The U.S. National Blend of Models for Statistical Postprocessing of Probability of Precipitation and Deterministic Precipitation Amount

THOMAS M. HAMILL

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

ERIC ENGLE, DAVID MYRICK, AND MATTHEW PEROUTKA

NOAA/NWS/Meteorological Development Laboratory, Silver Spring, Maryland

CHRISTINA FINAN

NCEP/Climate Prediction Center, College Park, and Innovim LLC, Greenbelt, Maryland

MICHAEL SCHEUERER

Cooperative Institute for Research in the Environmental Sciences and University of Colorado Boulder, Boulder, Colorado

(Manuscript received 26 August 2016, in final form 27 April 2017)

ABSTRACT

The U.S. National Blend of Models provides statistically postprocessed, high-resolution multimodel ensemble guidance, providing National Weather Service forecasters with a calibrated, downscaled starting point for producing digital forecasts.

Forecasts of 12-hourly probability of precipitation (POP12) over the contiguous United States are produced as follows: 1) Populate the forecast and analyze cumulative distribution functions (CDFs) to be used later in quantile mapping. Were every grid point processed without benefit of data from other points, 60 days of training data would likely be insufficient for estimating CDFs and adjusting the errors in the forecast. Accordingly, ‘‘supplemental’’ locations were identified for each grid point, and data from the supplemental locations were used to populate the forecast and analyzed CDFs used in the quantile mapping. 2) Load the real-time U.S. and Environment Canada (now known as Environment and Climate Change Canada) global deterministic and ensemble forecasts, interpolated to ⅛°. 3) Using CDFs from the past 60 days of data, apply a deterministic quantile mapping to the ensemble forecasts. 4) Dress the resulting ensemble with random noise. 5) Generate probabilities from the ensemble relative frequency. 6) Spatially smooth the forecast using a Savitzky–Golay smoother, applying more smoothing in flatter areas.

Forecasts of 6-hourly quantitative precipitation (QPF06) are more simply produced as follows: 1) Form a grand ensemble mean, again interpolated to ⅛°. 2) Quantile map the mean forecast using CDFs of the ensemble mean and analyzed distributions. 3) Spatially smooth the field, similar to POP12.

Results for spring 2016 are provided, demonstrating that the postprocessing improves POP12 reliability and skill, as well as the deterministic forecast bias, while maintaining sharpness and spatial detail.

1. Introduction

The forecast problem to be discussed in this article is the production of skillful, reliable, and geographically detailed precipitation guidance, leveraging the numerical guidance from one or more numerical weather prediction (NWP) systems and statistically postprocessed

Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/MWR-D-16-0331.s1.

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

DOI: 10.1175/MWR-D-16-0331.1

with short training datasets. Raw NWP precipitation forecasts, deterministic and probabilistic, are often less useful than they could be as a result of imperfections in the underlying prediction system. Forecast guidance may exhibit location-dependent and location-independent biases. Biases may also differ between light and heavy precipitation events, perhaps overforecasting light precipitation and underforecasting the heavier precipitation. Ensembles of precipitation predictions may be underspread and may not offer as much geographic detail as desired by users. For these reasons, statistical postprocessing is often relied on to adjust the current forecast using the discrepancies between past forecasts and observations and to downscale the guidance.

Several articles in the recent past have demonstrated the improvement of probabilistic forecast skill and reliability that results from multimodel ensemble combinations of precipitation forecasts (e.g., Hamill 2012, hereafter H12; Liu and Xie 2014). Presumably different centers, with different prediction systems, will produce guidance with somewhat different and compensating biases and an overall spread larger than that achieved from any one ensemble prediction system. H12 showed that 24-h accumulated probability of precipitation (POP24) from a multimodel ensemble (MME) formed from European Centre for Medium-Range Weather Forecasts (ECMWF), Met Office (UKMO), Canadian Meteorological Centre (CMC), and National Centers for Environmental Prediction (NCEP) global ensembles provided generally reliable and skillful forecasts relative to 1° precipitation analyses over the contiguous United States. Similar results using MMEs were found by Candille (2009) and Swinbank et al. (2016, and references therein).

Consensus forecasts have long been known to produce a more accurate forecast than any single individual model or forecaster when verified over an extended period of time (e.g., Vislocky and Fritsch 1997). Recent studies performed by National Weather Service (NWS) Central Region forecasters also showed that a consensus model blending approach could provide a skillful starting point for NWS digital forecasts (Craven et al. 2013). Accordingly, the National Weather Service instituted the National Blend of Models project, called simply the National Blend hereafter. Under the National Blend, the NWS desires to generate calibrated, high-resolution forecast guidance from statistically postprocessed multimodel ensembles for use in digital forecasting at weather forecast offices and national centers (Glahn and Ruth 2003).

While a straightforward estimation of probabilities from the four-center MME relative frequencies was shown in H12 to provide useful probabilistic precipitation guidance at 1° grid spacing, in the National Blend there are fewer available data (NCEP and CMC global forecast data only with this study), and the final desired output grid spacing over the contiguous United States (CONUS) is much finer, ~2.5 km. The hope and expectation is that statistical postprocessing can correct systematic errors in the mean and spread and can apply subgrid-scale spatial detail where it is realistic, such as statistically downscaling to introduce terrain-related precipitation variability in the western United States.

Can statistical postprocessing realistically improve upon multimodel ensemble precipitation guidance from coarser-resolution forecasts? Previously, several authors have demonstrated that with an extended time series of reforecast and high quality, high-resolution analyzed training data, it is possible to statistically postprocess a single model's output and thus to generate reliable, skillful, and downscaled probabilistic precipitation guidance (Hamill and Whitaker 2006; Hamill et al. 2008, 2013, 2015; Roulin and Vannitsem 2012; Verkade et al. 2013; Scheuerer and Hamill 2015; Baran and Nemoda 2016).

Regrettably, a method that has been demonstrated to produce high quality postprocessed guidance with a lengthy training dataset and a single-model ensemble will not necessarily perform optimally with multimodel ensembles and short training datasets. Hence, it is worth considering alternative methodologies. Bayesian model averaging was previously demonstrated using gamma-distribution kernels (Sloughter et al. 2007; Schmeits and Kok 2010), as well as gamma and censored gamma-distribution-fitting methods (Bentzien and Friederichs 2012; Scheuerer 2014). Logistic regression has been tried in many applications and with variations in formulation (Hamill et al. 2004; Wilks 2009; Schmeits and Kok 2010; Messner et al. 2014). However, some of these methodologies were developed with single models in mind; others were developed for higher-resolution systems where downscaling was less of a concern. Yet others may have been tested only with very limited data, not over a region such as the CONUS spanning many climatological regimes.

