across grid points with different climatologies.

Figure 5 illustrates the evolution of the predictive CSGD density with increasing mean precipitation $\overline{f}_s$ in a simplified setting where $\alpha_{3,s}$, $\alpha_{4,s}$, and $\alpha_{8,s}$ have been set to zero. It shows how the uncertainty increases with increasing $\overline{f}_s$; at the same time the skewness of the underlying gamma distribution becomes smaller and smaller. Choosing $\alpha_{1,s} = 0.05$ results in a moderate departure from a linear relation between $\overline{f}_s$ and μ_s.

Is the CSGD adequate for modeling conditional distributions of precipitation accumulations, and are the above regression equations for its parameters μ and σ adequate for describing the evolution of these parameters with increasing ensemble mean? To answer this we compare quantiles derived from predictive CSGDs with empirical conditional quantiles obtained without any parametric assumption. For this purpose, however, even the 12 years' worth of reforecast data are not enough if only data from a single grid point are

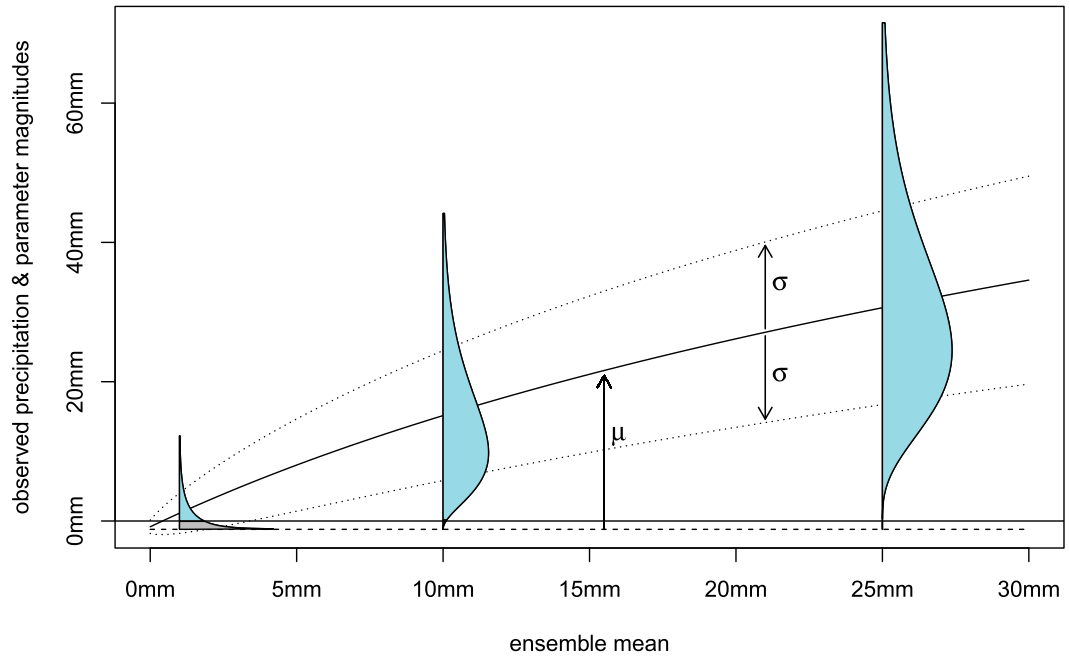

FIG. 5. Example of predictive CSGD densities, showing the evolution of the CSGD parameters μ and σ from (1) and (2) as a function of the ensemble-mean statistic $\overline{f}_s$.

considered. We focus on the analysis grid point corresponding to the city of Atlanta, Georgia, and we increase the corresponding dataset by selecting 200 additional analysis grid points within a radius of about 700 km around Atlanta that have a similar climatology (measured by comparing the empirical CDFs of analyzed precipitation amounts at the threshold values 0.5, 1.5, . . . , 99.5 mm) and are at least 40 km apart from each other. For each season, we then have about $91 \times 12 \times 201$ pairs of observations and quantile adjusted forecasts. We study again the simplified regression model with $\alpha_{3,s} = \alpha_{4,s} = \alpha_{8,s} = 0$ (i.e., with $\overline{f}_s$ as the only predictor). The conditional quantiles of the observation given $\overline{f}_s = x$ can then be approximated by considering all forecast–observation pairs for which $\overline{f}_s$ falls within a certain window $(x - \varepsilon, x + \varepsilon)$ around the precipitation amount x, and computing the quantiles of the corresponding observations. We let ε increase with x to account for the fact that the number of pairs with $\overline{f}_s \approx x$ decreases rapidly as x increases. For $x = 5$ mm and $x = 15$ mm our choice of ε is illustrated in Fig. 6. The crosses in each plot correspond to the empirical, conditional deciles (i.e., quantiles for the probabilities 0.1, . . . , 0.9) for each season and forecast lead times from +12 to +24 h and from +108 to +120 h. The solid lines are the quantiles obtained with our parametric regression model, fitted to the same training data. As for the unconditional CSGDs, the regression parameters are fitted by CRPS minimization using the LINCOA algorithm. Clearly, not every model-based quantile approximates the respective empirical quantile perfectly, and very irregular behavior cannot be captured. Yet one can see that the nonlinear relation between $\overline{f}_s$ and μ_s, which takes different forms depending on season and lead time, is accounted for by the logarithm in (14), allowing the red median curves to bend downward from a linear curve as $\overline{f}_s$ increases. Moreover, the increase of predictive uncertainty (distances between the blue decile curves) with increasing $\overline{f}_s$ is captured quite well by the model for σ_s given in (15). It is worth noting that our method for getting empirical estimates of conditional quantiles is quite similar to what is being done by analog approaches. Those techniques are much more flexible and avoid the approximation errors entailed by parametric methods. On the other hand, several of the plots in Fig. 6 also suggest that the empirical quantiles for large values of $\overline{f}_s$ are subject to quite substantial sampling error, even in the situation considered here where we choose the ''analogs'' from a training dataset of size $91 \times 12 \times 201$.

Finally, consider how the regression model in (14) and (15) for the predictive CSGDs approaches the parameters for the climatological CSGD in the limit where the raw ensemble forecasts have no skill. As the lead time increases, one can expect that the four predictors $\mathrm{POP}_{f,s}$, $\overline{f}_s$, $\overline{\chi}_s$, and $\mathrm{MD}_{f,s}$ become less and less informative about the true weather, and so the corresponding regression parameters $\alpha_{3,s}$, $\alpha_{4,s}$, $\alpha_{5,s}$, and $\alpha_{8,s}$

