Probabilistic Precipitation Forecast Postprocessing Using Quantile Mapping and Rank-Weighted Best-Member Dressing

THOMAS M. HAMILL

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

MICHAEL SCHEUERER

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

(Manuscript received 24 April 2018, in final form 23 August 2018)

ABSTRACT

Hamill et al. described a multimodel ensemble precipitation postprocessing algorithm that is used operationally by the U.S. National Weather Service (NWS). This article describes further changes that produce improved, reliable, and skillful probabilistic quantitative precipitation forecasts (PQPFs) for single or multimodel prediction systems. For multimodel systems, final probabilities are produced through the linear combination of PQPFs from the constituent models. The new methodology is applied to each prediction system. Prior to adjustment of the forecasts, parametric cumulative distribution functions (CDFs) of model and analyzed climatologies are generated using the previous 60 days' forecasts and analyses and supplemental locations. The CDFs, which can be stored with minimal disk space, are then used for quantile mapping to correct state-dependent bias for each member. In this stage, the ensemble is also enlarged using a stencil of forecast values from the 5×5 surrounding grid points. Different weights and dressing distributions are assigned to the sorted, quantile-mapped members, with generally larger weights for outlying members and broader dressing distributions for members with heavier precipitation. Probability distributions are generated from the weighted sum of the dressing distributions. The NWS Global Ensemble Forecast System (GEFS), the Canadian Meteorological Centre (CMC) global ensemble, and the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble forecast data are postprocessed for April–June 2016. Single prediction system postprocessed forecasts are generally reliable and skillful. Multimodel PQPFs are roughly as skillful as the ECMWF system alone. Postprocessed guidance was generally more skillful than guidance using the Gamma distribution approach of Scheuerer and Hamill, with coefficients generated from data pooled across the United States.

1. Introduction

The U.S. National Weather Service (NWS) recently instituted a program to generate multimodel ensemble postprocessed guidance for initializing its National Digital Forecast Database (NDFD; Glahn and Ruth 2003). The NDFD data provide high-resolution (2.5-km grid spacing) guidance over the contiguous United States, Alaska, Hawaii, and Puerto Rico. Its data can be found on NWS forecast office web pages and underlie the generation of its worded forecasts. The program to generate postprocessed guidance to initialize the NDFD is known as the National Blend of Models, or National Blend. A recent article by Hamill et al. (2017, hereafter H17) described an initial procedure for generation of deterministic 6-h quantitative precipitation forecasts (QPFs) and 12-h probability of precipitation (POP) that was made operational in late 2017 in the National Blend for medium-range forecasts. Aspects of the H17 postprocessing system that were novel or somewhat novel included the following:

1) Increasing the training sample size by augmenting the training data at a given grid point with data from other grid points with similar terrain and precipitation

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-18-0147.s1.

Corresponding author: Dr. Thomas M. Hamill, tom.hamill@noaa.gov

DOI: 10.1175/MWR-D-18-0147.1

climatology characteristics; this was called the ''supplemental location'' process.
2) Synthetically enlarging the multimodel ensemble size and addressing distributional bias by quantile mapping the precipitation forecast data from surrounding grid points, with the surrounding grid point's forecasts quantile mapped to be consistent with the center point's analyzed climatology.
3) Adding state-dependent random noise to each member to increase the spread, decrease forecast overconfidence, and improve reliability.
4) Decreasing spatial sampling variability through a terrain-roughness-dependent Savitzky–Golay smoothing (Press et al. 1992, section 14.8) of the resulting POPs.

Though H17 showed that postprocessed QPF guidance from the combination of Canadian Meteorological Centre (CMC) and National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) was skillful, and POPs were skillful and also reliable, there were several reasons to consider further modifications to the procedure. First, the postprocessing algorithm of H17 combined information from all potential prediction systems at an early stage of the processing, forming a superensemble of quantile-mapped amounts. Such a procedure, especially applying data-informed weighting techniques discussed below, would be challenging if the size of the ensemble varied from one day to the next as a result of data delays or data outages. Missing data are more likely to occur when the system includes predictions from other operational centers in an NWS production environment with strict data cutoff times. An alternative to H17 to be evaluated here is thus whether acceptable results can be obtained through a two-step postprocessing procedure, where each prediction system is postprocessed individually, and then resulting probabilities are linearly combined. In situations where guidance produced by such prediction systems is relatively independent, a further adjustment, such as suggested by Gneiting and Ranjan (2013), may provide an even better result.

A second deficiency was that the H17 procedure did not produce probabilistic forecasts for higher precipitation amount thresholds that were as reliable as the POP forecasts. A likely cause of the unreliability was that the artificial noise added to ensemble members only partly addressed the remaining issues of overconfidence in the enlarged, quantile-mapped ensemble.

To address this, we consider more objectively based algorithms for adjusting the probabilities than the addition of state-dependent noise in H17. Specifically, we consider variants on the approach known as ''best-member dressing'' (Roulston and Smith 2003). In standard dressing procedures, multiple realizations of noise are added to each member forecast, with the magnitude of the added noise consistent with the amount needed to ensure consistency between the ensemble-mean root-mean-square (RMS) error and the ensemble spread. The resulting ensemble had larger spread, and probabilities from the ensemble exhibited skill and improved reliability. An examination of subsequent literature, most notably Fortin et al. (2006; hereafter F06), suggested that an ensemble-weighting procedure may be able to improve upon the basic Roulston and Smith (2003) algorithm. The underlying concept discussed by F06 is as follows: except in the case of two ensemble members with the same value, such as both with zero precipitation, only one ensemble member will commonly have a value closest to the eventual analyzed state. The user will not know which one beforehand, but given a training dataset of previous cases of ensemble forecasts and the associated verification, it is possible to sort the ensemble, increment a counter associated with the rank of the closest member, and repeat the process over many past forecast dates and grid points. The resultant ''closest-member histogram'' is very similar to the rank-histogram concept discussed by Hamill (2001). The closest-member histogram statistics provide the necessary data for an objective reweighting of the sorted ensemble of forecasts before dressing and determining the probabilities. For example, perhaps the highest- and lowest-sorted ensemble members would be more highly weighted, given the overconfidence typical in ensemble prediction systems and the greater probability the analyzed state lies beyond the range of the ensemble (Hamill and Colucci 1998).

Inspired by F06, this article will examine whether this reweighting produces forecasts with improved skill and reliability relative to the performance benchmark set by H17. This article will also investigate how skill changes when European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble predictions are included in a multimodel ensemble from GEFS and CMC data. Past studies such as Hagedorn et al. (2008, 2012), Hamill et al. (2008), and Hamill (2012) have shown that ECMWF predictions, after their own postprocessing, set a high benchmark for skill, one that is hard to surpass even with postprocessed, multimodel guidance that includes ECMWF. This tentative conclusion will be reexamined with postprocessing that incorporates the existing quantile mapping together with the reweighting procedure suggested by F06.

The remainder of the article is organized as follows. Section 2 provides a brief description of the datasets and evaluation methodologies used in this article, which are mostly the same as in H17. Section 3 describes the modifications to the postprocessing procedure that will be evaluated here. Section 4 provides results, and section 5 discusses them and makes recommendations.

2. Datasets and evaluation methodologies

So that prediction results can be compared as directly as possible against the results discussed by H17, nearly the same verification data period is used, namely, forecasts initialized at 0000 UTC from 1 April to 30 June 2016. NCEP data were not available on 24, 25, 28, 29, and 30 June, so these dates were omitted for all systems. However, when including the training data, data were downloaded for the period back to 1 December 2015. As will be explained more in section 3, training data are needed for quantile mapping and for the estimation of "closest-member histograms." For a date of interest, data 61–120 days prior were used in the development of closest-member histograms. Data 1–60 days prior were used for the development of cumulative distribution functions used in the quantile mapping of the forecast during the verification period.

Precipitation forecast data during this period were obtained from the NCEP GEFS, the CMC ensemble, and, in this study, also from the ECMWF ensemble prediction system. These ensembles will be referred to simply as NCEP, CMC, and ECMWF, respectively. Data were downloaded from ECMWF's THORPEX Interactive Grand Global Ensemble (TIGGE; Bougeault et al. 2010; Buizza 2014; Swinbank et al. 2016) data portal. Twelve-hourly accumulated precipitation forecast data were downloaded at ½° grid spacing on a grid surrounding the contiguous United States (CONUS) and then bilinearly interpolated to the ⅛° grid of the analyzed data. Details on the NCEP and CMC ensembles were provided by H17. Details on the ECMWF ensemble in 2016 and its performance were documented by Haiden et al. (2016). Given that data were requested from the TIGGE portal at ½° and the ensemble prediction systems have higher native resolution, the lack of reliability of the raw ensembles, discussed later, is probably somewhat exaggerated through use of the degraded-resolution data.

As in H17, climatology-calibrated precipitation analysis data (Hou et al. 2014) at ⅛° grid spacing and 12-hourly temporal resolution over the CONUS are used for verification and training.

The evaluation metrics are the same as used by H17. In particular, Brier skill scores (BSS) and reliability diagrams will be the primary methods for diagnosing the raw and postprocessed guidance quality, calculated using data across all 1/8° grid points inside the CONUS and the Columbia River basin of Canada. Confidence intervals for selected tests are provided, as well as whether the test passed the 5% significance level. The procedure for developing confidence intervals follows the block bootstrap algorithm described by Hamill (1999). A case study will be included that also visually illustrates the characteristics of the guidance, from raw-model guidance through calibration and combination.

3. Description of the revised postprocessing procedure

a. *Review of the previously used procedure*

We start with a brief review of the postprocessing procedure for probabilistic precipitation forecasts in H17.