For our application, we seek a postprocessing methodology that is capable of providing reliable, highly skillful, highly detailed, probabilistic postprocessed guidance even when trained with limited, coarse-resolution training data. It must also produce fields that are visually acceptable to human forecasters; for example, the POP12 forecasts should not exhibit small-scale variations in probability in regions with relatively flat terrain. To achieve these criteria, we have built a procedure that combines several established methodologies together with some novel elements. At the heart of the procedure is an established technique known as ''quantile mapping'' (Hopson and Webster 2010; Voisin et al. 2010; Maraun 2013). To apply quantile mapping, one generates forecast and analyzed cumulative distribution functions (CDFs) from available forecast and analyzed data for the grid point of interest. Given today's precipitation forecast value at that grid point, one can determine the associated quantile from the forecast CDF and then replace the forecast with the analyzed value associated with that same quantile. Quantile mapping adjusts for bias conditioned on the forecast precipitation amount, and it does so in a way that avoids the collapse of spread common with regression approaches when there is little relationship between the forecast and observed (Wilks 2006, section 2.1.3). Quantile mapping, when leveraging higher-resolution precipitation analyses or observations, also implicitly produces a statistical downscaling. To deal with the small training sample size (the previous 60 days were used here) and other issues, the quantile-mapping procedure applied here will include the population of CDFs using ''supplemental locations.'' Variants of this approach have previously been described, such as in Hamill et al. (2008), Mass et al. (2008), Daly et al. (2008), Hamill et al. (2015), and Lerch and Baran (2017). Another feature to be applied here is quantile mapping of today's forecast values using surrounding grid points following Scheuerer and Hamill (2015), a procedure that enlarges the effective ensemble size and accounts in part for the overconfidence of ensemble prediction systems in the location of precipitation features. For POP12, there is an additional step, the addition of stochastic noise to the quantile-mapped ensemble values, inspired by dressing and kernel-density methods seen in articles such as Roulston and Smith (2003), Fortin et al. (2006), and Sloughter et al. (2007). A final smoothing is also applied to reduce the small-scale variations of probability in geographically flat regions.

In this article, we will also describe a methodology for QPF06 that renders multimodel ensemble mean deterministic precipitation forecasts less biased with respect to the observations, ameliorating the tendency of ensemble-mean forecasts to overforecast light precipitation and underforecast heavy amounts. It is based on a quantile mapping of the multimodel ensemble-mean forecast to analyzed data.

Below, section 2 will describe the forecast and analyzed precipitation datasets used in version 3.0 of the National Blend, operationally implemented 27 July 2017. Section 3 reviews the verification methodologies that are used in this article. Section 4 describes the methodology for increasing the sample size used to populate CDFs through the use of supplemental locations. Section 5 describes the POP12 algorithm, and section 6 describes the QPF06 algorithm. Section 7 provides objective verification statistics of the forecasts before and after postprocessing. Section 8 provides a discussion and conclusions.

2. Forecast and analyzed precipitation datasets

The National Blend project has been active for several years, though earlier versions mostly produced experimental postprocessed guidance, and they have not been described thoroughly in the peer-reviewed literature. The POP12 and QPF06 techniques described here will be incorporated into version 3.0 of in the National Blend, and they represent a more mature product. Because of the large volumes of data involved, we will examine the skill of raw and postprocessed guidance for the 0000 UTC cycle forecasts produced during the period of 1 April–6 July 2016, though data from 21 to 28 June were missing. Forecasts are also produced operationally from 1200 UTC guidance; data from this model cycle were not examined in this study, but we have no a priori reason to expect the performance would be different.

For each forecast lead time, the previous 60 days of coincident forecasts and analyses were used for training. The use of 60 days represented a compromise based on judgment. Larger sample sizes are generally desirable with statistical postprocessing, but since biases may be seasonally varying, use of, say, 120 days may actually provide a worse result and double the data storage. Ideally, were the model unchanged over a multiyear period, the training data would include the previous year's data during the same season. Writing code that could permit different training samples for different constituent models depending on the number of days since the last implementation was judged too difficult for the time being. NOAA also has limited storage on its supercomputers, and storage of more than 60 days' worth of training data was also judged to be impractical for the time being.

Precipitation forecast data was ''budget'' interpolated (Accadia et al. 2003) onto the ⅛° grid spacing of the analyzed data. In the current National Blend product, there is a final step of interpolating the postprocessed POP12 and QPF06 results onto a 2.5-km grid. In the future, postprocessing may occur directly on the 2.5-km grid, but for present purposes, results are presented only for the ⅛° output.

a. Forecast data

The primary data sources in this study were global ensemble forecasts from the NCEP Global Ensemble

Forecast System (GEFS) and the Canadian Meteorological Center (CMC) Ensemble Global Environmental Prediction System (GEPS). Deterministic forecasts from the NCEP Global Forecast System (GFS) and the CMC Global Deterministic Prediction System (GDPS) were also used. In the future, the NWS hopes to be able to leverage a greater range of data, including shorter-range forecasts from NWS models and possibly data from other prediction centers.

The version of the NCEP GEFS used in this experiment is described by Zhou et al. (2016, manuscript submitted to *Wea. Forecasting*) and this version went into operation on 2 December 2015. The GEFS used the NCEP Global Spectral Model (GSM) version 12.0.0 (described online at http://www.emc.ncep.noaa.gov/GFS/impl.php), and in turn the GSM used the model settings implemented on 14 January 2015. The GEFS horizontal resolution was T574 (spectral with triangular truncation at wavenumber 574) to day +8 and T382 thereafter to day +16, corresponding to respective grid spacings of approximately 27 and 40 km at 40°N. The GEFS system used 64 vertical levels. Twenty ensemble members were generated for each cycle. Model uncertainty was generated through the stochastic total tendency perturbations of Hou et al. (2008). Initial conditions were generated with ensemble Kalman filter (EnKF) perturbations centered around a control analysis generated with a hybrid EnKF–4D-variational analysis procedure (Kleist and Ide 2015a,b). The hybrid analysis used 75% weighting of EnKF covariance estimates and 25% weighting of static covariances.