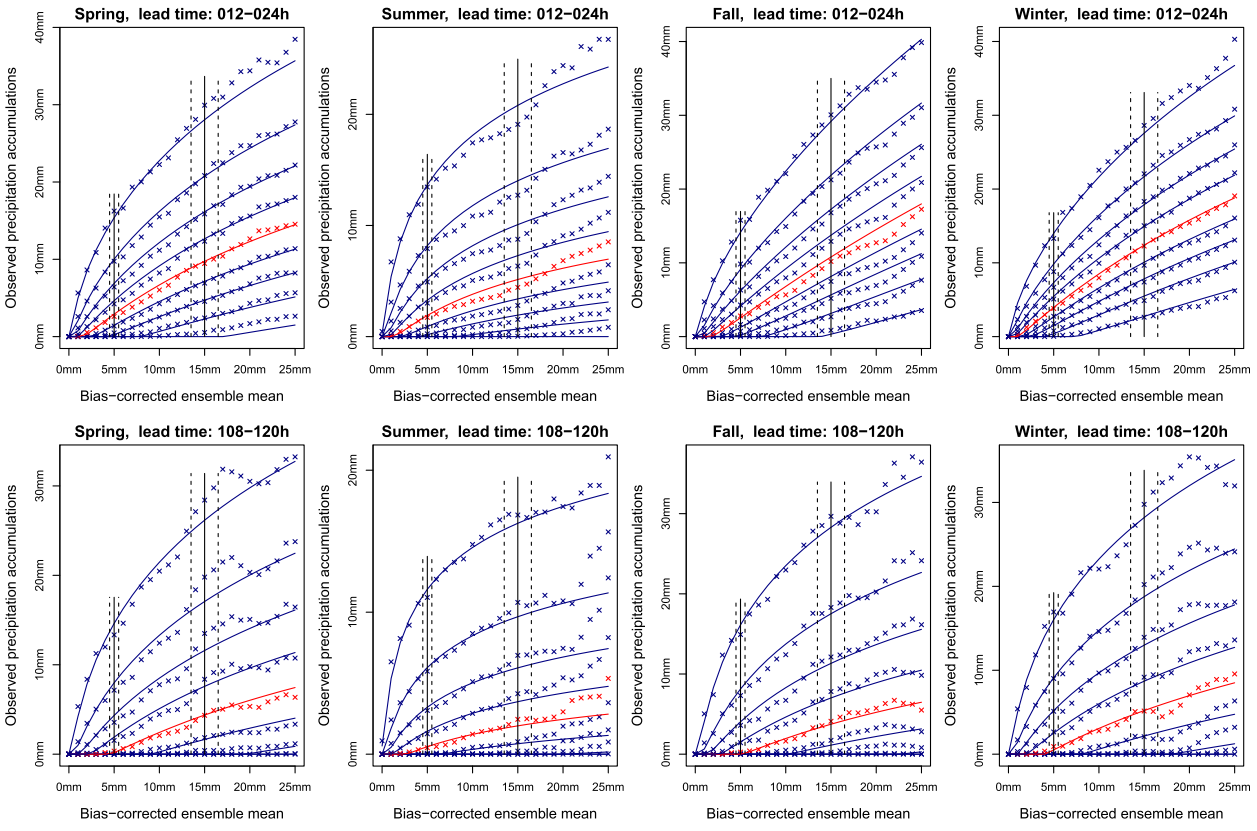

FIG. 6. Conditional deciles (median is highlighted in red) obtained with the augmented Atlanta dataset. Empirical deciles are depicted as crosses. For each conditioning value 1, 2, … , 25 mm they are obtained as empirical deciles of the observations corresponding to ensemble-mean statistics within a certain bin (with 5 and 15 mm depicted as vertical dashed lines) around this value. Deciles derived from the CSGD regression model are depicted as solid lines.

tend to zero. If at the same time $\alpha_{2,s}$ and $\alpha_{6,s}$ tend to one, then μ_s and σ_s tend to the climatological CSGD parameters $\mu_{\mathrm{cl},s}$ and $\sigma_{\mathrm{cl},s}$, whatever the values of $\alpha_{1,s}$ and $\alpha_{7,s}$, and so the climatological CSGD results as a limiting case. Including $\mu_{\mathrm{cl},s}$ and $\sigma_{\mathrm{cl},s}$ in the regression equations (14) and (15) can therefore be seen as a kind of normalization, which helps reduce the dependence of the regression parameters $\alpha_{1,s}, \ldots, \alpha_{8,s}$ on the climatology at location s and thus renders them more comparable across different grid points.

Modeling the conditional distributions as deviations from the climatological distributions requires some constraints of the latter. We found that this deviation concept does not work well at very dry locations if the shift parameter $\delta_{\mathrm{cl},s}$ of the climatological CSGD is large compared to $\mu_{\mathrm{cl},s}$. In this case, positive precipitation accumulations correspond to the tail end of the underlying gamma distribution, and deforming this distribution into a CSGD with a moderate to high probability of precipitation is rather unnatural. By introducing the constraint $\delta_{\mathrm{cl},s} \geq -\mu_{\mathrm{cl},s}$ on the climatology parameters in section 4b, we enforce a very small shape parameter k. The mass of the underlying gamma distribution is then concentrated near zero, and a very small shift is sufficient to obtain a high probability of values less than zero. Fitting a climatological CDF to the observation data under this constraint can result in a slightly suboptimal fit to the empirical, climatological CDF near zero, but this degradation is offset by the fact that the fitted CSGD permits a natural deformation into the predictive CSGD for any value of the predictors.

5. Validation of the CSGD method

We apply our CSGD regression method to the full dataset described in section 2. Now, every grid point of the CCPA grid (within the CONUS) is processed separately. Forecasts are cross validated; for example, 2002 forecasts are trained using 2003–13 data. To account for seasonal differences, a separate set of (both climatological and regression) parameters is fitted for each month; training data are composed of all forecasts and observations from +45 days around the 15th of the month under consideration and all years except the one

for which forecasts are sought. This results in a training sample size of 91 × 11 at each grid point. Compared to the amount of training data that are typically used for weather variables like wind speed or temperature, this training sample size seems fairly large. At very dry locations, however, the majority of both forecasts and observations are zero, and thus carry only limited information that can be leveraged for model fitting. For the parameters of the unconditional CSGDs we already described our special treatment of these dry cases in section 4b. For the regression parameters, we increase the training dataset of any grid point where the climatological probability of precipitation is less than 0.05 by considering also the data at adjacent grid points in the east–west and north–south direction. For grid points with a climatological probability of precipitation of less than 0.02, we additionally add the training data from diagonal neighbors. Parameters are estimated via CRPS minimization, subject to the following bounds:

$$0.001 \le \alpha_{1,s}, \quad \alpha_{2,s} \le 1, \quad 0 \le \alpha_{3,s}, \quad \alpha_{4,s}, \alpha_{5,s} \le 1.5,$$
$$0.1 \le \alpha_{6,s}, \quad \alpha_{7,s} \le 1, \quad \text{and} \quad 0 \le \alpha_{8,s} \le 1.5,$$

which are partly ad hoc and partly based on the discussion at the end of the previous section. In our experiments, CRPS minimization gave slightly better results than classical maximum likelihood estimation, which is nonrobust and tends to favor overdispersive predictive CSGDs. The same conclusion was reached by Gneiting et al. (2005) in the context of temperature postpostprocessing. Initially, we fix the radius of the neighborhood within which forecasts are considered as predictors (see section 4a) to $r = 2°$ (≈200 km).