Before any postprocessing occurred, for each grid point in the CONUS and for each month of the year, a set of "supplemental locations" had been determined. These locations were chosen based on similarity of terrain features and precipitation climatology. A minimum distance between supplemental locations was enforced so that training samples would have greater independence. Forecast and analyzed data at the supplemental locations were then used to populate the empirical cumulative distribution functions (CDFs) for precipitation that were used in the quantile mapping. Postprocessing was then performed grid point by grid point. First, quantile mapping was applied to each ensemble member to make its forecast more consistent with a draw from the analyzed precipitation climatology. In this step, the ensemble forecast was also synthetically enlarged ninefold by quantile mapping a 3×3 stencil of surrounding grid points' forecasts using each grid point's forecast distribution and the center grid point's analyzed distribution. Again, see H17 for details and figures that illustrate this procedure and provide more rationale for its use. The ninefold enlarged, quantile-mapped ensemble at this grid point was combined with the ninefold enlarged and quantile-mapped ensemble members from other prediction systems. A single realization of random Gaussian noise was added to each member to increase spread, a simplified form of dressing. The magnitude of this noise applied to a particular member was linearly related to that member's quantile-mapped precipitation amount, with larger noise associated with larger amount forecasts; this procedure is admittedly ad hoc, but it was inspired by previous experiments with precipitation forecast calibration such as Hamill and Colucci (1998), who showed that precipitation uncertainty tends to be linearly related to precipitation amount. Probabilities were then determined from the ensemble relative frequency, and as a final step, the gridded field of probabilities was smoothed using a Savitzky–Golay smoother.

b. *The new postprocessing procedure, with rank weighting of sorted members*

The major changes incorporated into the revised algorithm are now described, first at a high level and then

followed by a detailed description with equations and figures as needed. The changes include the following:

1) Postprocessing is applied separately to guidance from each prediction system.
2) CDFs for the forecast and analyzed distributions used in the quantile mapping are now estimated with a "fraction zero" and Gamma distributions (Wilks 2011) for positive amounts instead of empirical distributions. This revised approach radically shrinks the amount of training data information that needs to be stored prior to generation of the postprocessed guidance, and the algorithm runs more quickly. There is also a new adjustment to the quantile-mapping procedure to constrain the extent of forecast adjustment for precipitation amounts that are large relative to the gridpoint's climatological mean, presuming in such circumstances the quantile-mapping estimates are subject to larger sampling errors.
3) As the postprocessing for a particular ensemble system proceeds grid point by grid point, the previous synthetic ensemble enlargement and quantile mapping using a 3×3 stencil of surrounding grid points in H17 is replaced with a 5×5 stencil. This contributes to reduced sampling variability at each grid point and smoother spatial maps of ensemble probabilities, especially if the technique is applied to generate probabilistic forecasts from a deterministic prediction. As a consequence of this and the more sophisticated dressing procedure in the subsequent step that provides some noise reduction, the Savitzky–Golay smoothing is omitted.
4) Formation of a postprocessed forecast CDF through the summation of objectively weighted Gaussian dressing CDFs, as opposed to the H17 algorithm of adding one realization of random noise to each quantile-mapped member and forming probabilities from ensemble relative frequency.
5) Determination of the objective weights in step 4 with "closest-member histograms."
6) When multimodel ensemble probabilities are desired, the final product is generated from a weighted combination of the single-model postprocessed PQPFs.

We now describe each of the algorithm revisions in more detail.

1) Independent processing of each model

The first algorithmic revision is straightforward; each prediction system is processed independently.

2) Changes to the quantile-mapping procedure

The second revision is to fit parametric forecast and analyzed CDFs to be used in the quantile mapping, as opposed to the empirical CDFs used by H17. Use of parametric instead of empirical CDFs saves much disk storage of the training data. Three CDF parameters are estimated: a fraction zero (FZ) and the shape α and scale β of a Gamma distribution for positive precipitation amounts. These parameters are fit separately for analyzed and forecast data. The parameters are estimated individually for each grid point using the data from that grid point and from the supplemental locations using the previous 60 days of forecasts and analyses. Assume we have m samples to estimate the parameters of a variable y, which could be the quantile-mapped ensemble forecast information or analyzed information for a particular grid point. Define an indicator function I for whether the ith of the m samples of y is greater than zero:

$$I(i) = \begin{cases} 0, & \text{if} \quad y_i = 0 \\ 1, & \text{if} \quad y_i > 0 \end{cases}. \tag{1}$$

Then, the estimated fraction zero parameter $\hat{F}Z$ is estimated from the relative frequency of zeros in the sample:

$$\hat{F}Z = 1 - \frac{\sum_{i=1}^{m} I(i)}{m}. \tag{2}$$

Suppose from the original m samples of y, we have a set of n remaining samples with positive precipitation amounts, which we denote as y^+. For samples with nonzero precipitation, α and β are estimated using the method of maximum likelihood and the Thom (1958) estimator as described by Wilks (2011, section 4.4.3). The sample statistic D is calculated as

$$D = \ln(\overline{y}^+) - \frac{1}{n}\sum_{i=1}^{n} \ln(y_i^+) = \ln\left(\frac{1}{n}\sum_{i=1}^{n} y_i^+\right) - \frac{1}{n}\sum_{i=1}^{n} \ln(y_i^+), \tag{3}$$

where the overbar denotes an arithmetic average. The appealing characteristic of estimating CDFs with a parametric distribution is that minimal storage is required, so the parameters can be estimated rapidly. For each of the preceding 60 days and each grid point (including data from the supplemental locations), we tally m, n, $\sum_{i=1}^{n} y_i^+$, and $\sum_{i=1}^{n} \ln(y_i^+)$. Using this, we can sum the appropriate information over the 60 training days, generate the D statistic from Eq. (3), and then estimate fitted parameters $\hat{F}Z$, $\hat{\alpha}$, and $\hat{\beta}$:

$$\hat{\alpha} = 1 + \frac{\sqrt{1 + 4D/3}}{4D}, \tag{4}$$

and

$$\hat{\beta} = \overline{y}^+ / \hat{\alpha}. \tag{5}$$

Quantile mapping in most circumstances then proceeds as described by H17's Eqs. (8) and (9). However, one final modification has been made to the quantile-mapping procedure. Suppose the precipitation forecast for a particular grid point is unusually large relative to that point's climatology as expressed by the forecast CDF. In such a circumstance, we may not have sufficient trust that the tails of the fitted gamma distributions and the resulting mapping functions are adequate. In this case, we use a slight modification of the procedure described by Scheuerer and Hamill (2015, their appendix A). Under that procedure, if the nonexceedance probability of today's forecast relative forecast's climatological CDF exceeds 0.9, a regression slope correction b is applied to estimate the quantile-mapped values. Let x_i^f be the ith member's raw forecast amount, and let $[q_{0.90}^f, \ldots, q_{0.99}^f]$ and $[q_{0.90}^a, \ldots, q_{0.99}^a]$ represent vectors of the quantiles associated with the 90th–99th quantiles of the forecast and analyzed distribution every 1%. The ith quantile-mapped forecast $\tilde{x}_i{}^f$ then is

$$\tilde{x}_i^f = \begin{cases} q_{0.90}^a + b\,(x_i^f - q_{0.90}^f) & \text{if} \quad q_{0.90}^f \le x_i^f < q_{0.99}^f \\ q_{0.90}^a + b\,(q_{0.99}^f - q_{0.90}^f) + (x_i^f - q_{0.99}^f) & \text{if} \quad x_i^f \ge q_{0.99}^f \end{cases}. \tag{6}$$

In other words, if the forecast is between the 90th and 99th percentiles of the forecast CDF, a straightforward regression slope correction is applied following Scheuerer and Hamill (2015). If the forecast is beyond the 99th percentile, the difference between today's forecast and the 99th percentile of the forecast distribution is also added. This permits extremely large forecast values to retain some of their anomalous nature but to retain a bias correction estimated for data between the 90th and 99th percentiles.

3) Use of 5 × 5 stencil of surrounding grid points

We now consider the postprocessing of a particular single ensemble prediction system at one grid point in the domain of interest and at one particular lead time. To deal with systematic position errors and synthetically increase the sample size, the H17 procedure enlarged the ensemble by quantile mapping a surrounding 3 × 3 stencil of grid points, using the CDF unique to each stencil point and the analysis CDF at the central point of interest. In the revised procedure, a 5 × 5 stencil is used, providing a larger sample. For subsequent testing against the previous version of the algorithm, the distance between grid points in the H17 3 × 3 stencil is double that of the current 5 × 5 stencil, which ensures that the 3 × 3 and 5 × 5 stencils cover the same area, just with a denser grid for the 5 × 5 stencil. The spacing between grid points in the stencil depends on the forecast lead time, increasing linearly from 1/8° grid for +12-h lead forecasts to 5/8° at +156–168 h. The increase of spacing with lead time is an ad hoc way of dealing with the potential increase of position biases in ensemble systems with increasing lead time (Scheuerer and Hamill 2015, Fig. 14).

4) Dressing with weighted CDFs

The revised dressing procedure at a particular grid point and lead time is now described. The vector $\tilde{\mathbf{x}}^f$ represents sorted, quantile-mapped, 25-fold enlarged ensemble members at a grid point of interest:

$$\tilde{\mathbf{x}}^f = [\tilde{x}_{(1)}^f, \ldots, \tilde{x}_{(n\times 25)}^f]\,. \tag{7}$$

These provide estimates of the random variable x, the unknown true precipitation amount.

The index subscript (i) in Eq. (7) now denotes the ith-*sorted* member. The mean of the quantile-mapped and enlarged forecasts will also be used to later set the value of an index in the closest-member histogram:

$$\tilde{x}^f = \frac{1}{n\times 25}\sum_{(i)=1}^{n\times 25} \tilde{x}_{(i)}^f\,. \tag{8}$$

The CDF $\Phi(x)$ of postprocessed precipitation amount is estimated through a weighted combination of Gaussian-distributed dressing cumulative probability distributions associated with each sorted ensemble member:

$$\Phi(x) = \sum_{(i)=1}^{n\times 25} h_{(i)} \times \Phi_N\left[\frac{(x - \tilde{x}_{(i)}^f)^2}{2\sigma_{(i)}^2}\right], \tag{9}$$

where $h_{(i)}$ is the "closest-member histogram" weight associated with the ith-sorted member, described in more detail later. Parameter $\Phi_N[.]$ in Eq. (9) is the Gaussian-distributed dressing CDF for the ith-sorted member, a distribution whose associated PDF is centered on the sorted, quantile-mapped member with associated standard deviation

$$\sigma_{(i)} = \begin{cases} 0, & \text{if} \quad \tilde{x}^f_{(i)} = 0 \\ 0.15 + \dfrac{\tilde{x}^f_{(i)}}{0.15}, & \text{if} \quad \tilde{x}^f_{(i)} > 0 \end{cases}. \tag{10}$$

For amounts greater than zero, standard deviation starts at an initially small nonzero value and increases linearly with precipitation amount. The chosen standard deviation of the distribution is admittedly ad hoc, but through extensive testing, including the objective fitting of Gamma dressing distributions (not shown), it was determined that the results are not very sensitive to the choice of dressing distribution parameters for the ensemble systems examined here.

5) Estimation of the closest-member histogram weights

Consider now how the closest-member histograms are estimated. The vector of closest-member histogram weights

$$\mathbf{h} = [h_{(1)}, \ldots, h_{(n\times 25)}] \tag{11}$$

are estimated directly from quantile-mapped training data during the past 60 days, accumulated over the CONUS. To permit a dependence of the closest-member histogram weights on precipitation amount, the ith-sorted member's weight is estimated as a function of the rank of the sorted member and an index $M(\overline{\tilde{x}}^f)$ of the quantile-mapped mean precipitation amount:

$$h_{(i)} = \mathcal{H}[(i), M(\overline{\tilde{x}}^f)], \tag{12}$$

where

$$M(\overline{\tilde{x}}^f) = \begin{cases} 1, & \text{if} \quad \overline{\tilde{x}}^f \le 0.01\,\text{mm} \\ 2, & \text{if} \quad 0.01 \le \overline{\tilde{x}}^f < 2.0\,\text{mm} \\ 3, & \text{if} \quad 2.0 \le \overline{\tilde{x}}^f < 6.0\,\text{mm} \\ 4, & \text{if} \quad 6.0\,\text{mm} \le \overline{\tilde{x}}^f \end{cases}. \tag{13}$$

In the case where $M(\overline{\tilde{x}}^f) = 1$, the weights for each member are set equally to $1/(n \times 25)$. When $M(\overline{\tilde{x}}^f) > 1$, the statistics are estimated objectively from the closest-member histogram statistics, which are stratified by the quantile-mapped ensemble-mean amount. When tallying closest-member histogram statistics, should one or more quantile-mapped members and the analyzed state have the same value such as zero, the closest-member rank is assigned randomly between the sorted members with equal values.