The deterministic forecasts from the NCEP GFS used the GSM version 12.0.0, which provided forecast data at T1534 resolution on a reduced Gaussian grid with a grid spacing of approximately 10 km at 40°N for forecasts to +240-h lead. Semi-Lagrangian time stepping was used. For forecasts from +240 to +384 h, the horizontal resolution was T574. Sixty-four vertical layers are used, with a model top at 0.3 hPa. Assimilation was the same as described for the GEFS above. Other model changes are described online (http://www.emc.ncep.noaa.gov/GFS/impl.php).

The CMC GEPS dataset used in this study was produced using version 4.1.1 of their software, described in Gagnon et al. (2014, 2015). The GEPS used in turn the Canadian Global Environmental Multiscale Model (GEM) version 4.6.3, with the basic computational dynamics described in Cote et al. (1998a,b). Data assimilation for defining the GEPS initial conditions used a 256-member EnKF described in Houtekamer et al. (2014). The horizontal grid for assimilation was 800 × 400, providing a grid spacing of ~38 km at 40°N. The ensemble forecast system was computed on the same grid with 40 vertical levels. A model change log for the Canadian prediction systems is available online (http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/changes_e.html). One notable feature of the GEPS system was the use of multiple parameterizations that differed from one member to another (Gagnon et al. 2014, 2015), instituted to provide a larger amount of spread. A consequence of this was that CMC GEPS ensemble members were not exchangeable (i.e., different members may have different precipitation forecast biases). This required the archival of separate forecast CDFs for each GEPS member and member-by-member quantile mapping, discussed in section 5c.

b. *Analyzed precipitation data*

Precipitation analyses for training and verification were obtained from the Climatology-Calibrated Precipitation Analysis (CCPA) dataset described in Hou et al. (2014). As described in the article, the CCPA makes statistical adjustments to NWS Stage IV precipitation products so that they are more consistent with NCEP Climate Prediction Center gauge-based analyses. The CCPA was judged to be of high quality, though the ⅛° grid spacing was larger than desirable, given the ultimate forecast product was requested on a 2.5-km grid. In the future, it is expected that ⅛° CCPA data will be replaced by other, higher-resolution precipitation analyses where available.

This study produced 12-hourly probability of precipitation forecasts and 6-hourly accumulated deterministic precipitation forecasts. Accordingly, we extracted both 6- and 12-hourly accumulated precipitation from the CCPA dataset for similar time periods to match the QPF06 and POP12 forecasts. For this study, precipitation was extracted, processed, and validated on an ⅛° grid over the CONUS. Only grid points inside the CONUS or over the Columbia River basin in Canada and a few other small basins north of the U.S.–Canadian border were considered here.

In the definition of supplemental locations discussed in the next section, a longer time series of CCPA data was used, spanning 2002–15, as opposed to the prior 60 days used for the quantile mapping conditional bias correction procedure to be described in section 5c. The longer time series of CCPA data permitted a more accurate estimate of the precipitation climatology.

3. Verification methodologies

a. *Probabilistic forecast verification*

A common way of evaluating probabilistic predictions from ensembles is through the use of reliability diagrams

(Wilks 2011, section 8.4.4). The reliability diagrams assess the relationships between the forecast probability (in this case, for the event of 12-h accumulated precipitation ≥ 0.254 mm)[1] and the observed frequency. A 1 to 1 relationship is desirable (i.e., in instances when a 30% probability is forecast, that analyzed event frequency should be 30%). In this study, reliability was evaluated for 21 bins between 0% and 100%. Along with reliability curves, inset histograms are included that provide the frequency with which forecasts of various probabilities were issued. The inset histograms provide information on how sharp the forecasts were (i.e., the extent to which they issued more ones or zeros and fewer intermediate probability forecasts). The ultimate goal is maximal forecast sharpness subject to reliability (Gneiting et al. 2007).

Additionally, Brier skill scores (BSSs) for the POP12 event are provided. The BSS is calculated relative to climatology, where climatology is determined uniquely for each month of the year and each ⅛° grid point inside the CONUS from the 2002–15 CCPA dataset. BSS is calculated here in a rather conventional way (Wilks 2011, section 8.4.2), rather than more involved methods such as Hamill and Juras (2006) that guard against falsely attributing skill due to geographic variations in climatological event frequency. This simpler version of the BSS is acceptable for evaluating relative changes of skill with changes in the system, as we are interested in here. The more involved calculations are preferred if quantifications of absolute skill are desired.

The conventional method for calculating BSS is as follows. For a particular forecast lead time, let $P(l, c)$ be the forecast of POP12 at location l (a tuple of the i and j indices) and the case day c, $0 \leq P(l, c) \leq 1$. There are L overall locations in the CONUS and C case days, and associated with each location l is a latitude $\phi(l)$, which is used to provide a weight to the sample proportional to the grid box area so that grid boxes at northern latitudes do not have undue influence on the results. The analyzed event $O(l, c)$ is set to 1.0 if the analyzed precipitation is $\geq$0.254 mm, and it is set to 0.0 if the analyzed precipitation is <0.254 mm. The Brier score of the forecast BS_f is then calculated as a grid-box-size weighted sum of the average squared error of the probability forecast:

$$\mathrm{BS}_f = \frac{\frac{1}{LC}\sum_{l=1}^{L}\sum_{c=1}^{C}\cos[\phi(l)][P(l,c)-O(l,c)]^2}{\frac{1}{LC}\sum_{l=1}^{L}\sum_{c=1}^{C}\cos[\phi(l)]}. \qquad (1)$$

The Brier score of climatology (BS_c) is calculated similarly. We define $P_c(l, c)$ as the climatological event probability of $\geq$0.254 mm in the 12-h period, as determined from 2002–15 CCPA data. This climatology is determined separately for each grid point l and for each month of the year. Then, $P_c(l, c)$ replaces $P(l, c)$ in Eq. (1) when calculating BS_c. The BSS is then calculated as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_f}{\mathrm{BS}_c}. \qquad (2)$$

A perfect probability forecast has a value of 1.0, and a forecast of 0.0 has the same skill as climatology. Some studies also provide a decomposition of the BS_f into components of reliability, resolution, and uncertainty (Wilks 2011, section 8.4.2). However, since this decomposition is only strictly valid when samples are drawn from a population with the same underlying distribution (Hamill and Juras 2006), plots based on this decomposition are omitted.