a. *Overall performance and model complexity*

First, we take a look at the overall predictive performance of our CSGD method, measured by the continuous ranked probability skill score (CRPSS), which quantifies the improvement of the CRPS of the predictive CSGDs over climatological forecasts. We also study in how far the different predictors and the nonlinear and heteroscedastic components of our model contribute to this overall performance. As a benchmark we use the basic model defined by (11) and (12), which has only three parameters and is comparable in terms of model complexity with the basic ExLR model by Wilks (2009). Direct comparisons of the parametric postprocessing approaches mentioned in section 1 (ExLR, BMA, EMOS) suggest that their predictive performance is quite similar (Schmeits and Kok 2010; Scheuerer 2014), so how much extra skill can be gained by adding additional predictors or permitting certain forms of nonlinearity? We increase model complexity step by step as follows:

1) Use the nonlinear model (14) for μ_s but still use $\overline{f}_s$ as the only predictor.
2) Relax the assumption $\alpha_{7,s} = 0.5$ in (12) about the rate of increase of σ_s with increasing μ_s.
3) Use the full model (15) for σ_s by adding $\mathrm{MD}_{f,s}$ as a predictor for forecast uncertainty.
4) Add $\overline{\chi}_s$ (precipitable water) as a predictor for μ_s.
5) Add $\mathrm{POP}_{f,s}$ as a predictor for μ_s.

Figure 7 depicts the overall CRPSS (for the full model) for different lead times and the CRPSS increase that results from adding step by step the extra components described above. The first thing to note is the pronounced seasonal cycle of the CRPSS. Summertime convection is more difficult to forecast than synoptic-scale winter precipitation, and so forecast skill during the cool season is substantially higher than during the warm season. This pattern is inherited from the raw ensemble predictions, the corresponding results can be found in Hamill et al. (2015). The increase in skill due to the different refinements of the basic model is rather moderate for each individual extension, but sums up to a cumulative increase of about 0.01–0.015. The biggest benefit results from allowing a nonlinear increase of μ_s with $\overline{f}_s$, especially for longer lead times (see right panel of Fig. 7). The predictor $\mathrm{POP}_{f,s}$ yields a rather constant improvement in skill over all months of the year, while the predictor $\overline{\chi}_s$ (precipitable water) becomes especially useful in the warm season but adds no information to the ensemble precipitation forecasts during the more predictable cool season. The converse is true for the $\mathrm{MD}_{f,s}$ predictor that measures the spread of the forecasts between different ensemble members and forecast grid points within $N(s)$: it provides useful information about flow-dependent forecast uncertainty during the cool season, but does not improve (or even degrades, for longer lead times) probabilistic forecast skill during the warm season. The degradation is presumably a result of overfitting, to which the $\mathrm{MD}_{f,s}$ predictor is particularly prone, and that becomes a more serious concern as the signal-to-noise ratio in the training dataset decreases. Finally, we note that estimating the rate of increase of σ_s with increasing μ_s rather than fixing $\alpha_{7,s} = 0.5$ adds some flexibility, but the resulting benefit on predictive performance is quite marginal.

Figure 8 depicts maps of CRPSS values of the CSGD predictions to provide an impression about regional differences. Forecast skill is largest along the east and especially the west coast of the CONUS, which we believe is due to the relatively high predictability of orographically induced, synoptically forced precipitation. The general spatial pattern of forecast skill resembles

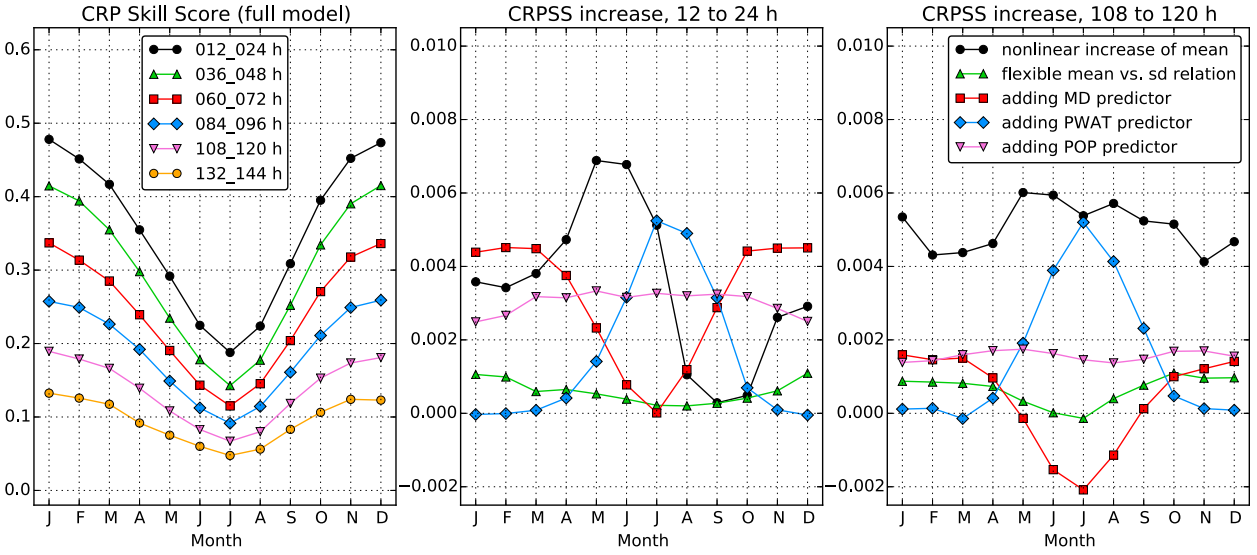

FIG. 7. (left) CRP skill scores for different lead times, separately for each month, but aggregated over all analysis grid points within the CONUS. Increases of CRPSSs due to increased model complexity are shown for (middle) +12- to +24-h lead time and (right) +108- to +120-h lead time.

that of the raw ensemble (see Hamill et al. 2015) while the skill is significantly better.

b. *CSGD versus analog method: Brier skill scores and reliability*

The CRPS studied so far is a useful and common measure of the overall skill, but it does not allow any conclusion about how skillful the CSGD forecasts are for predicting light, intermediate, and heavy precipitation events. To answer this question, we study Brier skill scores [Wilks 2011, see his Eqs. (7.34) and (7.35)] for the three thresholds of 1, 10, and 25 mm $(12\,\mathrm{h})^{-1}$. We further compare the predictive performance of the CSGD approach to a recently proposed variant of the rank analog approach by Hamill and Whitaker (2006), where supplemental locations are used to augment the training dataset at each analysis grid point. This adjustment to the rank analog procedure was shown to substantially improve probabilistic forecasts for heavy precipitation events (Hamill et al. 2015). Can the same or even more improvement be achieved by a parametric postprocessing scheme? Figure 9 depicts the monthly Brier skill scores (BSSs) for both methods, the three different thresholds, and forecast lead times up to +6 days. Even for the >1 mm $(12\,\mathrm{h})^{-1}$ event, the CSGD method can still improve upon the analog method, despite the fact that this is a rather common event at most grid points, and we should expect that sufficiently close analogs can usually be found. The fact that the CSGD method can compete with the analog approach in this situation suggests that