Figure 1 provides an example of the closest-member histogram statistics, in this case for the training data for quantile-mapped and 5×5 stencil-enlarged ensembles for +36–48-h forecasts with an initial date of 0000 UTC 1 May 2016. To permit three prediction systems with ensembles of different numbers of members to be plotted on the same axis, the closest-member histogram abscissa is scaled 0 to 1 for each system. The ECMWF system's histograms are displaced below the CMC and NCEP ensembles because of the larger number of ensemble members in the ECMWF system and lower expected fraction per bin.

The closest-member histograms illustrate the potential advantage to be gained by weighting the members based on both the members' sorted rank (the abscissa) and the mean precipitation amount (the panel in Fig. 1). No matter the precipitation amount, the sorted members with extreme ranks of each system were generally more likely to be the closest member, to varying degrees. For heavy precipitation in Fig. 1c, the lowest-ranked member in the NCEP ensemble was ~8.4% likely to be the closest member, which was approximately two orders of magnitude more likely than some of the upper-ranked members. The shapes of the histograms also varied substantially with the mean precipitation amount. The largest weights (histogram values) with light mean precipitation in Fig. 1a were applied to the top-ranked members, whereas the largest weights with heavy mean precipitation in Fig. 1c were applied to the lowest-ranked members. This indicates a general tendency of the quantile-mapped forecasts to underforecast precipitation amounts with light-mean-forecast precipitation and to overforecast precipitation amounts with heavy-mean-forecast precipitation, though it is noted that stratification of data can sometimes lead to misleading results (Siegert et al. 2012). The NCEP system had more weight for outlying members, even after accounting for differences in ensemble size. This indicates overconfidence (lack of spread) in the NCEP ensemble, even after quantile mapping and the 25-fold expansion. General characteristics of these closest-member histograms were similar at other leads, though the greater weight at the extreme ranks was typically more severe for shorter-lead forecasts than for longer-lead ones (not shown); though shorter-lead forecasts had lower ensemble-mean errors, they were also more overconfident. This may be a consequence of model spinup issues or suboptimal ensemble design.

Extensive testing was performed to develop a method for objectively estimating the characteristics

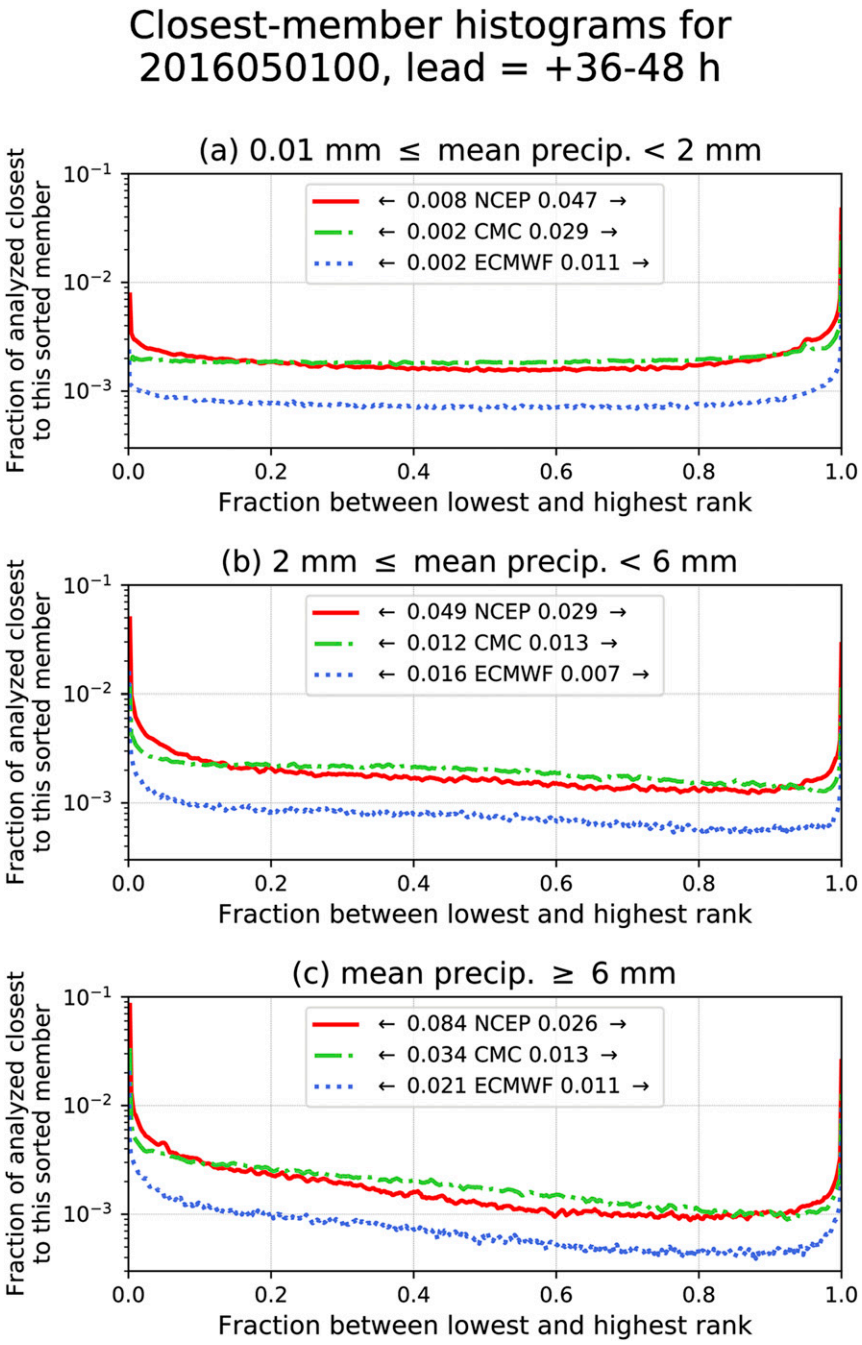

FIG. 1. Closest-member histograms for +36–48-h quantile-mapped precipitation forecasts valid for initial time of 0000 UTC 1 May 2016. The abscissa is the fraction between the lowest and highest ranks. The three panels are for indices of M, conditioned on the mean precipitation amount, from lighter to heavier. The fractional values associated with the lowest and highest ranks are indicated by the text in the legend. Interior ranks of the histograms were smoothed with a Savitzky–Golay smoother using a window length of nine ranks and fitting the coefficients of a second-order polynomial.

of Gamma-distributed dressing distributions. Upon comparison with much simpler, ad hoc Gaussian-distributed dressing distributions, only small differences in skill and reliability were found. Accordingly, the algorithm here uses the simpler Gaussian dressing distributions, and a detailed description of the method for generating objective Gamma-distributed dressing distributions is not included here.

6) Estimation of multimodel ensemble probabilities

The final algorithmic change relative to H17 is that the final estimate of the probabilities of exceeding an event threshold are generated from a weighted linear combination of the event probabilities estimated from each prediction system. The weights that are applied in the later figures of this paper are somewhat arbitrary, but systems with larger postprocessed skill receive higher weights.

c. *The censored, shifted Gamma distribution method of postprocessing*

The probabilistic forecasts obtained with the new weighted kernel dressing approach described above are compared with those obtained with the heteroscedastic regression approach based on censored, shifted Gamma distributions (CSGDs) proposed by Scheuerer and Hamill (2015). The CSGDs define a three-parameter distribution family that is able to model the occurrence and amount of precipitation simultaneously. Unlike the kernel dressing methodology described above, the CSGD approach uses three statistics (ensemble probability of precipitation, ensemble mean, and ensemble-mean absolute difference) to summarize the information in the quantile-mapped ensemble rather than the individual ensemble member forecasts. Nonlinear regression equations link the predictive CSGD parameters to the ensemble statistics, and the parameterization is chosen in such a way that the calibrated forecast distribution converges to the climatological distribution of the analyzed precipitation amounts as the skill of the underlying NWP forecasts tends to zero.

With large datasets and locally fitted CSGD coefficients, this approach was demonstrated to yield reliable and highly skillful probabilistic forecasts. A recent study (Zhang et al. 2017) confirmed these conclusions and found that the CSGD approach compares favorably with the mixed-type meta-Gaussian distribution (MMGD) model, which has been an integral part of the National Weather Service's Hydrologic Ensemble Forecast System. In all studies where the CSGD method was tested so far, however, a reforecast dataset was available and provided a sufficiently large training sample for locally fitting the CSGD model parameters. In the present setup, where the training sample size is limited, a number of modifications of the original approach described by Scheuerer and Hamill (2015) are required to prevent overfitting (see, e.g., their Fig. 13, for an illustration of the adverse effects of overfitting). As applied here, the CSGD significantly reduces the total number of model parameters that need to be estimated by assuming the regression parameters are constant across the CONUS domain. A spatially varying predictor of NWP model skill is introduced in addition to the spatially varying climatology parameters to address local forecast characteristics. Technical details about these modifications are provided in the online supplemental material to this paper. Several variants of this chosen algorithm were also tried; descriptions of these alternatives are omitted, given their somewhat reduced skill and reliability.

4. Verification of probabilistic forecasts

Figures 2 and 3 show the skill of constituent center's predictions for the POP and 10-mm threshold, respectively, as modifications are sequentially added to the base H17 algorithm for each model. Table 1 describes the various experiments that are plotted in these figures and the abbreviations used in the figure captions. Both figures illustrate that NCEP raw guidance was less skillful than either CMC or ECMWF guidance and is improved more through the postprocessing. Whereas for POP, at shorter leads the CMC guidance was more skillful than ECMWF, at longer leads and for the 10-mm threshold, ECMWF guidance was more skillful. Figure 2 also shows the profound impact of quantile mapping of light precipitation with the surrounding stencil of grid points, with especially pronounced impact for the NCEP system and the ECMWF system at the earlier forecast leads. In comparison, the other improvements, such as adding dressing and closest-histogram weighting, had much smaller impact for POP than the quantile mapping. They had more of an impact for the more poorly performing NCEP system and virtually no impact with the weighed multimodel ensemble. However, examining skill for the $\geq$10 mm 12 h^{-1} event in Fig. 3, we see the positive impact of the closest-histogram weighting, which addresses remaining issues of forecast overconfidence. At these higher amounts, quantile mapping had a smaller impact relative to the closest-member histogram rank-based weighting. Apparently, both quantile mapping and the closest-histogram weighting were necessary to achieve significant skill improvements simultaneously for both smaller and larger events. The primary deficiency at light precipitation amounts was apparently bias, addressed through the quantile mapping, and the primary deficiency of forecasts of heavier amounts was overconfidence, addressed through the closest-histogram rank weighting.