b. Deterministic forecast verification

Equitable threat scores (ETSs) and bias (BIA) are determined in standard fashion following Wilks (2011, Eqs. 8.18 and 8.10 respectively), though again they may overestimate the magnitude of the actual skill by neglecting variations in climatological event probabilities (Hamill and Juras 2006). These standard scores are presented here, as their use is common in the NWS. ETS and BIA are generated by populating a contingency table with bins for (forecast, event) pairs. Let $F(l, c)$ denote the ensemble-mean forecast at location l and case day c, and $O(l, c)$ is the analyzed value. The event threshold again is $T = 0.254$ mm. Now, define an indicator variable for the event of the forecast and analyzed being greater than or equal to the threshold amount:

$$\begin{aligned} 1_{F\geq T, O\geq T}(l,c) &= 1 \quad \text{if } F(l,c) \geq T \\ &= 0 \quad \text{if } F(l,c) \leq T. \end{aligned} \qquad (3)$$

We then denote a as the grid-box-size weighted number of samples associated with both the forecast and analyzed exceeding a particular event threshold:

$$a = \sum_{l=1}^{L}\sum_{c=1}^{C} 1_{F\geq T, O\geq T}(l,c)\cos[\phi(l)]. \qquad (4)$$

[1] Eventually the National Blend is expected to provide more full PQPFs (i.e., exceedance probabilities for many other events with different precipitation thresholds). For brevity and because the full PQPF products are still in development, verification of other precipitation thresholds will be omitted in this study.

Similarly, b is defined as the grid-box-size weighted number of samples when the forecast was equal to or exceeded T but the analyzed did not. Here, c is the weighted number of samples when the forecast did not exceed T but the analyzed equaled or exceeded T, and d is the weighted number of samples when both the forecast and analyzed were below T. Then, the ETS is calculated as

$$\mathrm{ETS} = \frac{a - a_{\mathrm{ref}}}{a - a_{\mathrm{ref}} + b + c}, \tag{5}$$

where $a_{\mathrm{ref}} = (a + b)(a + c)/(a + b + c + d)$. If all samples are drawn from a population with the same underlying climatology, then ETS = 1 for a perfect forecast and 0 for a forecast with the skill of climatology. In this situation, we are computing ETS using samples from locations with wide variations in their climatologies, and hence positive ETS may be reported even when skill is zero (Hamill and Juras 2006).

Bias is also calculated using contingency table elements. Bias is defined as follows:

$$\mathrm{BIA} = \frac{a + b}{a + c}. \tag{6}$$

An unbiased forecast has a value of 1.0; bias exceeding 1.0 indicates the events are overforecast on average, and bias below 1.0 indicates the events are underforecast on average.

4. Augmenting training sample size with supplemental locations

Before describing the supplemental location methodology, we propose a rationale for doing so. Suppose we had a practically infinite time series of reforecasts and associated high quality analyses, say four or more decades of daily gridded forecast and analysis data, and with the underlying observing system stable during the period. Were we to examine the systematic errors for a particular member of that prediction system, we would notice biases that were both location dependent and location independent. Location-dependent bias might be related to the local climatology and terrain elevation and orientation (or aspect). For precipitation, it is likely that the bias has some seasonality, varying from winter to summer. Bias may be conditional upon the forecast amount (i.e., different for 1- versus 10-mm forecasts). With additional exploratory data analysis, we might discover further conditional biases—perhaps the model overforecasts precipitation on a westerly wind and underforecasts precipitation on an easterly wind. Perhaps bias differs with the phase of El Niño–La Niña or with other low-frequency modes of oscillation.

Additionally, the forecast guidance may exhibit bias that is largely independent of location. It has previously been noted (e.g., H12) that many global forecast models often systematically overforecast the occurrence of light precipitation amounts and underforecasts heavy amounts (e.g., Moore et al. 2015). These may be due to, say, deficiencies in the model numerics or its parameterization suite.

We are presented with a practical problem in statistical postprocessing—the challenge of providing meaningful statistical adjustment to the raw guidance with limited training data. In this application, a practical constraint was to provide as much improvement as possible using only the last 60 days of forecasts; more would be desirable to increase the sample size, while less might be desirable given seasonal changes in bias characteristics and frequent model changes with associated changes in systematic errors. While the use of longer training datasets such as reforecasts is preferable, in their absence we aim to still be able to make meaningful statistical adjustments.

What from the above list of potential biases is *practical* to estimate with the most recent 60 days' worth of data? What from the above list is *important* to estimate with 60 days of data? Assuming we are limited to training with only the most recent data, estimating bias conditional upon low-frequency phenomena like El Niño and La Niña is not practical, and probably there are not enough data to estimate the possible bias dependency on weather aspects like wind direction. It is clearly important to correct for widespread systematic biases. Arguably, too, it is important to try to correctly estimate gross location-dependent biases if they exist. To illustrate location dependence of systematic error, let us examine CDFs for two nearby locations (Fig. 1). The CDF of the forecast at a particular grid point location (i, j) is defined as

$$F_f(A_f) = P(X_f \le A_f), \tag{7}$$

where A_f is a particular precipitation amount and X_f is a random variable for the forecast and analyzed amounts. A CDF for the analyzed amount $F_a(A_a)$ is defined similarly. The CDFs are estimated from the long-term event-relative frequency using GEFS reforecasts and only the data specifically at that grid point without data from supplemental locations so that we can be confident that systematic errors reflect that particular location. The percentile associated with a particular precipitation amount is commonly known as a "quantile." Figure 1 shows forecast and analyzed CDFs for two nearby points