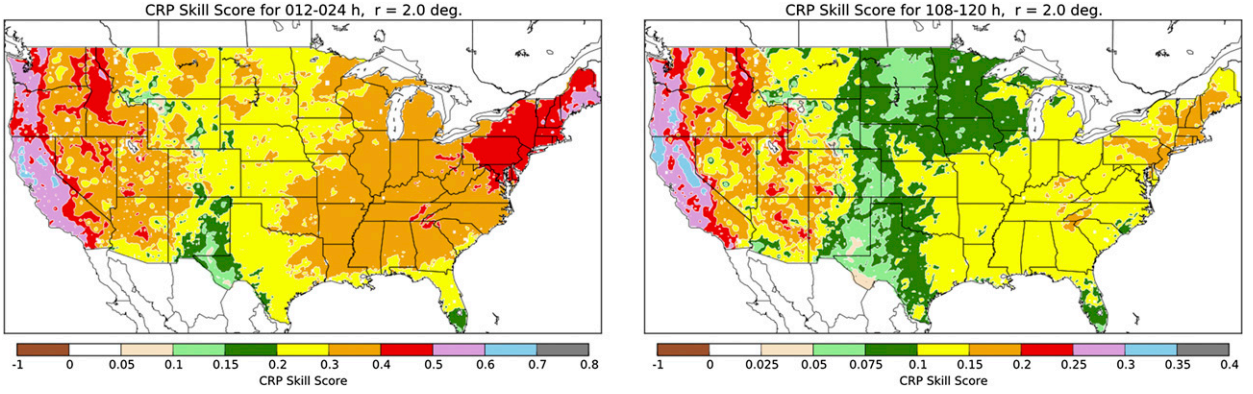

FIG. 8. Map of CRPSS values, aggregated over all months and all cross-validated years: (left) for +12- to +24-h lead time and (right) for +108- to +120-h lead time.

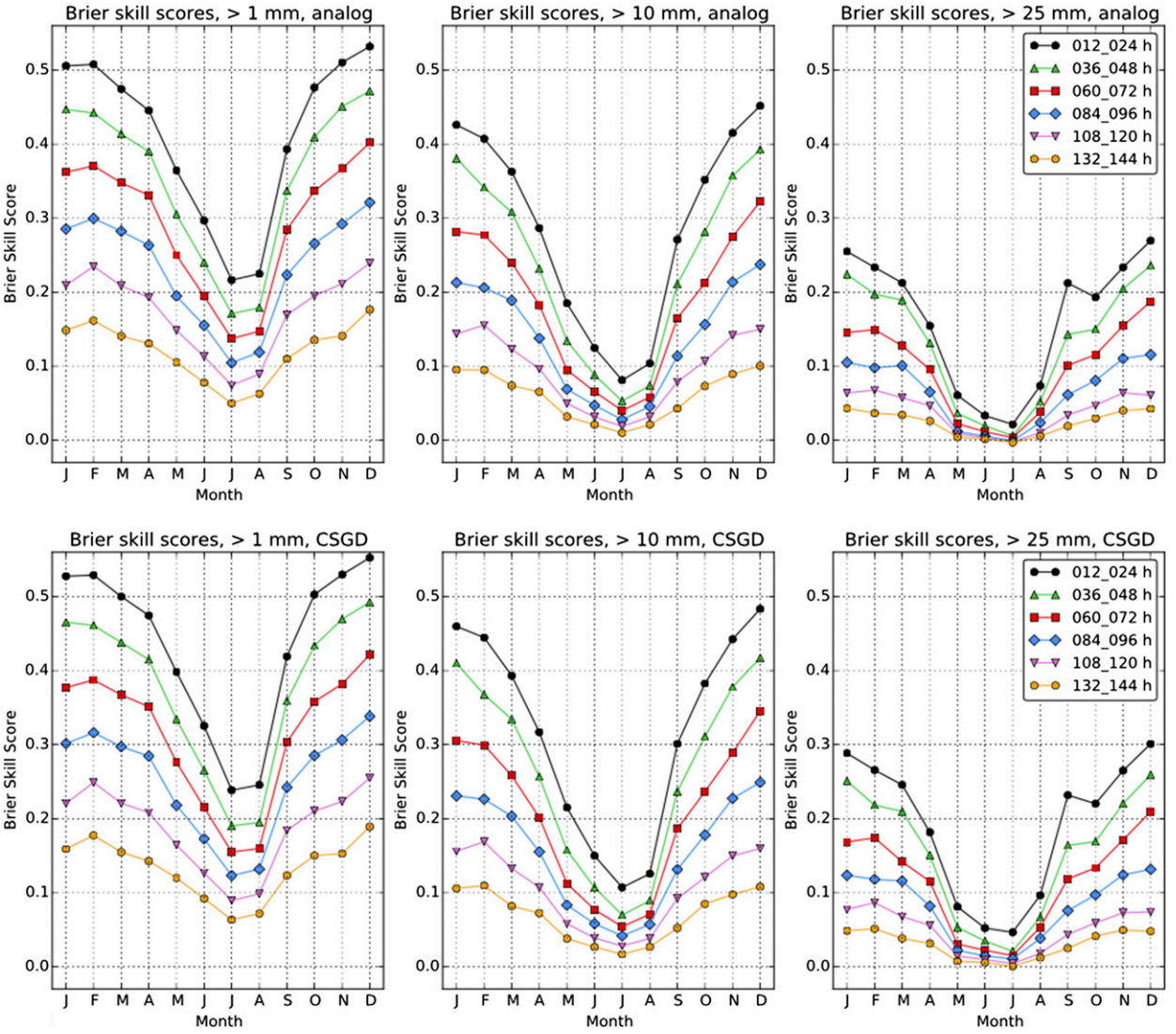

FIG. 9. Brier skill scores for different lead times and different event thresholds, separately for each month, but aggregated over all analysis grid points within the CONUS: (top) Results for the rank-analog method and (bottom) results for the CSGD regression approach.

our parametric approximation does not degrade predictive performance even when analog methods can be expected to perform very well. Comparing results for higher thresholds, we find that the probabilistic CSGD forecasts are again able to improve upon the forecasts by the analog method. The event >25 mm $(12\,\text{h})^{-1}$ is relatively rare, making it difficult to find a sufficient number of suitable analogs, even if supplemental locations are added to increase the training datasets. Our parametric method, on the contrary, can extrapolate relations found for more common situations and thus yield superior predictions of rare events.

To provide some understanding about the causes of the better performance of our parametric method compared to the nonparametric analog approach, we consider reliability diagrams for the same events as above [thresholds 1, 10, and 25 mm $(12\,\text{h})^{-1}$] and lead times from +12 to +24 h and from +108 to +120 h. The plots in Figs. 10 and 11 suggest that both methods yield reliable probabilistic forecasts at short lead times. At longer lead times, they are still sufficiently accurate, though somewhat less reliable. By comparing the inset frequency histograms, one can see that the performance gain of our CSGD method is mainly due to increased resolution; it issues high probabilities for observing heavy precipitation more frequently without degrading the reliability compared to the analog approach, which does not rely on parametric assumptions.