What if the data from one prediction system were not available, perhaps due to a communications outage? In this case, predictions would be made through a linear combination of the remaining data. Figure 4 shows the skill of forecasts when all of the systems were available

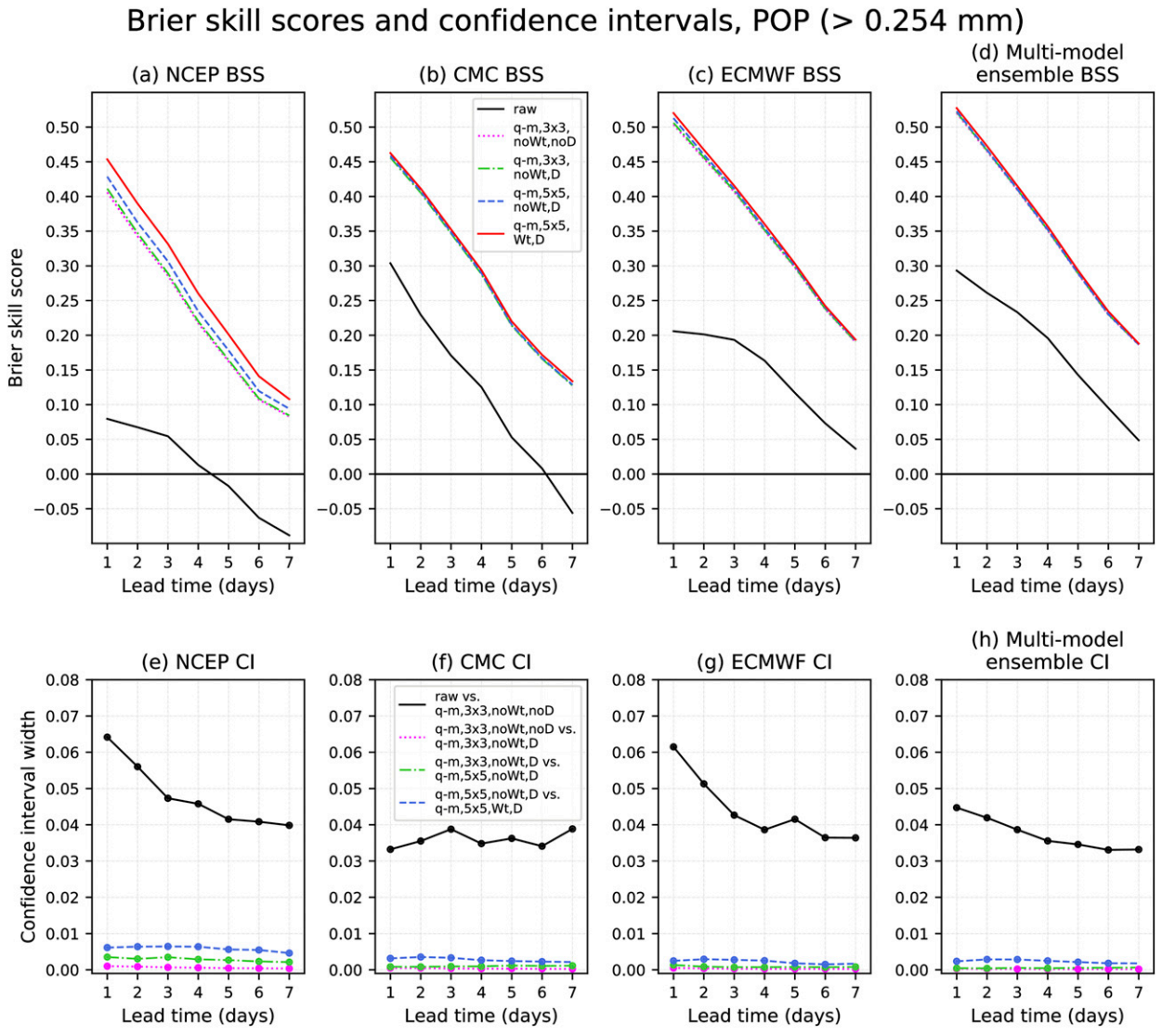

FIG. 2. BSSs for exceeding the POP threshold for various postprocessing configurations and as a function of lead time. (a) NCEP, (b) CMC, (c) ECMWF, and (d) multimodel ensemble with 20% weight for NCEP, 30% weight for CMC, and 50% weight for ECMWF. The experiment configurations are described in Table 1. (e)–(g) The absolute difference between the 50th and 95th percentiles of a block bootstrap distribution (i.e., a confidence interval for the hypothesis test of each forecast modification relative to the step before). Dots indicate that the difference was statistically significant at the 5% level.

and also when one system was missing. Confidence intervals are not presented, given the large number of possible permutations. The relative weights assigned were based roughly on the forecast accuracy, and weights are indicated in the figure legend. For POP, if ECMWF data were missing, there was a degradation of skill amounting to roughly ½ day of forecast lost lead time. The loss of data from either the CMC or the NCEP system was relatively unimportant when ECMWF data were available. For the $\geq$10 mm 12 h^{-1} event, forecasts again were most profoundly affected by the loss of ECMWF data, and forecasts were actually improved in skill very slightly when NCEP data are not used. This suggests that after postprocessing, the NCEP data did not provide much information that was independent of the information already provided by the postprocessed CMC and ECMWF systems or that was much less accurate. With a major overhaul of the NCEP prediction systems in 2018/19 and incorporation of a new dynamical core, one should not expect this characteristic to continue indefinitely.

Figures 5 and 6 provide comparison against another reference standard, the CSGD methodology of Scheuerer and Hamill (2015). In most situations, the quantile mapping and closest-histogram weighted dressing algorithm

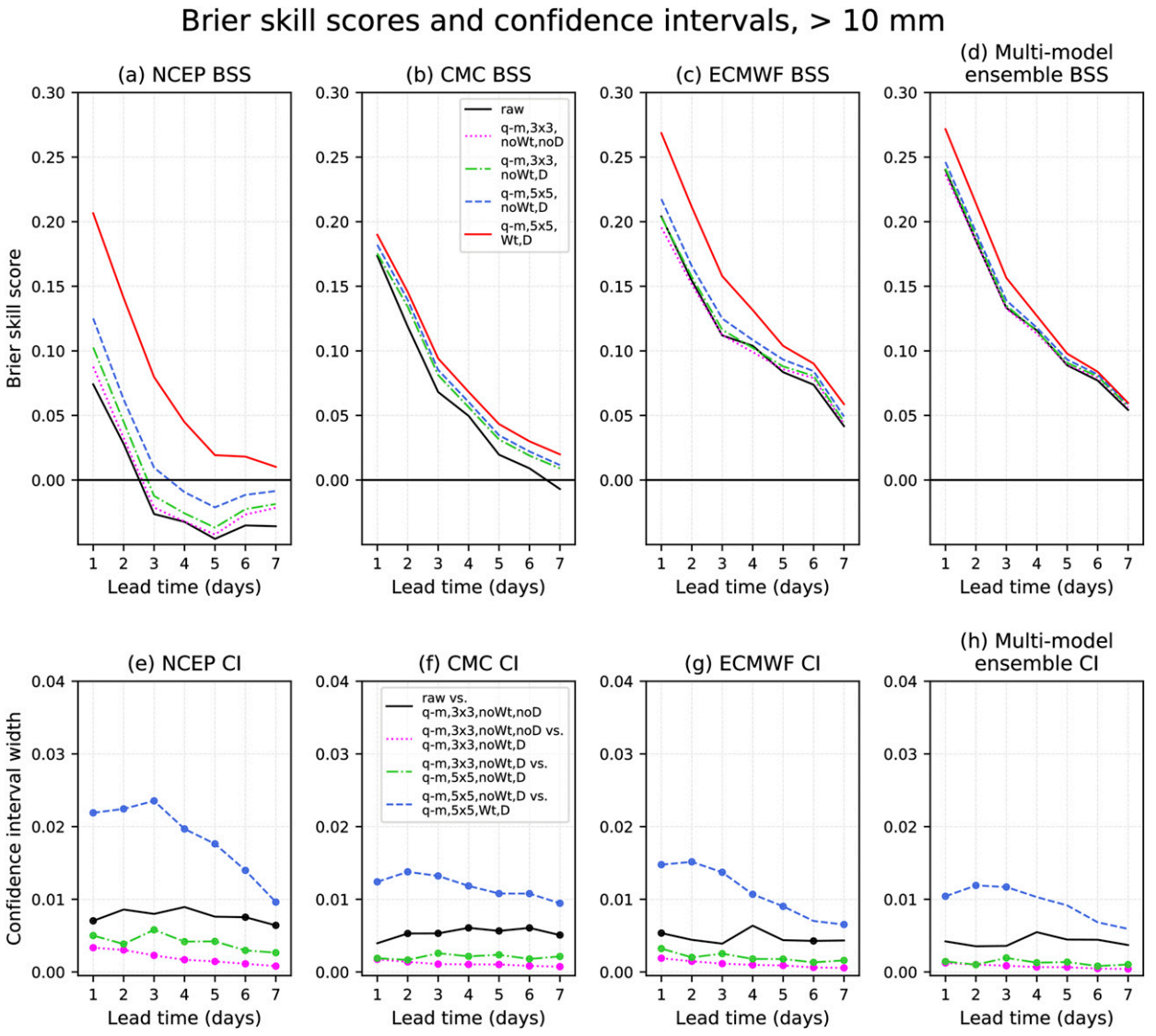

FIG. 3. As in Fig. 2, but for the >10 mm 12 h^{-1} threshold.

described here outperformed the CSGD methodology. One notable exception was that the performance of NCEP's ≥10 mm 12 h^{-1} forecasts at longer leads was improved through the use of the CSGD algorithm. We conjecture that the overall mediocre performance of the CSGD method in the present setup was due to the simplifications that were necessary to address the limited training data. By including local climatological and skill information in the CSGD regression equations, we tried to account for local characteristics and make the assumption of spatially constant regression parameters more justifiable. Still, for a domain like the CONUS that

TABLE 1. Experiment names for various permutations of quantile mapping and dressing algorithm.

Experiment name	Quantile mapping?	3 × 3 stencil?	5 × 5 stencil?	Closest-histogram weighting?	Dressing?
raw	No	No	No	No	No
q-m,3 × 3,noWt,noD	Yes	Yes	No	No	No
q-m,3 × 3,noWt,D	Yes	Yes	No	No	Yes
q-m,5 × 5,noWt,D	Yes	No	Yes	No	Yes
q-m,5 × 5,Wt,D	Yes	No	Yes	Yes	No

FIG. 4. Weighted BSSs for postprocessed forecast skill for the q-m, 5 × 5,Wt,D experiment (see Table 1) but excluding one of the ensemble prediction centers. (a) POP (>0.254 mm 12 h^{-1}) threshold and (b) >10 mm 12 h^{-1} threshold. Percentage weights for the multimodel combination are indicated in the legend, where N = NCEP, C = CMC, and E = ECMWF.

contains diverse climatologies and different predictability, this simplification appears to have significant adverse effects on the performance of the resulting forecasts. We conjecture that the improved performance of the CSGD approach, when applied to NCEP's higher precipitation amount forecasts and longer lead time, is explained by the desirable convergence of CSGD forecasts to the climatological distributions in situations with little predictive skill, as explained by Scheuerer and Hamill (2015).

We turn now to an examination of the reliability diagrams for the forecasts, with POP data shown in Fig. 7 and ≥10 mm 12 h^{-1} data in Fig. 8. Figures 7a and 8a show the reliability and frequency of usage for each individual system's raw ensemble and for the raw multimodel ensemble. All forecast systems were quite unreliable; for POP, the CMC system had the greatest reliability, but its forecasts were not as sharp at high probabilities, compared to the ECMWF system, so ECMWF skills were higher.