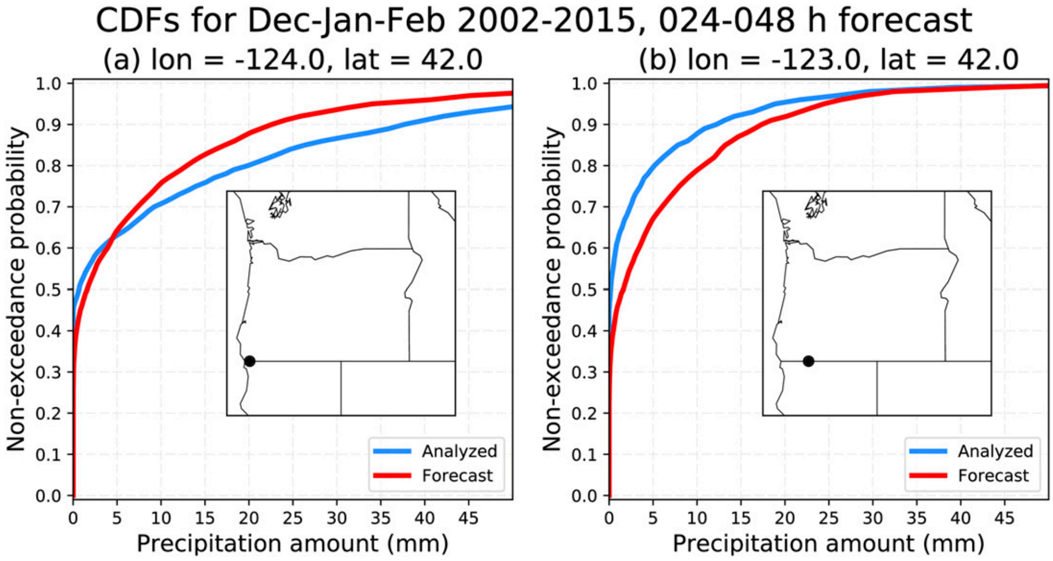

FIG. 1. Illustration of CDFs for December–February 2002–15 using CCPA analysis data (blue) and GEFS reforecast data [red; see Hamill et al. (2013) for more on reforecasts: CDFs (a) for 42°N, 124°W and (b) for 42°N, 123°W. Locations are denoted by the black dots on the inset maps.

in southern Oregon, developed with a relatively large sample: 14 winter seasons of GEFS reforecast data (Hamill et al. 2013) and accumulated precipitation analyses for +24- to +48-h forecasts. A 24-h accumulation period was used here so that biases could not be attributed to the diurnal cycle. As noted by the large differences in analyzed and forecast CDFs at the two locations, these two points have greatly different biases conditional upon forecast amount. For example, for the western point in Fig. 1a, the quantile associated with the 15-mm forecasts is associated with the ~23-mm analysis. At the eastern point in Fig. 1b, the quantile associated with the forecast value of 15 mm is associated with ~10-mm analysis. That is, the western grid point has a dry bias for this forecast amount, and the eastern grid point a wet bias. Perhaps the smoothed terrain representation in the GEFS system was partly responsible. In any case, this is the sort of repeatable first-order location-dependent bias that a human forecaster would prefer to see automatically corrected. Without this correction, the resulting POP12 forecasts would not have appropriate terrain-related variability, and a forecaster would likely feel compelled to reintroduce it through manual modification.

In summary, with a short training dataset we would like at least to be able to correct for spatially invariant systematic bias and gross location-dependent bias. The potential location dependence of bias would suggest performing statistical corrections on a gridpoint-by-gridpoint basis, not using data from other points. Consistency of forecast biases due to endemic model problems argues for the pooling of training data across broader sets of locations to minimize sampling error. Pooling of training data (e.g., Charba and Samplatsky 2011a,b; Hamill et al. 2008, 2015; Mass et al. 2008; Daly et al. 2008; Kleiber et al. 2011; Lerch and Baran 2017) may also be a practical necessity when the training sample size is small, as it is in this application.

We now discuss the details of the supplemental location methodology, first briefly below and then in more detail in the online supplemental material (see file A). The presumption underlying the methodology is that biases can be identified that have commonalities related to terrain elevation, terrain orientation, and to some extent on the climatological distribution of precipitation. If these assumptions are met, then for a given grid point, it will be possible to identify supplemental locations with similar precipitation climatologies and terrain characteristics, and augmenting the training data with information from these locations will result in reduced sampling error for the resultant CDFs with minimal diminution of the capacity to correctly infer the location-dependent bias. Presumably, the population of CDFs using the forecast and analyzed data at additional supplemental locations will help ameliorate sampling error while still preserving the ability to correctly estimate location-dependent biases. It is also hoped that the resulting CDFs will span a larger range of weather conditions despite the use of a short training sample.

Hamill et al. (2015) described an earlier version of an algorithm to determine the supplemental locations

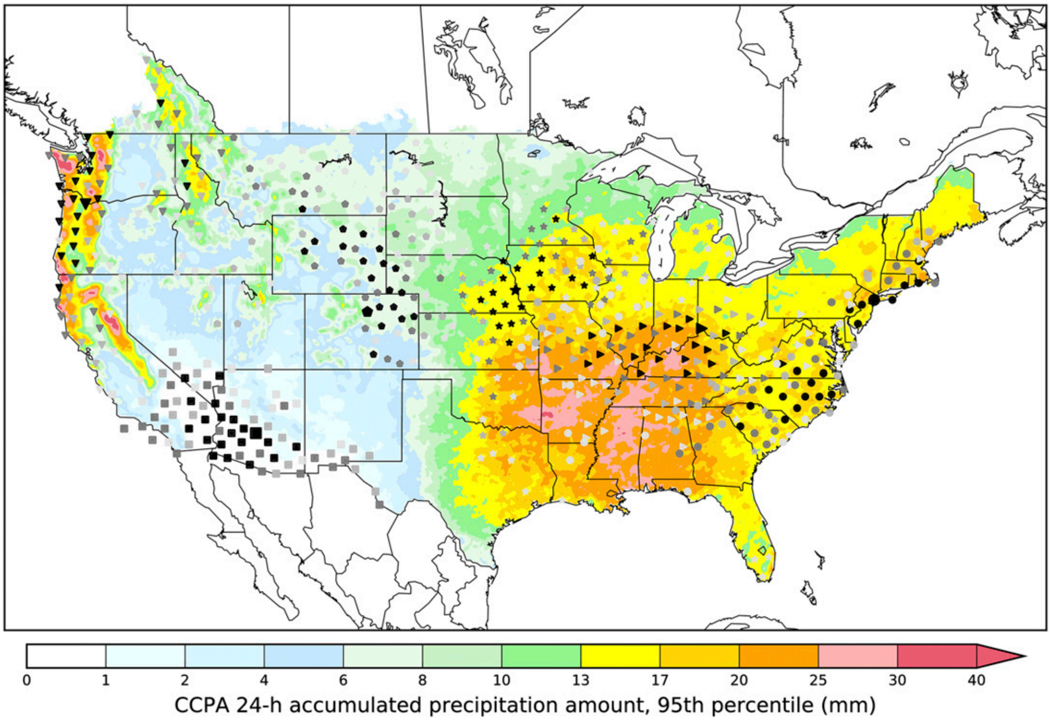
0
1
2
4
6
8
10
13
17
20
25
30
40
CCPA 24-h accumulated precipitation amount, 95th percentile (mm)