We illustrate the last point by considering a heavy precipitation event that took place over Washington

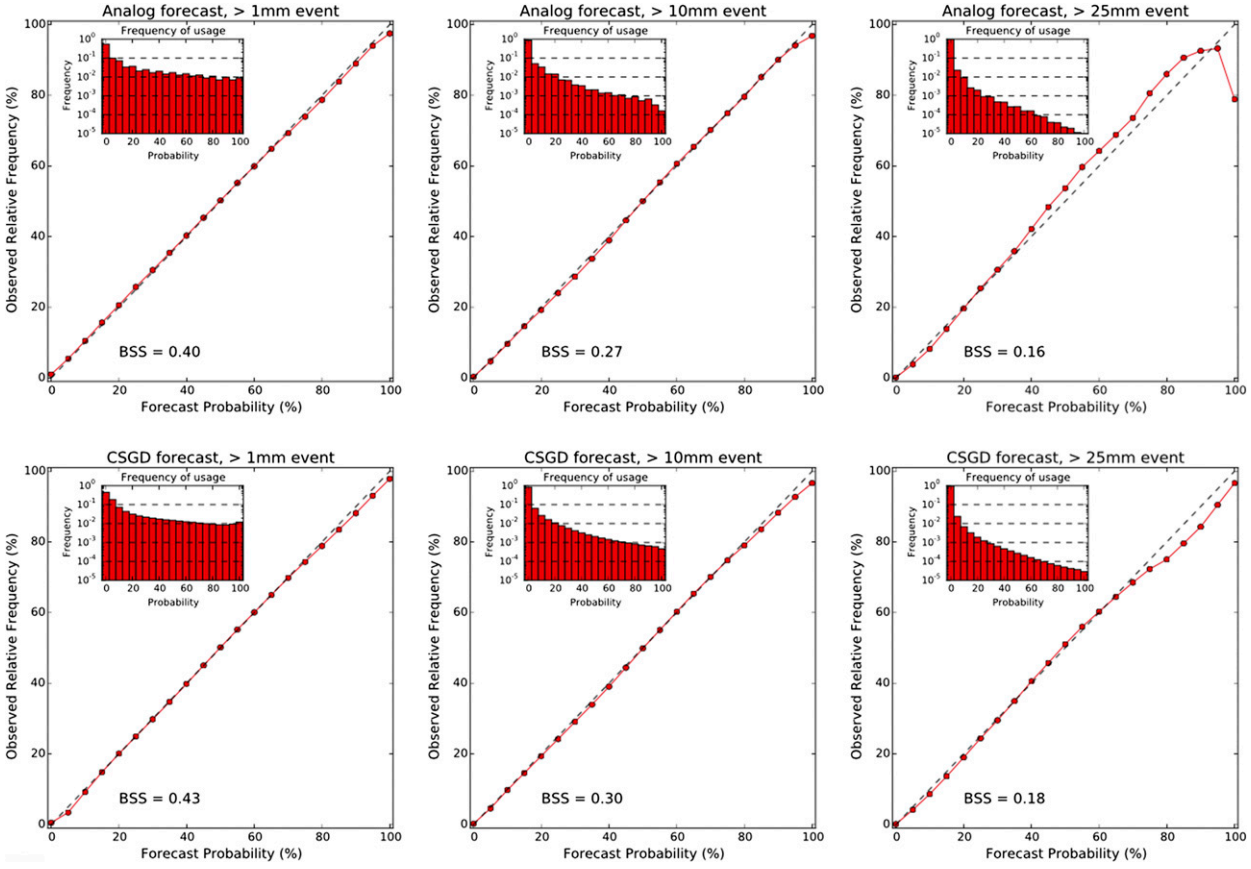

FIG. 10. Reliability diagrams for +12- to +24-h lead time and different event thresholds, calculated with forecast–observation pairs of all months, all cross-validated years, and all analysis grid points within the CONUS. (top) Results for the rank-analog method and (bottom) results for the CSGD regression approach. The inset histograms depict the frequency with which each category was predicted.

State between 1200 UTC 6 November and 0000 UTC 7 November 2006. Figure 12 shows the analyzed precipitation accumulations for that period, as well as +12- to +24-h lead predicted probabilities for exceeding 25 mm $(12\,\text{h})^{-1}$ of precipitation by the raw ensemble, the analog approach, and the CSGD regression method. The raw ensemble forecasts for that day were quite accurate, but since this is not always the case, one can expect that calibrated probabilistic forecasts modulate the high forecast probabilities. The analog approach modulates them more strongly, issuing rather moderate probabilities. On the other hand, the CSGD method largely retains the strong signal from the raw ensemble, and hence provides decision-makers with a more unequivocal expectation of heavy precipitation.

c. *Performance with a greatly reduced training dataset*

The results by Hamill et al. (2015) underscore the importance of a sufficiently large training dataset for statistical postprocessing, especially when the interest is in heavy precipitation events. What if a large reforecast dataset is not available? Can the CSGD approach retain its strong performance, or will it lose a large amount of skill as a result of overfitted regression parameters? Can possible overfitting be avoided by supplementing the training dataset at each grid point with training data from other grid points? To answer these questions we repeat the entire procedure (quantile mapping, calculating ensemble statistics, fitting the regression model) described above, this time using, for each of the 12 verification years, only forecast data from the preceding year or from the preceding three years (defining 2013 to be the year that precedes 2002) for training. We are thus left with only 91 and 273 training days, respectively, for quantile mapping and model fitting. Using just a single year of training data mimics the situation where no reforecasts are produced, but one year of training data are available from a preoperational test phase after a major update of the NWP system. Such an update would only have a limited or no effect on the verification/calibration data, and we, therefore, use the same CCPA datasets as