Raw ECMWF usage frequencies exhibit a sawtooth pattern that the other systems did not. Why is this? The reliability diagram assigns a range of probabilities to a discrete bin number. For example, the forecast percent probabilities [0, 2.5), [2.5, 7.5), and [7.5–12.5) are assigned to bins 1, 2, and 3. Here the "[" indicates that the lower bound is included, and ")" indicates that the upper bound is excluded. With its 50 members, ECMWF probabilities are 0/50, 1/50, 2/50, and so forth. By inspection, one can see that the 2/50 and 3/50 (two possibilities) are assigned to the second bin, but 4/50, 5/50, and 6/50 (three possibilities) are assigned to the third bin. This oscillation of the number of possible outcomes

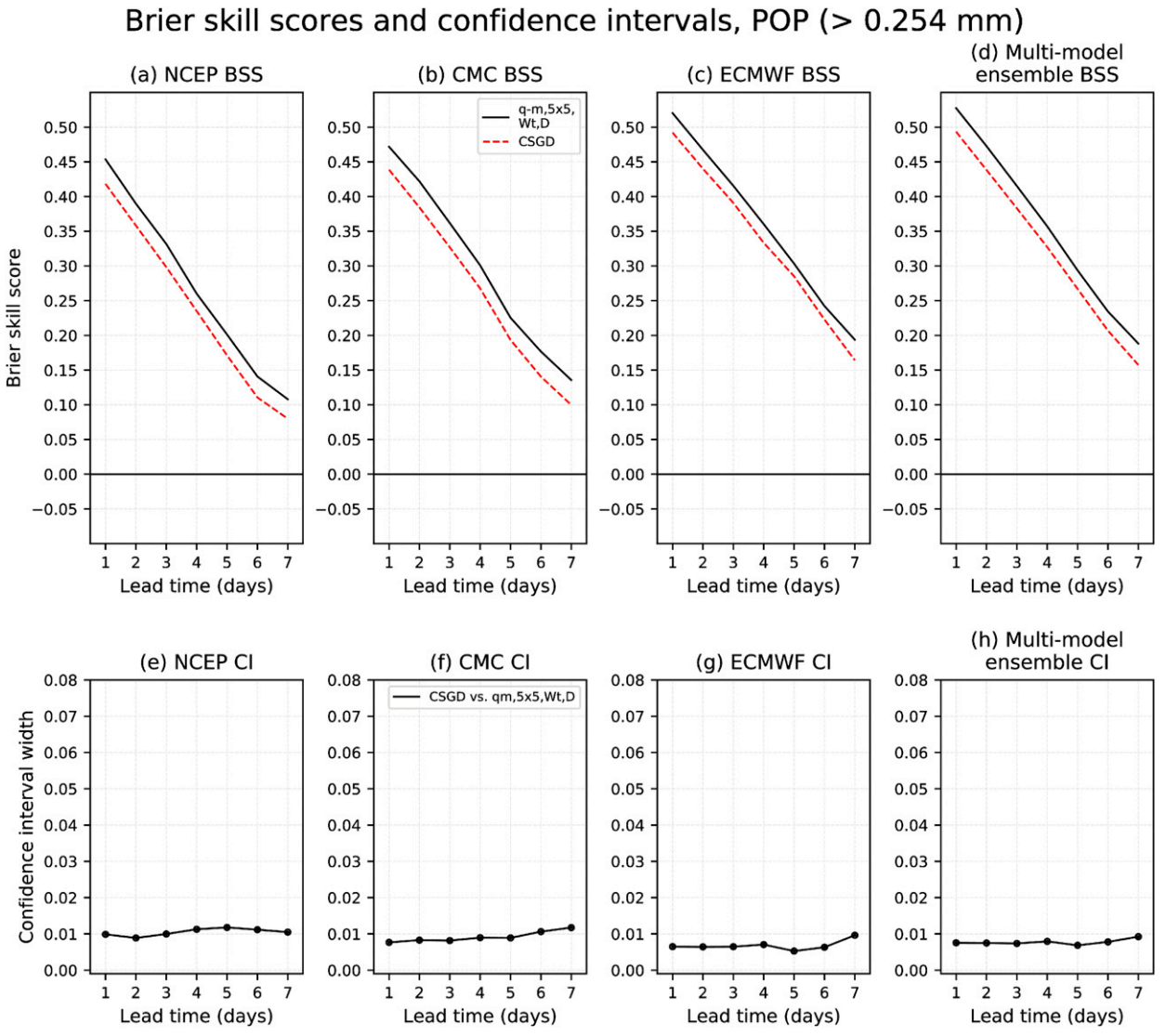

FIG. 5. Comparison of POP postprocessed forecast skill for the q-m, 5 × 5,Wt,D experiment vs the CSGD methodology. (a) NCEP, (b) CMC, (c) ECMWF, and (d) multimodel ensemble. (e)–(h) Confidence intervals and significance following the description in the Fig. 2 caption.

assigned to a particular bin explains ECMWF's sawtooth frequency-of-usage pattern.

Figures 7b and 8b show the effects of quantile mapping using the older 3 × 3 stencil of points and no dressing. Forecasts were made much more reliable for POP through the quantile mapping and only slightly more reliable at ≥10 mm 12 h^{-1}, which was consistent with the greater skill improvement for POP than for ≥10 mm 12 h^{-1} previously shown in Figs. 2 and 3. There was still some remaining unreliability of the forecasts after quantile mapping, especially at ≥10 mm. Reliability and skill were only slightly improved through the use of dressing (Figs. 7c, 8c) and the use of the 5 × 5 stencil (Figs. 7d, 8d). Only after the application of the closest-histogram weighting (Figs. 7e, 8e) were reliabilities significantly improved further. There were still some issues with unreliability at high probabilities, but it is apparent from the frequency of usage histograms that these high probabilities were issued quite infrequently; at lower probabilities that were much more common, the forecasts were quite reliable.

Figure 7e shows an odd characteristic of the postprocessed NCEP guidance of POP after the closest-histogram weighting. Before (Fig. 7d), forecast probabilities in the range of 2.5%–7.5% were issued slightly less than 1% of the time. After the closest-histogram weighting, forecasts in this range were issued much less frequently, roughly two times in 1000. Why did this happen? After the 5 × 5 stencil quantile mapping, the NCEP ensemble is expanded in size to 20 × 25 = 500

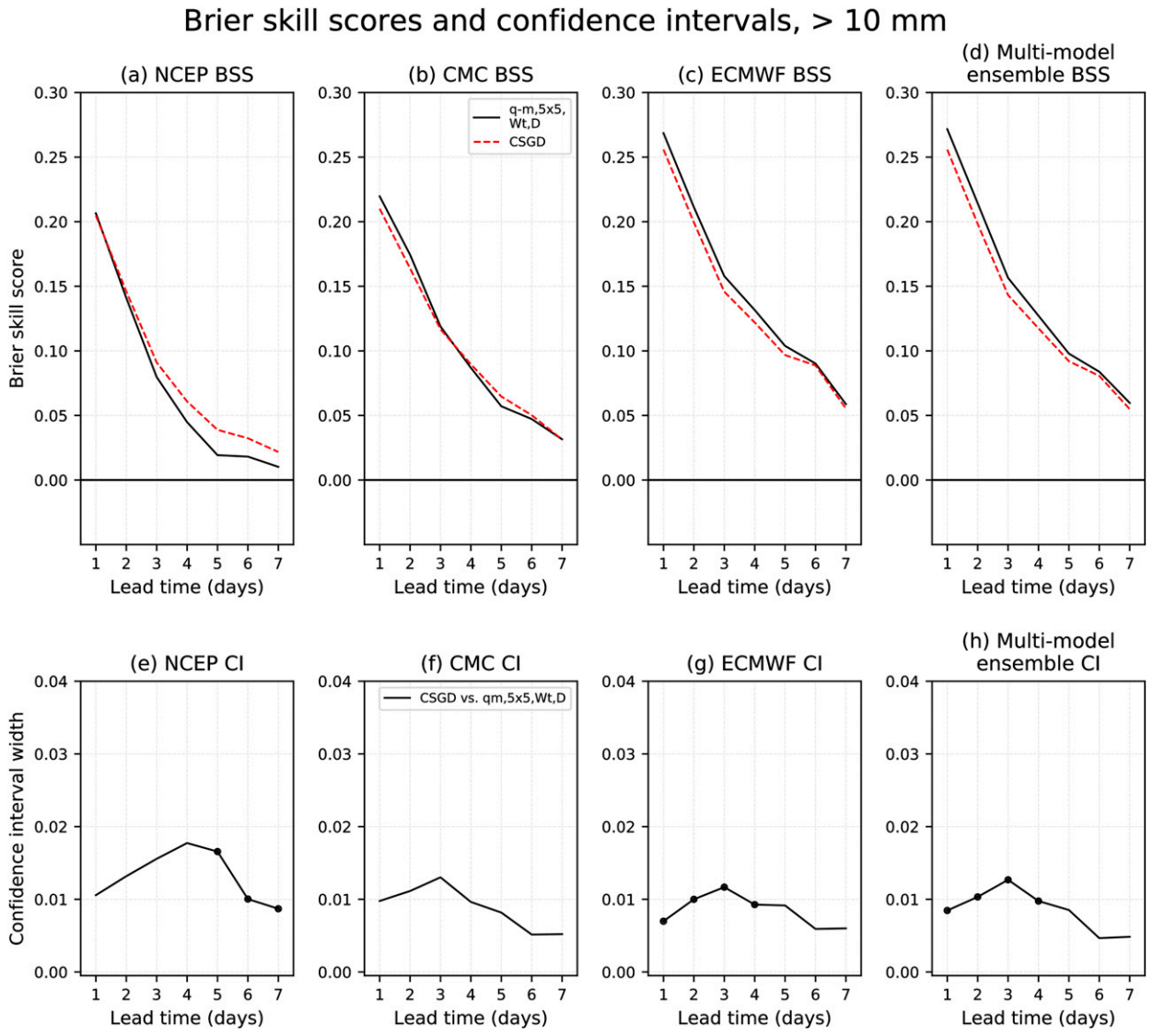

FIG. 6. As in Fig. 5, but for the >10 mm 12 h^{-1} threshold.

members. Neglecting the effects of dressing, forecasts between 2.5% and 7.5% probability would occur with 13–37 members of 500 exceeding the POP threshold. Let us assume that these low probabilities are associated with a relatively low ensemble-mean precipitation amount, between 0.01 and 2 mm, indicating that the closest-histogram weighting would be associated with the data presented in Fig. 1a. From inspection, we see there that the highest-ranking forecast member will have its probability changed from 1/500 to 0.047. That is, when the quantile-mapped forecasts are in the range of 2.5%–7.5%, typically another 5% probability is added to these forecasts through the closest-histogram weighting, dramatically reducing the fraction of situations when forecasts of this probability range are issued. CMC and ECMWF do not exhibit this problem as much because their highest-sorted member has a much lower probability of being the closest member, again shown in Fig. 1a.