FIG. 2. Illustration of supplemental locations for the month of April. Larger symbols denote the locations for which supplemental locations were calculated (roughly Portland, OR; Phoenix, AZ; Boulder, CO; Omaha, Cincinnati, OH; and New York City, NY). Smaller symbols indicate the supplemental locations. Darker symbols indicate a better match, and lighter symbols a poorer match. The colors on the map denote the 95th percentile of the 24-h accumulated precipitation amounts for the month, determined from 2002–15 CCPA data.

tailored toward the larger samples available with reforecasts. In that application, postprocessing at a particular forecast was based on the training data not only from the forecast grid point but also using data from 20 other supplemental locations. Postprocessed skill was increased, especially for higher amounts. Also, Hamill et al. (2008; see Fig. 2) demonstrated how omitting the use of supplemental locations resulted in undesirable small-scale variability in probabilistic quantitative precipitation forecasts (PQPFs) when postprocessed with logistic regression. For the current application, where a much smaller training dataset is available (again, only the last 60 days), a modified algorithm is presented that specifies a greater number of supplemental locations. It is similar in concept to the methodology in Hamill et al. (2015), but different in a few details; the specific algorithmic details of the supplemental locations are described in full in file A in the online supplement.

As was illustrated in Fig. 1, indiscriminate use, say, of other surrounding data points based solely on proximity can provide substandard adjustments for conditional bias. In the current procedure, for each "target" grid point where a postprocessed forecast was produced, a set of supplemental locations was defined based on the similarity of terrain characteristics, analyzed precipitation climatologies, and horizontal distance. Supplemental locations were also required to be spaced some minimum distance from the target point and from each other to provide more independent samples. The selection of supplemental locations was based on the minimization of a penalty function. The first supplemental location was defined as the location with the smallest penalty (i.e., the smallest weighted difference in precipitation climatology, terrain characteristics, and distance), while maintaining a minimum distance from the original grid point. The second supplemental location was similarly defined, but it was also required to be a minimum distance from both the target and the first supplemental location. Definitions of the third, fourth, and subsequent locations proceeded similarly in an iterative manner. Each grid point in the contiguous United States had a minimum of 50 supplemental locations defined, though for regions where a larger number could be found with relatively small penalty functions, up to 100 supplemental locations were defined. This typically increased the number of supplemental locations in flatter, drier areas.

Figure 2 illustrates the supplemental locations that were defined for several preselected grid points for the month of April; separate supplemental locations were calculated for each month. The 95th percentile of the 24-h accumulated precipitation climatology determined from the 2002–15 CCPA dataset is underlain for reference. As shown in Fig. 2, the supplemental locations are prevented from being too close to each other by design so as to potentially provide a more independent set of samples spanning a larger range of weather conditions. The algorithm, as intended, exhibits the tendency to define the supplemental locations based on the similarity of precipitation climatology. Notice, for example, that supplemental locations for Omaha, Nebraska, and Portland, Oregon, tend to occur more preferentially at locations with similar 95th percentiles of climatology. Though not shown, terrain height and orientation (facet) also were factors in the selection of the particular locations. Similar plots for different months are shown in online file A.

Supplemental locations are not a panacea for all of the problems of precipitation bias correction. The current algorithm may still not provide realistic estimates in situations where the recent past has encountered unusually widespread dry or wet weather. For example, should the region surrounding a grid point of interest be experiencing a large-scale extended drought, the CDFs populated with the last 60 days of data and supplemental locations will not resemble the CDFs that would be generated with multiple decades of data. Consequently, the accuracy of the subsequent quantile mapping should be considered suspect if today's weather is unlike anything that has occurred in the area in the past 60 days. Another problem is that the CDFs implicitly reflect the forecast bias of the previous 60 days, which may or may not reflect the current forecast bias; after all, precipitation biases may change with seasons, from winter seasons over the CONUS dominated by nonconvective rainfall to summer seasons more affected by convective rainfall. Such challenges are largely unavoidable when postprocessing with a short training dataset.

5. The 12-hourly probability of precipitation algorithm

The overall process for generating a gridded POP12 forecast is as follows: 1) Populate the CDFs for the forecast and analyzed data using the past 60 days and the supplemental location data. 2) Read in most recent deterministic and ensemble forecasts from the NCEP and CMC systems, and interpolate them onto the ⅛° grid of the precipitation analyses. 3) Perform a deterministic quantile-mapping procedure to correct for the conditional biases of each member. 4) Dress each member the resulting ensemble with random noise to introduce some additional ensemble spread. 5) Form an initial POP12 forecast from the ensemble relative frequency. 6) Perform a Savitzky–Golay smoothing of the POP12 forecast, applying more smoothing in flatter regions. We now describe each of the steps in more detail.

a. *Populate CDFs*

The first step, updating of CDFs, is typically performed prior to the arrival of the most recent forecast data. The updating is dependent on the arrival of the most recent CCPA analysis estimates, which of course arrive after the forecast data for the same time period. A rolling archive is maintained of the last 60 days of paired forecast and precipitation analysis data; when a new day's data are ready, the oldest data are discarded. For the GEFS, forecast CDFs are generated using all members. For the CMC ensemble, separated CDFs are computed for each member given the use of different parameterizations or parameters for individual members, with potentially different forecast biases. Separate archives of forecast–observation pairs are maintained for each forecast lead time in question (e.g., 12, 24, 36 h, and so forth) to permit estimation of lead-time dependent CDFs, since biases may vary with forecast lead or over the diurnal cycle. With the last 60 days of valid data and with the supplemental locations predetermined, CDFs can now be generated. For a particular lead time, the last 60 days of forecast and analyzed data are input, as are the files of supplemental locations. For each ⅛° grid point determined to be inside the CONUS or the Columbia River basin in Canada, a CDF is populated separately for the interpolated forecast and the analyzed data, using data from that grid point and from the predefined supplemental locations. The resulting CDF data for each grid point consists of a pair of vectors: one denoting ordered accumulated precipitation amounts and the other indicating the cumulative nonexceedance probability (i.e., the probability that the analyzed or forecast result is lower than or equal to the amount; see Eq. (7)]. This is based simply on the relative frequency in the training data.

b. *Interpolation of forecast data to ⅛° grid*

The second step, the input of forecast data and interpolation onto the ⅛° grid of the CCPA, is rather straightforward. "Budget" interpolation is used (Accadia et al. 2003).