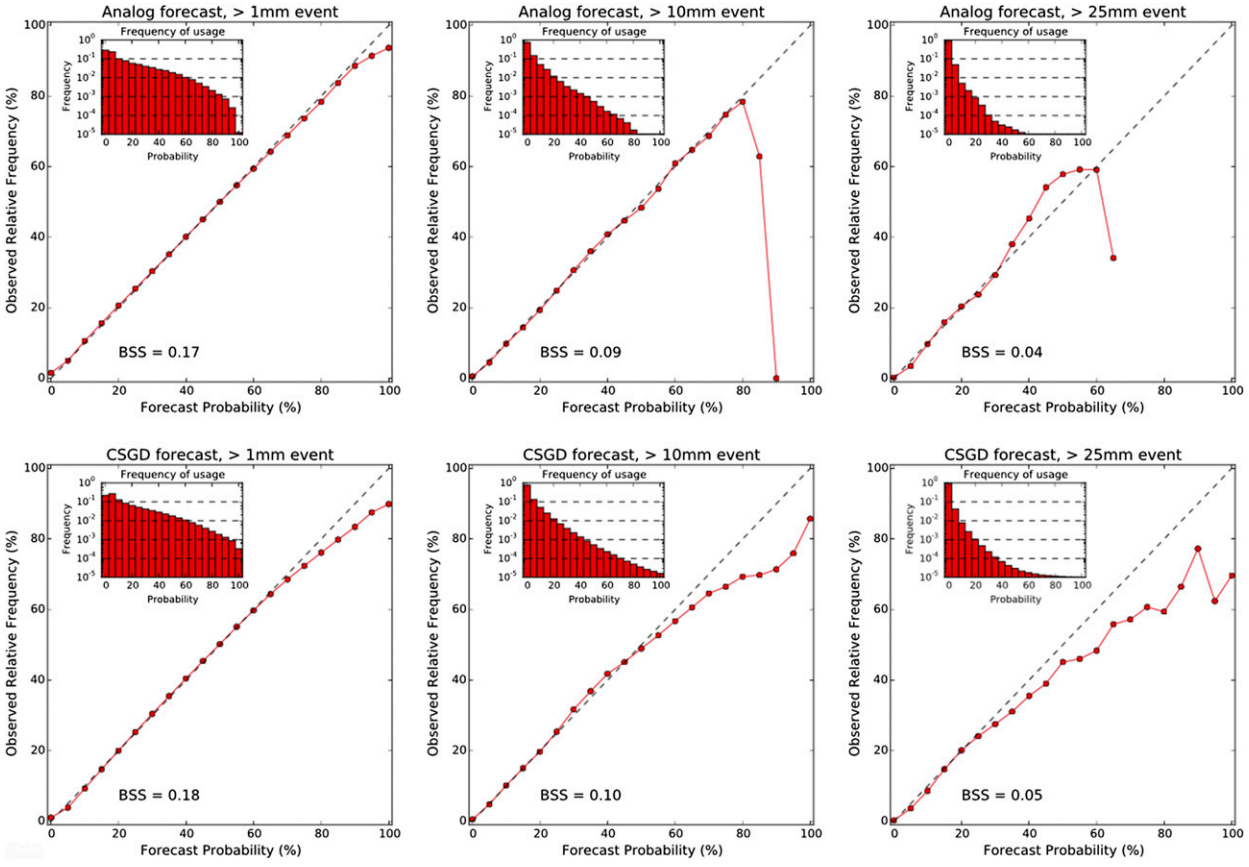

FIG. 11. As in Fig. 10, but for +108- to +120-h lead time.

before (11 training years for each verification year) for calculating the CCPA quantiles (section 4a) and fitting the unconditional CSGD model (section 4b).

Since fixing $\alpha_{7,s} = 0.5$ hardly affected the predictive performance with the large training dataset used above, and since the uncertainty parameter $\alpha_{8,s}$ is particularly prone to overfitting, we fit reduced CSGD regression models with $\alpha_{7,s} = 0.5$ and $\alpha_{8,s} = 0$ in the present setup. Even the estimation of the remaining six parameters might be difficult with only 91 or 273 training days (the majority of which are typically dry days). We, therefore, consider a further setup where we use again just 1 year/3 years of training data (forecasts) but supplement the dataset at each analysis grid point with data from 19 other analysis grid points with similar climatologies and terrain characteristics, and a certain minimal distance to each other to make sure that their forecast error characteristic are largely independent. A detailed description of the algorithm for selecting the supplemental locations is given in the paper by Hamill et al. (2015; also see appendix A in the online supplemental material), where supplemental locations are used quite successfully to improve the predictive performance of the analog method for higher precipitation events. In their setup, the supplemental data *complement* the reforecast data; here, we study how well data from other locations can *substitute* reforecast data.

For the calculation of the forecast quantiles as required for the quantile mapping step, there is no straightforward way to pool data across different grid locations. In this context we must hope that there is sufficient independent information in the ensemble (recall that all ensemble member forecasts are pooled for the purpose of calculating the forecast quantiles) to warrant an adequate estimation of the forecast climatology.

Figure 13 depicts the decrease of the Brier skill scores obtained with the setups described above compared to the full model fitted with 11 years of training data. The effects of reducing the training sample size are dramatic, especially for the prediction of the 25 mm $(12\,\text{h})^{-1}$ event. Brier skill scores for +12- to +24-h lead time go down by up to 0.1 when only one year of training data are used. Inspection of the corresponding reliability diagrams (see appendix C in the online supplemental material) reveals that the reliability of CSGD forecasts suffers