The reliability diagrams in Fig. 7f reinforce the previous discussion about the limitations of the modified CSGD approach. The dramatic reduction in the overall degrees of freedom entailed by the assumption of spatially constant regression coefficients made it difficult to obtain high-quality regression equations valid across all grid points and all thresholds. The general reliability reported in previous applications of the CSGD method, where the regression parameters were specific to each analysis grid point, is regrettably no longer valid under the assumption of spatially constant regression coefficients.

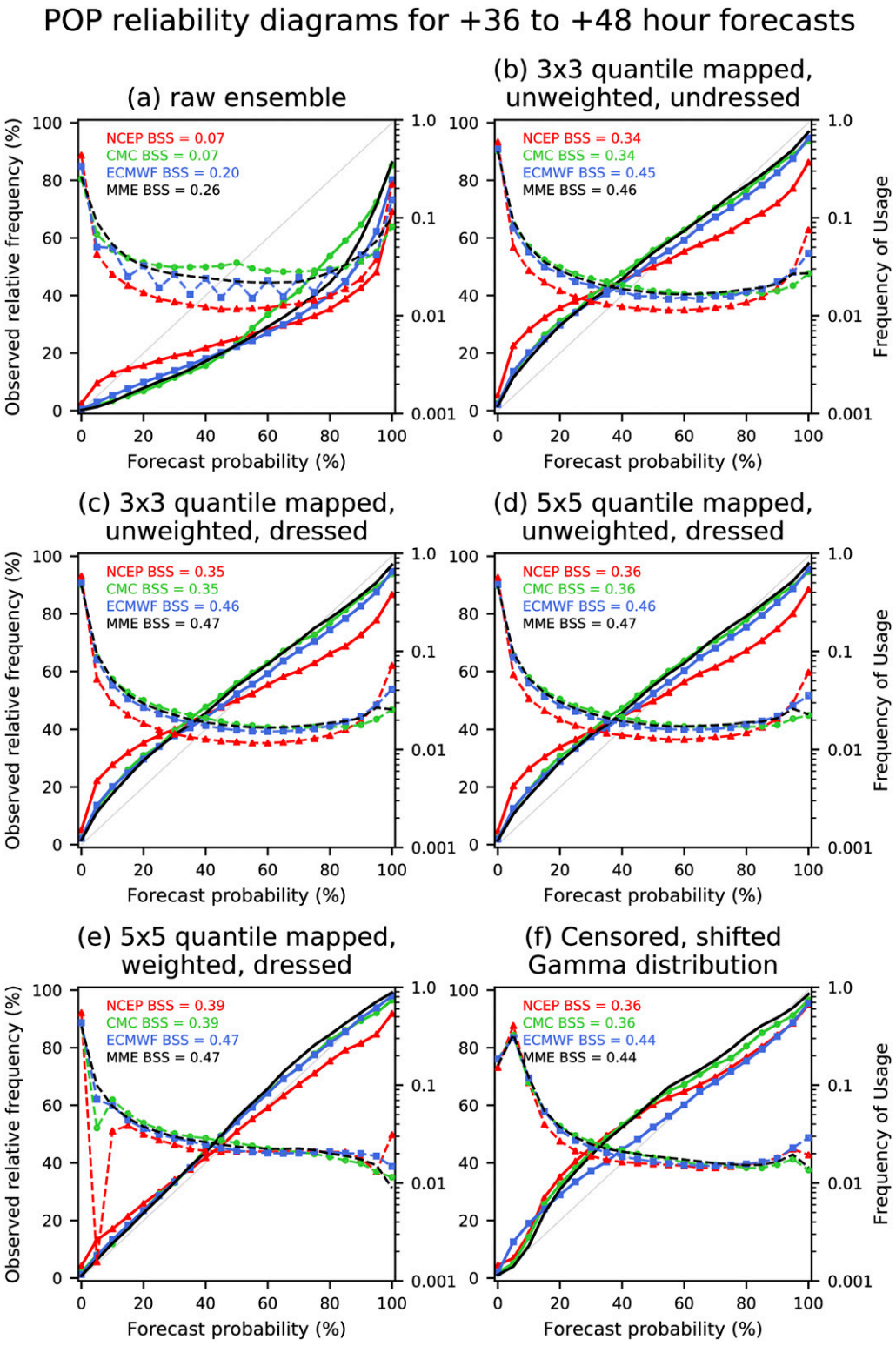

FIG. 7. POP reliability diagrams (left axis label) and logarithmic frequency (right axis label) of forecast usage for CMC, NCEP, ECMWF, and multimodel ensembles (MME, with 20% NCEP, 30% CMC, and 50% ECMWF weighting). Data in various panels described in titles.

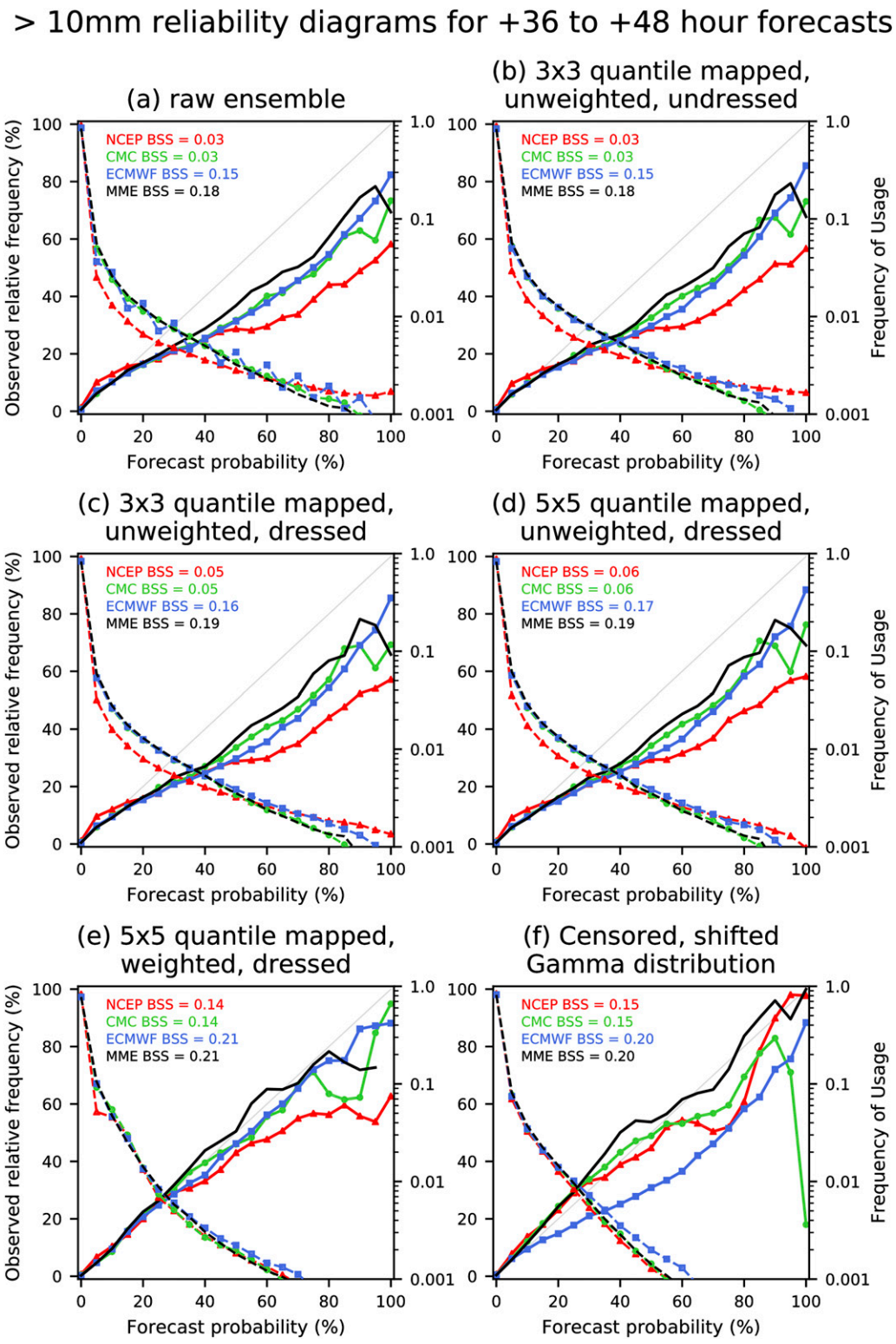

FIG. 8. As in Fig. 7, but for the >10 mm 12 h^{-1} threshold.

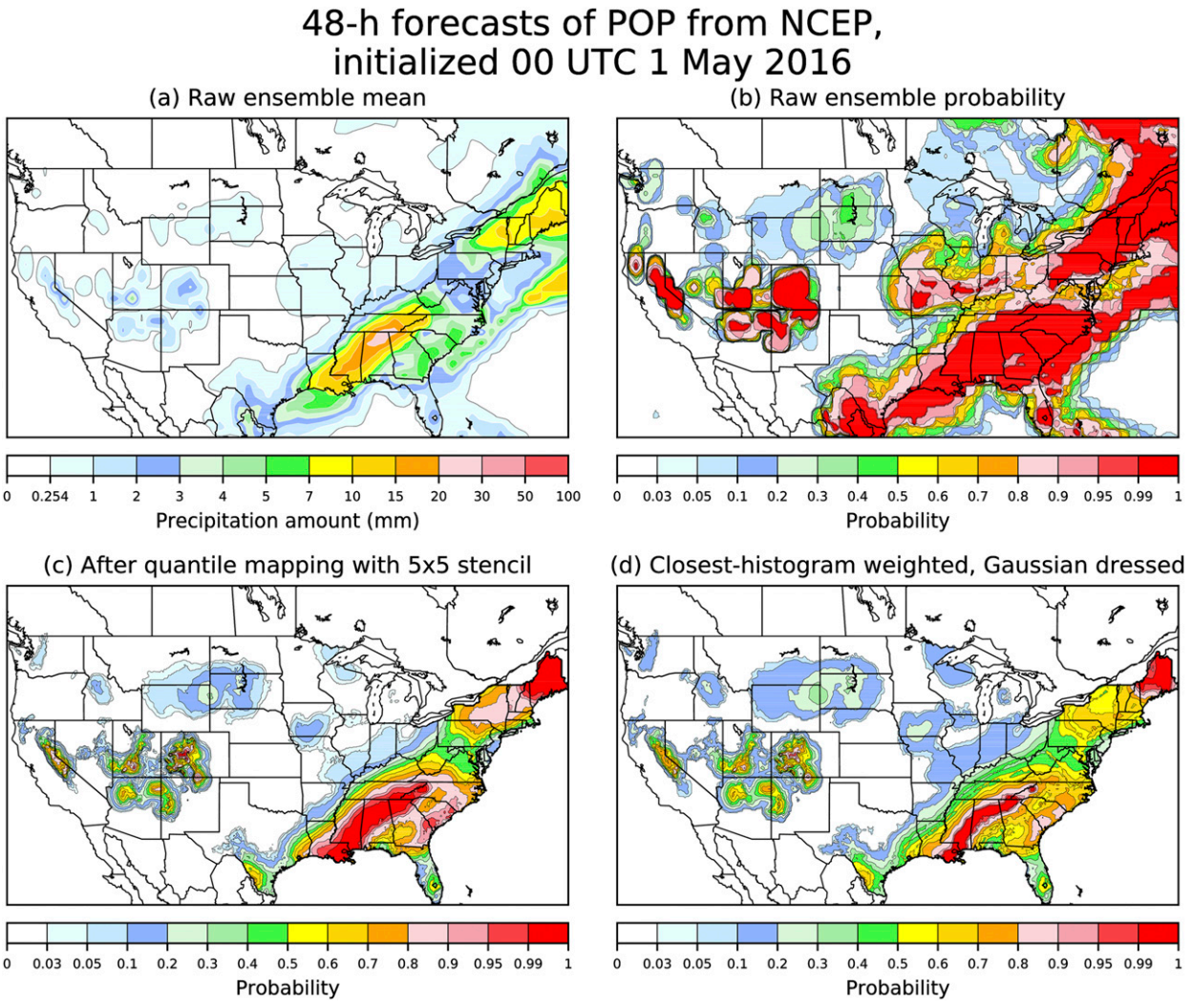

FIG. 9. NCEP POP forecast guidance for +36–48-h forecasts initialized at 0000 UTC 1 May 2016. (a) Raw ensemble-mean precipitation amounts; (b) raw ensemble POP forecast; (c) POP after the quantile mapping; and (d) POP after quantile mapping and weighted dressing.