c. *Quantile mapping of each member*

The third step, a deterministic quantile-mapping procedure, is more involved. First, we review the concept of quantile mapping, as this lies at the heart of the algorithm. Equation (7) provided a definition of the CDF. There is also an inverse distribution function, also known as the quantile function, which maps from a given cumulative probability back to a precipitation amount. For example, for a precipitation forecast, the quantile function is

$$A_f = F_f^{-1}(p), \quad p \in [0,1]. \tag{8}$$

Quantile mapping, which generates a bias-corrected precipitation amount conditional on the input forecast, $A_{f\rightarrow a}$, can thus be expressed as

$$A_{f\rightarrow a} = F_a^{-1}[F_f(A_f)]. \tag{9}$$

The quantile of the CDF associated with today's forecast amount is identified, and then the bias-corrected forecast amount is determined from the analyzed amount associated with the same quantile [see Fig. 2 in Hamill and Whitaker (2006) for an illustration of the concept]. In the application to POP12, the quantile mapping procedure is repeated for every member of the multimodel, multicenter forecast ensemble. After quantile mapping, the resulting ensemble has members that have more exchangeable error characteristics and that are more consistent with draws from the analyzed climatological distribution. Ensemble variability that is due purely to bias differences between members has been reduced (Eckel and Mass 2005).

For POP12, we introduce some modifications to the basic quantile-mapping procedure of Eq. (3). A first modification is introduced to deal with the potential for the large sampling variability of quantiles at the extremes. In situations where the quantile associated with the forecast value is greater than 0.95, quantile mapping is performed with a regression analysis modification following Scheuerer and Hamill (2015); see also file A in the online supplement and Eqs. (1) and (2) above.

Another modification is in the synthetic enlargement of the ensemble size. When processing a given grid point (i, j), quantile mapping uses not only data at (i, j) but forecast data at eight surrounding grid points as well (Figs. 3 and 4). Forecast CDFs and precipitation forecast amounts are used from each of the locations in the 3 × 3 array, but the quantile mapping only uses the analyzed CDF from the center grid point, thereby attempting to render the 3 × 3 array of forecasts to be plausible as draws from the central point's analyzed climatology. A related application of this underlying technology was first described in Scheuerer and Hamill (2015); see also Fig. 2 and the associated text. Some underlying rationales for the use of the 3 × 3 array in this application are that 1) ensemble size is limited, and methods for realistically increasing ensemble size may ameliorate sampling error, and 2) ensembles frequently have small-to-moderate position errors, and quantile mapping using forecasts from other nearby positions provides some robustness against position errors. In the National Blend version 3.0 described here, the 3 × 3 arrays of points were spaced 3/8° apart from each other at +12-h lead, linearly increasing to 1¾° apart at +240-h lead. Ideally, the optimal grid spacing would be identified through extensive testing of multiple separation values. The chosen values were instead informed by trial and error.

Figures 3 and 4 provide a practical application of the methodology for a grid point in Northern California. The surrounding grid points have forecast CDFs that vary from rather dry (Fig. 3c) to nearly as wet as the analyzed state (Fig. 3e). Consequently, there are a variety of mapping functions, shown in Fig. 4, reflecting the spatially dependent adjustments that are made for each of the nine grid points as they are individually adjusted to be consistent with the analyzed sample climatology at the central point.

By quantile mapping using this 3 × 3 stencil of data points, one can imagine very different resulting ensembles and POP12s for differing weather scenarios. Consider first a situation with a widespread area of moderate precipitation. In this case, a high forecast quantile will be identified for each of the nine locations, with presumably a moderate or higher number of quantile-mapped analyzed amounts across the 3 × 3 stencil since Fig. 4 shows an underforecast bias. This will result in a POP12 presumably near 1.0 estimated from the enlarged, quantile-mapped ensemble relative frequency. In a second scenario, forecast precipitation is spatially scattered, with perhaps a high amount at the central point but zero precipitation at several other surrounding points. Quantile mapping using the 3 × 3 stencil of forecast values will thus produce a larger ensemble with many zero quantile-mapped values. A lower POP12 will be generated from the quantile-mapped ensemble relative frequency relative to the previous scenario with its widespread moderate precipitation. In this way, the spatial consistency of the quantiles associated with the forecast precipitation becomes an implicit predictor of POP12.

d. Ensemble dressing

Because of sampling error and possible remaining conditional biases even after a deterministic quantile