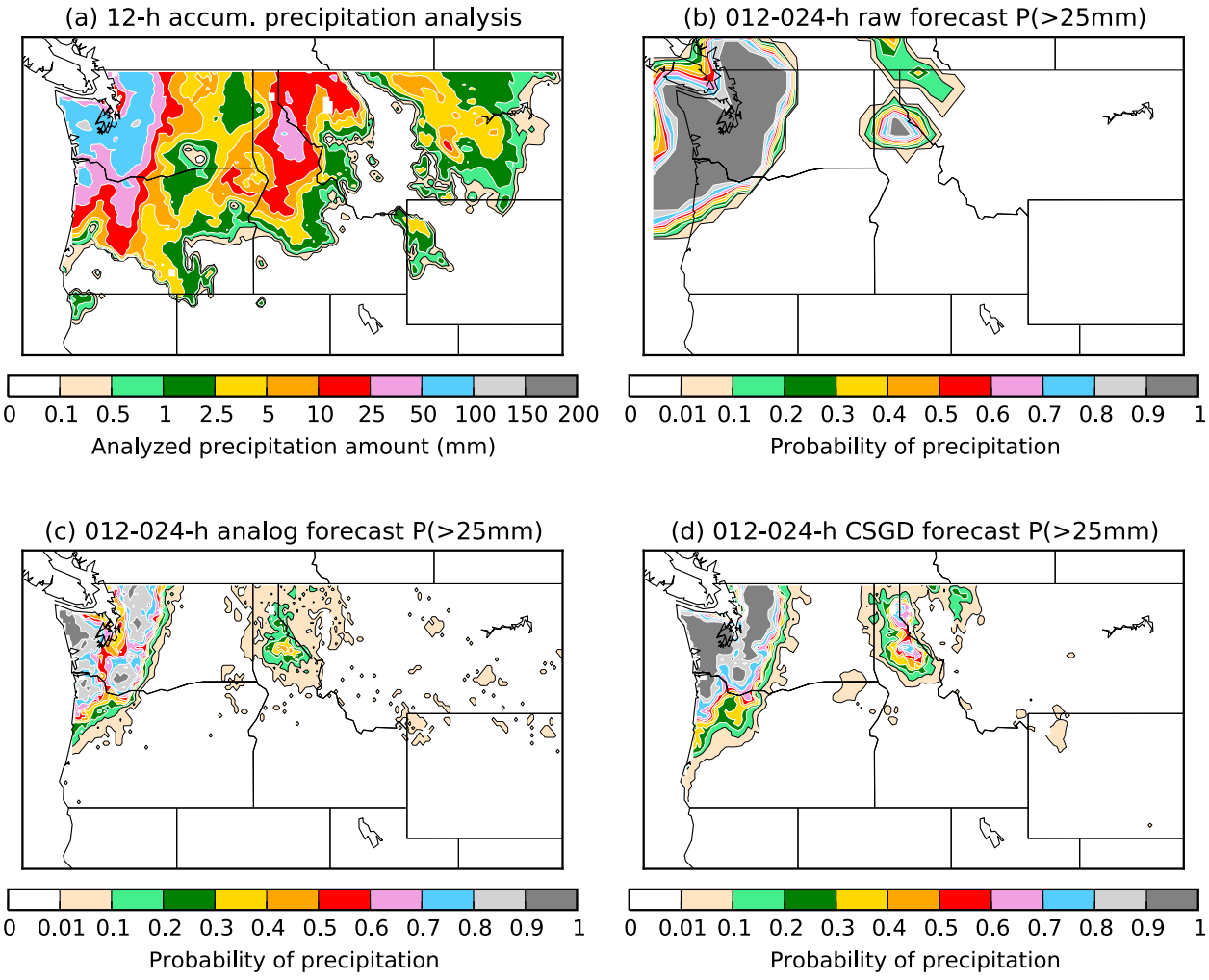

FIG. 12. (a) Analyzed precipitation between 1200 UTC 6 Nov and 0000 UTC 7 Nov 2006 and corresponding +12- to +24-h lead probability forecasts for exceeding 25 mm $(12\,\text{h})^{-1}$ of precipitation by (b) the raw ensemble, (c) the analog method, and (d) the CSGD regression approach.

substantially. As a result of overfitting, the predictive CSGDs become overconfident (underdispersive), and this overconfidence particularly affects the higher thresholds. The use of supplemental locations can mitigate but not entirely compensate for the lack of reforecast data; although the training dataset corresponding to the ''one year reforecast plus supplemental locations'' setup is almost twice as large as the training dataset with 11 years of reforecasts (but no supplemental data), the resulting CSGD predictions are still inferior. However, they nearly match at least the performance of the CSGD model fitted to a training dataset consisting of three years of reforecasts but no supplemental data. On the one hand, this highlights the benefits that a lengthy reforecast can provide, with its greater variety of weather events covered. On the other hand, it shows that the strategy of increasing the training dataset by considering supplemental locations can substantially reduce the performance loss due overfitting.

d. Role of the neighborhood size considered for the ensemble statistics

So far, all results for the CSGD method were obtained with a radius $r = 2°$ around each analysis grid point, within which forecasts were used as predictors. This is an ad hoc choice, and the question suggests itself as to how much of an impact the choice of the neighborhood size has on the predictive performance, and what the optimal radius would be for each lead time. To study this, we use again the maximal training dataset (11 years of forecasts and analyses), the full regression model (14) and (15), and calculate the CRPSS of the predictive CSGDs for different choices of r. The smallest possible radius $r = 0.5°$ (the resolution of the forecast grid) serves as a

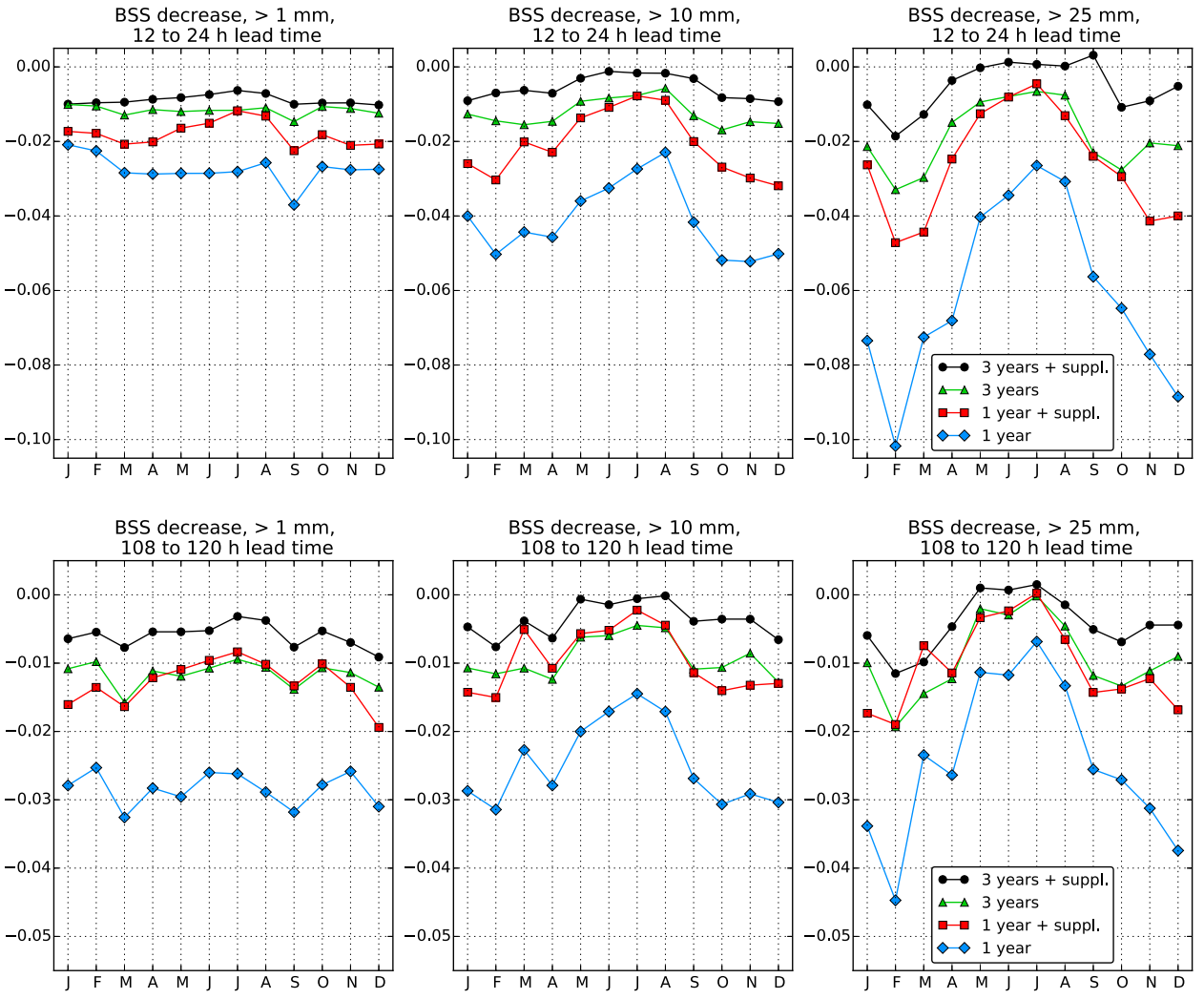

FIG. 13. Decrease of the Brier skill scores (aggregated over all analysis grid points within the CONUS) due to a reduction of the training sample to 1 or 3 years of training data. In both cases we also give results for CSGD distributions fitted with additional training data from 19 supplemental locations.

benchmark and corresponds to neighborhoods that only contain the closest forecast grid points. Extremely large neighborhoods were not tested due to the increased computational expense. In Fig. 14 we depict the change in CRPSS relative to this benchmark value for larger neighborhood sizes. As might be expected, the optimal radius changes with lead time: for the longest (days 4.5–5) lead time considered here, the largest radius $r = 3$ yields the best results, while for the shortest (days 0.5–1) lead time an initial increase in predictive performance is eventually reversed when r is increased beyond 2°. This case further shows that it is not just lead time, but more generally predictability that determines the optimal radius: the more predictable precipitation generating processes during the cool season favor smaller neighborhood sizes than the less predictable processes during the warm season. The overall increase of skill resulting from an adequate choice of r (larger than the minimal choice of $r = 0.5°$) is similar or even larger in magnitude than the increase resulting from more sophisticated regression equations as studied in section 5a.