5. A representative case study

We briefly show a typical case that illustrates the changes that occur in the major steps from raw guidance through to quantile-mapped and dressed postprocessed guidance and finally multimodel combination. For brevity, we do not show the CSGD forecasts.

The 36–48-h predictions from the NCEP, CMC, and ECMWF systems are shown in Figs. 9–11, respectively. In each system, there was a predicted maximum raw ensemble-mean amount, extending northeast from Louisiana to roughly western North Carolina, and a second maximum in the northeast United States. Smaller forecast mean precipitation amounts occurred across the mountains of the western United States. Smaller-scale details differed between the prediction systems. Considering the NCEP raw POP forecasts in Fig. 9b, we see a large area of red indicating probabilities near 1.0 and a general blocky pattern due in part to the limited resolution of the forecast model and storage of data at reduced resolution in the TIGGE archive. After the 5 × 5 stencil quantile mapping shown in Fig. 9c, there was additional terrain-related enhancement of probabilities in the western United States and a decrease in the area with high POP in the eastern United States. The closest-histogram weighting and dressing further depressed high probabilities in the eastern United States and turned many regions in the upper Great Plains and the Ohio River valley with forecasts between 3% and 5% to between 10% and 20%. The enhancement of terrain-related detail and the desharpening of forecasts can also be seen in the CMC forecasts (Fig. 10) and ECMWF forecasts (Fig. 11). After the postprocessing, the three systems'

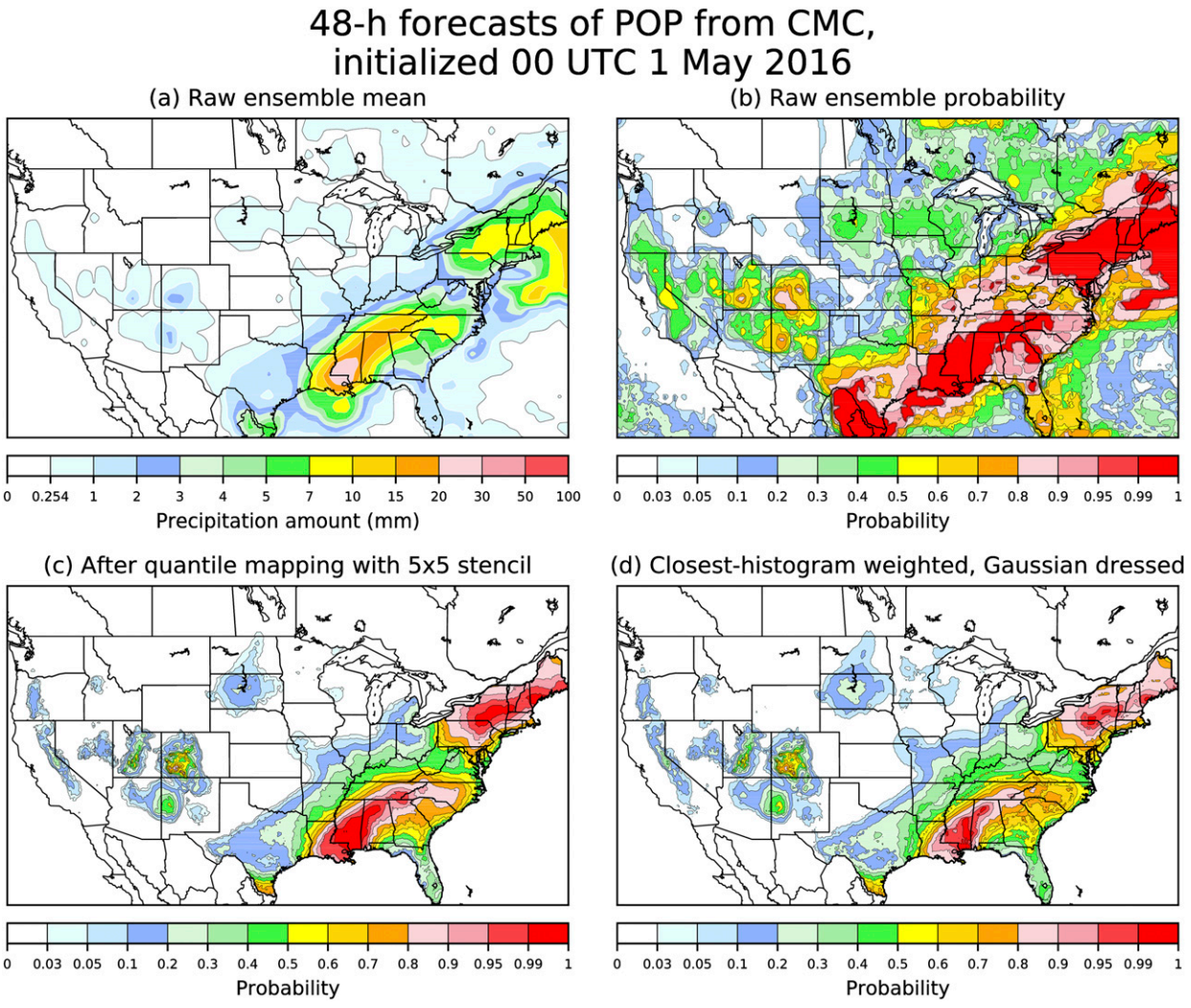

FIG. 10. As in Fig. 9, but for the CMC ensemble.

POP guidances resemble each other much more than the original raw POP guidance did.

Figure 12a provides the final, weighted multimodel POP forecast synthesized from 20% NCEP, 30% CMC, and 50% ECMWF forecast data. As there was some difference in the positions of probability maxima and minima among the three systems, there was an additional slight loss of sharpness in the forecasts. Considering the verifying analysis in Fig. 12b, nearly every area with precipitation greater than the POP threshold was covered by nonzero probabilities, with higher probabilities generally associated with the locations that had higher verifying precipitation amounts.

To provide a quick glimpse of probabilities at a higher threshold, Fig. 13 also provides the final, synthesized weighted multimodel guidance of probabilities of exceeding the 10-mm threshold. Higher probabilities were confined to Louisiana, Mississippi, and northern Alabama. There was somewhat less correspondence between the areas of higher probability and the locations exceeding 10 mm, though again, in almost all circumstances, the locations with greater than 10 mm were covered by nonzero probabilities.

6. Conclusions

This article describes proposed changes to the probabilistic precipitation forecast algorithm that NOAA's research arm proposes to transfer to operational use in the National Weather Service under its National Blend of Models program. The major algorithmic changes from those described by H17 are as follows:

1) The separate postprocessing of each prediction system's guidance, followed by the weighted combination

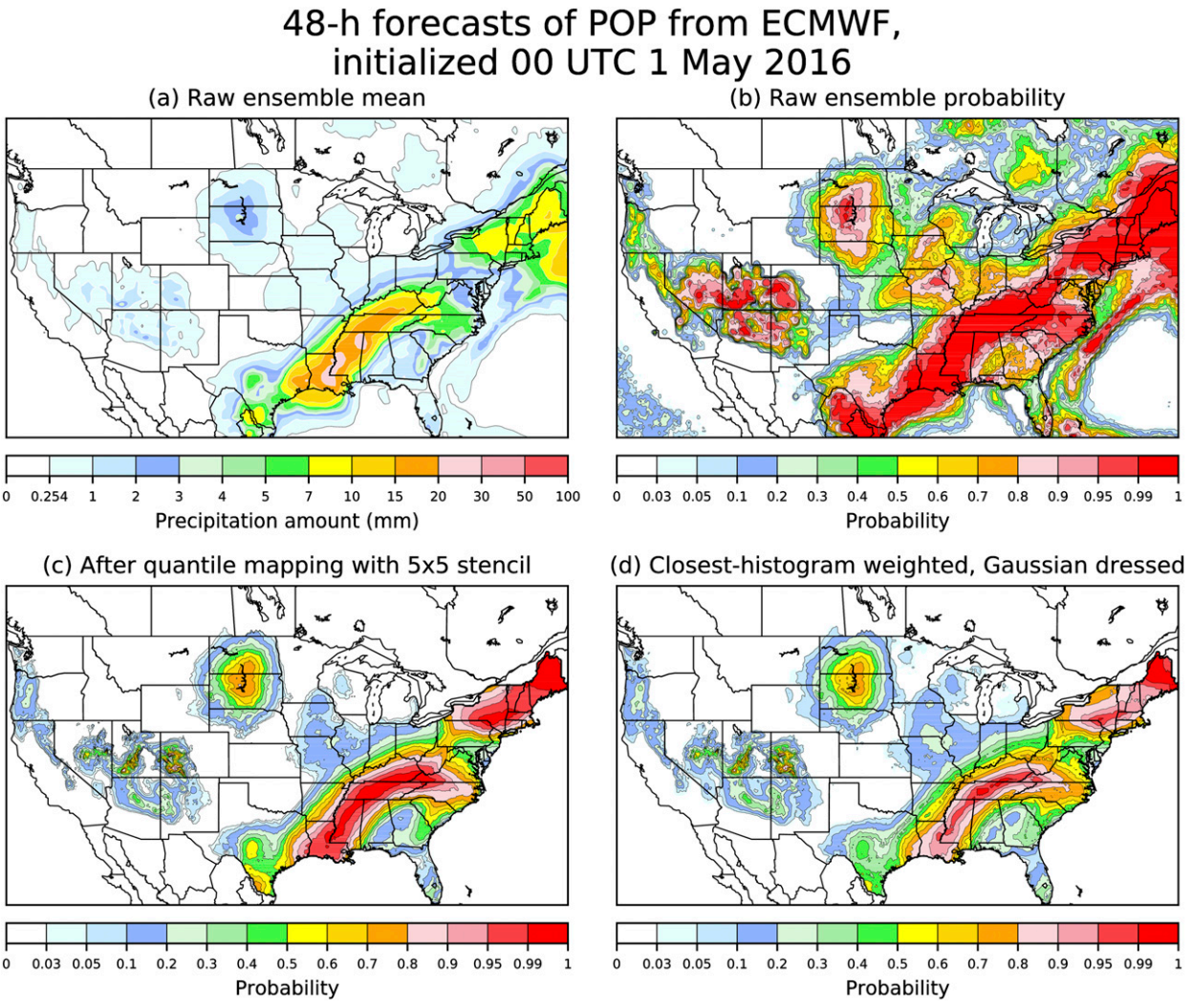

FIG. 11. As in Fig. 9, but for the ECMWF ensemble.

of guidance from all available systems. This facilitates dealing with data delays or outages of the individual prediction systems used in the National Blend.

2) Changes to the quantile-mapping procedure that is used to ameliorate biases in the mean forecast state. In particular, the revised procedure now estimates the forecast and analyzed CDFs used in quantile mapping with a fraction zero and a Gamma distribution for positive amounts. The advantage of this approach is that much less information needs be stored relative to the previous procedure where empirical CDFs were used. The procedure also runs faster.
3) A revised procedure for the quantile mapping at the highest precipitation amounts relative to that grid point's climatology. This procedure addresses the limitations of training sample size in populating the CDFs through the use of a regression approach when today's forecast is between the 90th and 99th percentiles of the forecast CDF. If greater than the 99th percentile, an additional correction is added, the difference of today's forecast from the 99th percentile.
4) The variable weighting of sorted ensemble members according to closest-member histogram statistics, defined through training data of forecasts across the domain during the previous 60 days.

The results presented also include ECMWF ensemble forecast data in this comparison, following a negotiated agreement between NOAA and ECMWF to permit the ECMWF data to be used in the National Blend. The ECMWF forecast system, consistent with prior results, is much more skillful than the other systems in most

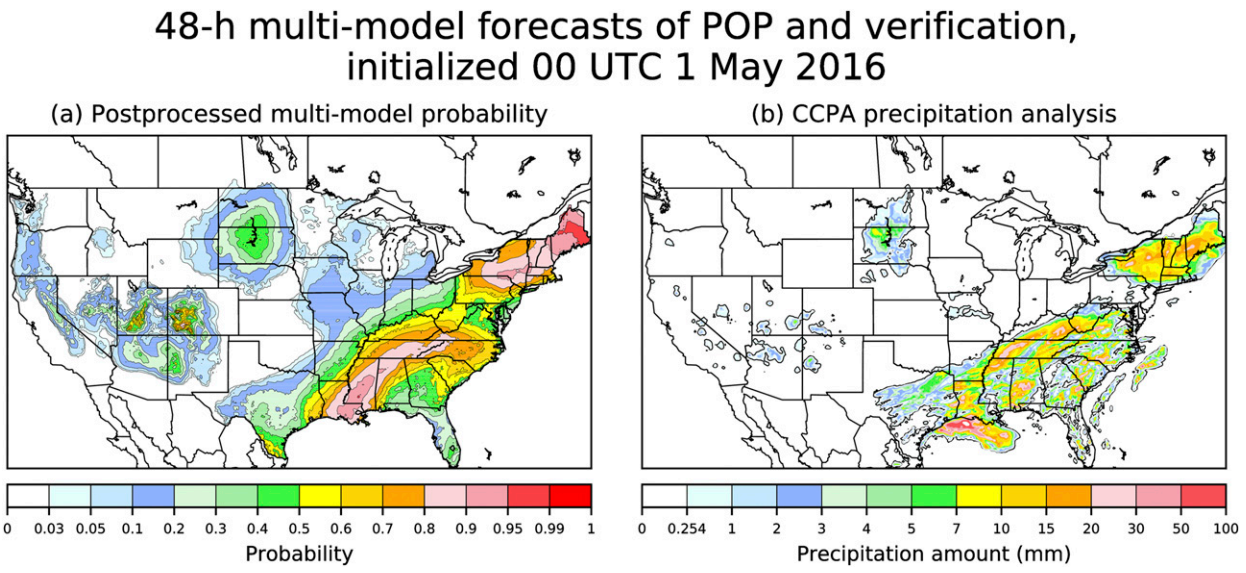

FIG. 12. (a) Multimodel ensemble POP for +36–48-h forecast initialized at 0000 UTC 12 May 2016. (b) Corresponding verifying precipitation analysis for accumulated precipitation for the 12 h ending 0000 UTC 14 May 2016. All areas inside the black contour in (b) verify the event as having occurred.

circumstances and adds substantial skill to the postprocessed guidance.

In general, with the proposed new precipitation postprocessing algorithm, it was demonstrated that additional forecast skill and improved reliability can be added beyond that demonstrated by H17, particularly for heavier precipitation events (the >10 mm 12 h^{-1} results were shown here). These system improvements suggest that with the use of this algorithm, the National Blend probabilistic precipitation forecasts will be of sufficient skill and reliability that they can and should be disseminated more widely. Currently, National Blend guidance does not include fully probabilistic quantitative precipitation forecast guidance, only the probability of nonzero precipitation (POP).

The authors of this article intend to work with National Weather Service colleagues to implement these algorithms in future versions of the National Blend. Still, there are many avenues for continued improvement of the system. One area of current research is the development of methodologies for creating synthetic, high-resolution ensemble forecasts that are consistent with the high-resolution

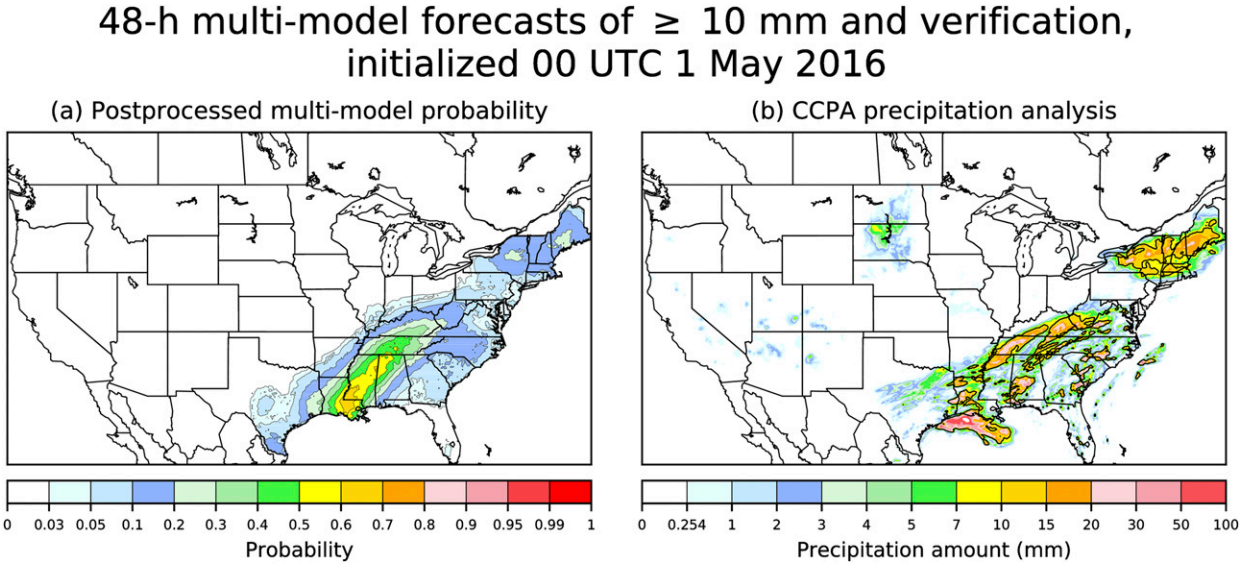

FIG. 13. As in Fig. 12, but for the >10 mm 12 h^{-1} event. All areas inside the black contour in (b) verify the event as having occurred.

postprocessed precipitation forecast guidance created here. Such ensembles with realistic space and time variability are commonly necessary as forcings to ensemble hydrologic prediction systems. We have explored (Scheuerer et al. 2017; Scheuerer and Hamill 2018) approaches suitable for small basins, and we intend to explore methodologies suitable for large basins in the months and years to come.

In the future, some prediction systems, in particular ECMWF and the NCEP global ensemble, will be accompanied by large training datasets. A future National Blend precipitation postprocessing algorithm should be designed to leverage these larger training datasets to improve product quality, and we intend to work with National Weather Service partners on such algorithms in the coming years.

Acknowledgments. Funding for this research was provided by the National Weather Service's Office of Science and Technology Integration via MDL and account S8MWFES-P4Y, intended to support the technology transition of precipitation forecast products to the National Blend.

REFERENCES

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

Buizza, R., 2014: The TIGGE global, medium-range ensembles. ECMWF Tech. Memo. 739, 53 pp., https://www.ecmwf.int/en/elibrary/7529-tigge-global-medium-range-ensembles.

Fortin, V., A.-C. Favre, and M. Saïd, 2006: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quart. J. Roy. Meteor. Soc.*, **132**, 1349–1369, https://doi.org/10.1256/qj.05.167.

Glahn, H. R., and D. P. Ruth, 2003: The new Digital Forecast Database of the National Weather Service. *Bull. Amer. Meteor. Soc.*, **84**, 195–202, https://doi.org/10.1175/BAMS-84-2-195.

Gneiting, T., and R. Ranjan, 2013: Combining predictive distributions. *Electron. J. Stat.*, **7**, 1747–1782, https://doi.org/10.1214/13-EJS823.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, https://doi.org/10.1175/2007MWR2410.1.

——, R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1814–1827, https://doi.org/10.1002/qj.1895.

Haiden, T., M. Janousek, J. Bidlot, L. Ferranti, F. Prates, F. Vitart, P. Bauer, and D. S. Richardson, 2016: Evaluation of ECMWF forecasts, including the 2016 resolution upgrade. ECMWF Tech. Memo. 792, 55 pp., https://www.ecmwf.int/en/elibrary/16924-evaluation-ecmwf-forecasts-including-2016-resolution-upgrade.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

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

——, 2012: Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Mon. Wea. Rev.*, **140**, 2232–2252, https://doi.org/10.1175/MWR-D-11-00220.1.

——, and S. J. Colucci, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724, https://doi.org/10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2.

——, 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, https://doi.org/10.1175/2007MWR2411.1.

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

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, https://doi.org/10.1175/JHM-D-11-0140.1.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran*. 2nd ed. Cambridge University Press, 963 pp.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, https://doi.org/10.1034/j.1600-0870.2003.201378.x.

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

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

——, ——, B. Whitin, M. He, and A. Henkel, 2017: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.*, **53**, 3029–3046, https://doi.org/10.1002/2016WR020133.

Siegert, S., J. Bröcker, and H. Kantz, 2012: Rank histograms of stratified Monte Carlo ensembles. *Mon. Wea. Rev.*, **140**, 1558–1571, https://doi.org/10.1175/MWR-D-11-00302.1.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Thom, H. C. S., 1958: A note on the Gamma distribution. *Mon. Wea. Rev.*, **86**, 117–122, https://doi.org/10.1175/1520-0493(1958)086<0117:ANOTGD>2.0.CO;2.

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

Zhang, Y., L. Wu, M. Scheuerer, J. Schaake, and C. Kongoli, 2017: Comparison of probabilistic quantitative precipitation forecasts from two postprocessing mechanisms. *J. Hydrometeor.*, **18**, 2873–2891, https://doi.org/10.1175/JHM-D-16-0293.1.