6. Discussion

We have discussed a parametric postprocessing approach that uses statistics of the raw ensemble forecasts as predictors for the parameters of a censored, shifted Gamma distribution (CSGD). Exploratory analysis (see Figs. 3, 4, and 6) showed that CSGDs can approximate both climatological distributions of observed precipitation and distributions conditional on the ensemble forecasts reasonably well. Ensemble mean and

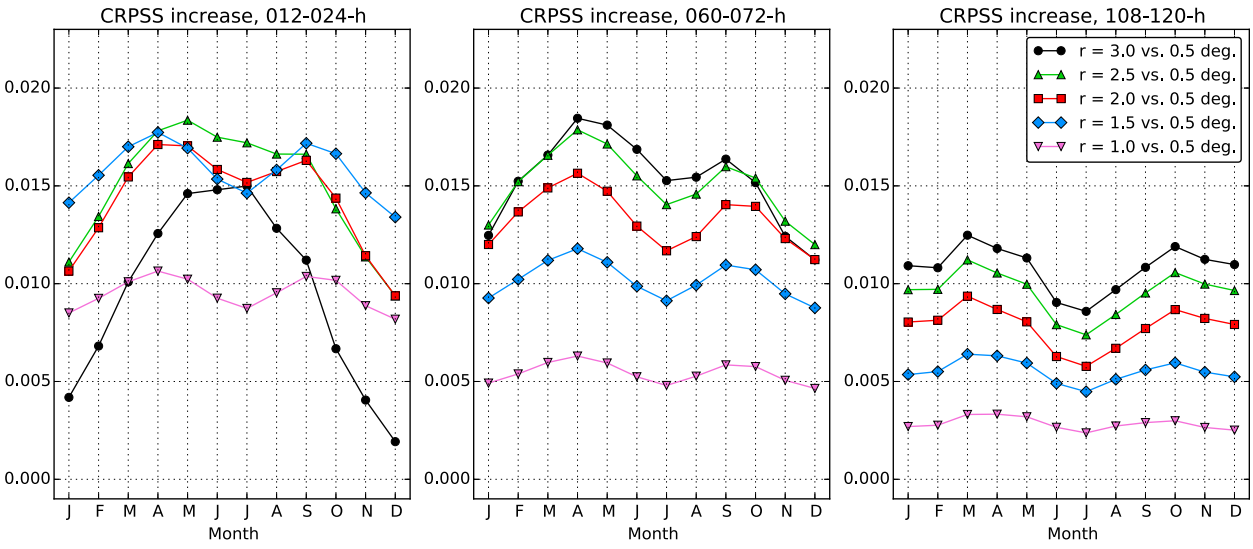

FIG. 14. Increase of CRP skill scores for different neighborhood sizes, relative to the smallest possible neighborhood with $r = 0.5°$.

dispersion predictors were estimated from the ensemble at the observation location and in a surrounding area. Ensemble mean precipitable water was used as a further predictor. These statistics were used to drive a heteroscedastic regression model, which was demonstrated to be capable of modeling the relation between ensemble forecasts and parameters of the predictive CSGDs. Verification results showed that the CSGD regression approach yields probabilistic forecast that were sufficiently reliable at all lead times and had better resolution than the forecasts obtained by a state-of-the-art analog approach. This was especially true for forecasts of extreme events, which are of particular interest due to their socioeconomic impact.

The CSGD approach presented here adopted the Scheuerer (2014) procedure of utilizing forecasts within a larger neighborhood of the location of interest as predictors. Accordingly, we also studied the connection between the optimal neighborhood size and predictive skill, finding that very large neighborhoods with a radius of over 300 km performed best with longer lead times and for predictions during the warm season. For short lead times and synoptic-scale winter precipitation, a smaller radius was more appropriate. The improvement in skill with larger neighborhoods compared to a model that only used forecasts at the nearest forecast grid points was similar in magnitude as the improvement due to more complex and flexible regression equations that permitted a nonlinear relation between the predictors and the predictive mean.

Finally, we studied the effect of training sample size on the predictive performance of the fitted CSGD model. For the analog method, a large training dataset is preferred because only then can good analogs always be found over all cases. The results presented here suggest that the predictive performance of a parametric approach also suffers substantially if the model is fitted with an insufficiently large dataset. Supplemental data from close-by grid points can partly compensate for a lack of reforecasts, but more efficient ways to share information between different locations need to be found to ensure good predictive skill of forecast of more extreme events even with a limited amount of reforecasts. These results affirm the positive value of lengthy training datasets that reforecasts can provide.

Acknowledgements. This research was supported by funding from the NOAA/National Weather Service Sandy Supplemental and NOAA/OAR High-Impact Weather Prediction Project, both funded via the Disaster Relief Appropriations Act of 2013.

REFERENCES

Ben Bouallègue, Z., 2013: Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Wea. Forecasting*, **28**, 515–524, doi:10.1175/WAF-D-12-00062.1.

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

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901, doi:10.1175/2009MWR3187.1.

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

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

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, doi:10.1175/2007MWR2411.1.

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

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

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, doi:10.1175/JHM-D-11-0140.1.

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.*, **123**, 2181–2196, doi:10.1175/1520-0493(1995)123<2181:MFEP>2.0.CO;2.

Messner, J. W., and G. J. Mayr, 2014: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, doi:10.1175/MWR-D-13-00271.1.

Molteni, F., R. Buizza, T. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, doi:10.1002/qj.49712252905.

Powell, M. J. D., 2014: On fast trust region methods for quadratic models with linear constraints. Tech. Rep. DAMTP 2014/NA02, Department of Applied Mathematics and Theoretical Physics, Cambridge University, 31 pp. [Available online at http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2014_02.pdf.]

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

Schmeits, M. J., and K. J. Kok, 2010: A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Mon. Wea. Rev.*, **138**, 4199–4211, doi:10.1175/2010MWR3285.1.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, doi:10.1175/MWR3441.1.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl, 2013: Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncertainty Quantif.*, **1**, 522–534, doi:10.1137/130907550.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, doi:10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, doi:10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, doi:10.1002/met.134.

——, 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. International Geophysics Series, Vol. 100, Elsevier Academic Press, 704 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, doi:10.1175/MWR3402.1.