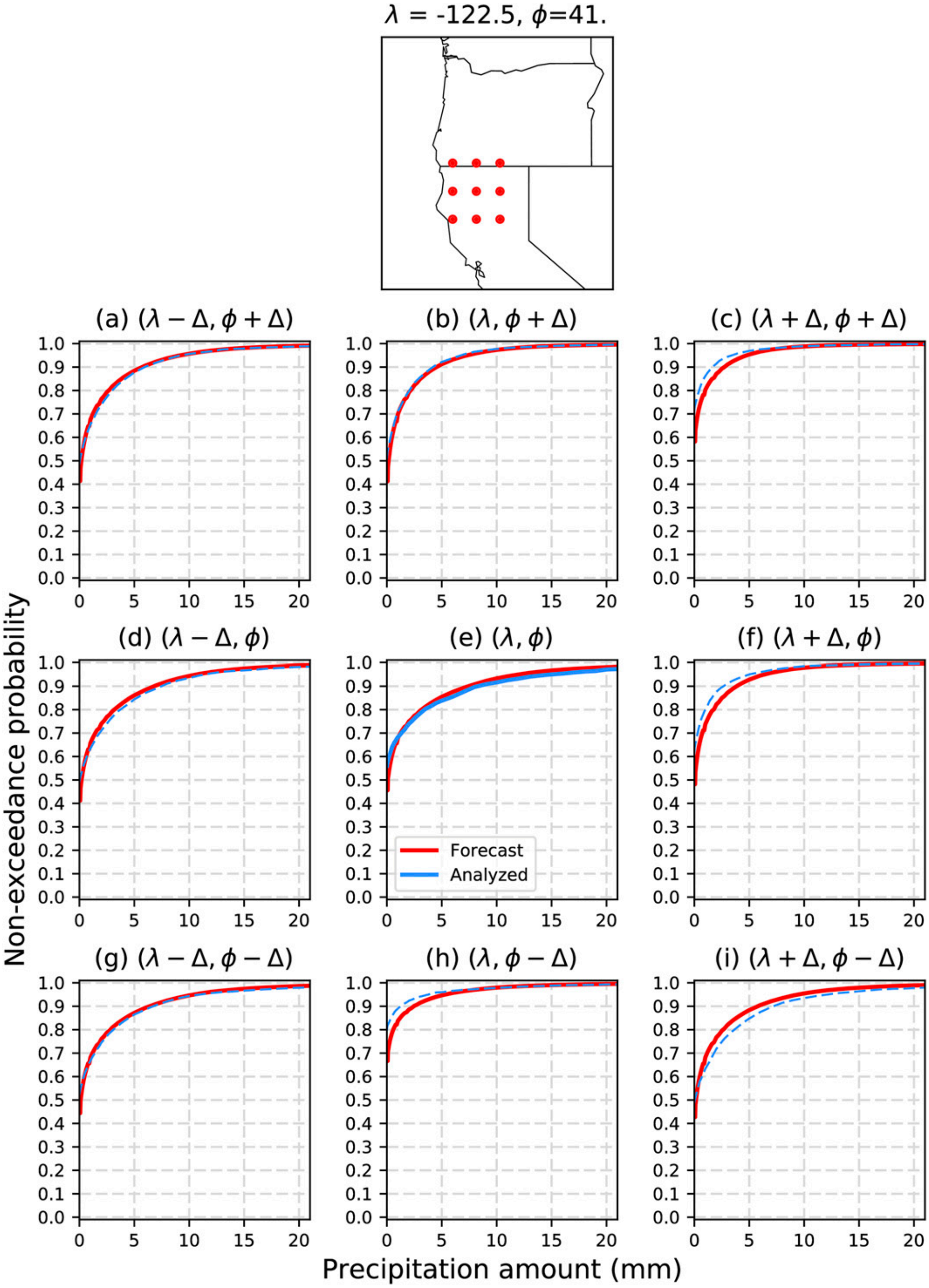

FIG. 3. Illustration of the quantile mapping of data from multiple locations surrounding a grid point of interest, here in Northern California. In this case, we are producing a quantile-mapping adjustment to ensemble members for the center grid point of the nine red dots shown in the inset map. CDF data are for +120- to +132-h member forecasts from the NCEP GEFS system initialized at 0000 UTC 6 Apr 2016 using the past 60 days and supplemental locations. Rather than quantile mapping the 20 GEFS members only at the center point [in (e)], we quantile map the 20 GEFS forecasts at each of the nine locations with red dots. The quantile mapping uses the CDF of the forecast at each of the nine grid points [red curves in (a)–(i)] and the analyzed data at the center grid point in (e). The resulting ensemble at the center point has ninefold more members.

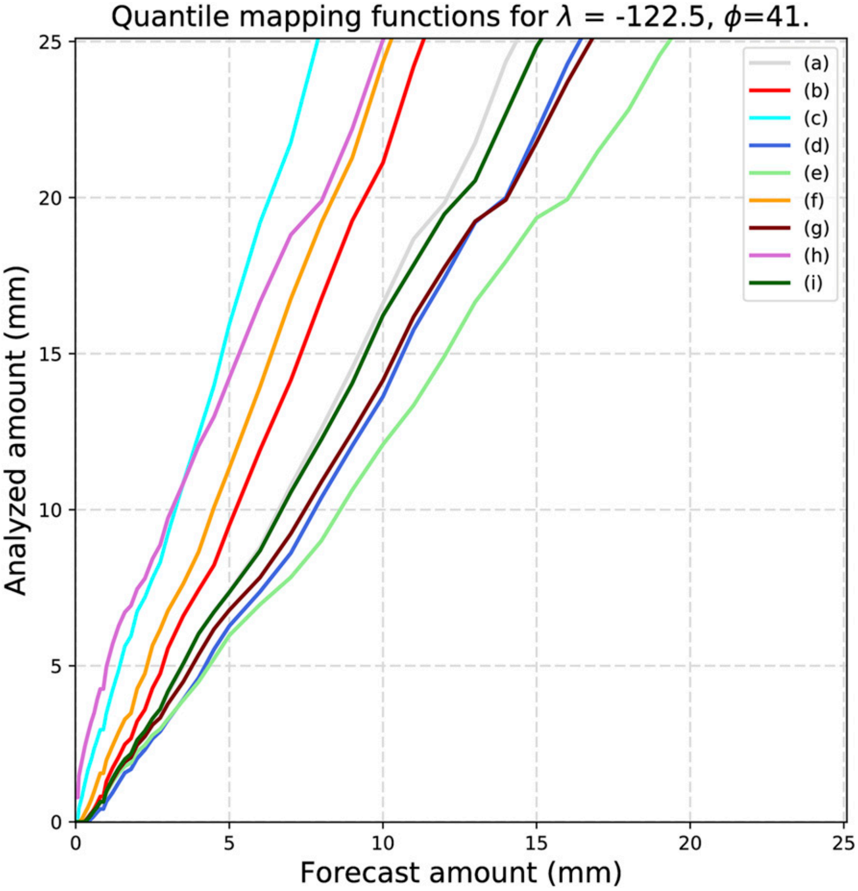

FIG. 4. Deterministic quantile mapping functions for each of the locations in Fig. 3.

mapping, the range of the ensemble can still be too narrow, resulting in suboptimal POP12 reliability. The fourth step in the POP12 algorithm is thus the dressing of the ensemble with random, amount-dependent noise. Adding noise to each forecast member consistent with the statistics of errors of the member that is closest to the verification will improve the POP12 generated from the resulting dressed ensemble. The procedure was inspired by the ''best member'' ensemble dressing concepts of Roulston and Smith (2003) and Fortin et al. (2006), as well as the kernel fitting in the precipitation BMA algorithm of Sloughter et al. (2007).

For this application, random noise is drawn from a normal distribution whose mean is zero and whose standard deviation is 0.25 + 0.5 × the quantile-mapped value, that is, $\sim N[0, (0.25 + 0.5A_{f\rightarrow a})^2]$. If the resulting dressed forecast has a negative value, it is reset to zero. The noise magnitudes were arrived at through both trial and error and through their similarity to objectively determined Gamma-distribution dressing statistics that may be incorporated into a future version of the National Blend (not shown). The dressing procedure is illustrated in Fig. 5. This shows the shifts in mean position due to the deterministic quantile mapping for several forecast values and the implied pdf of the dressing distributions.

e. Estimating POP12 from ensemble relative frequency

The fifth step is simple. Estimate the POP12 from the enlarged ensemble's relative frequency. The NWS threshold for POP12 is 0.01 in., or ~0.254 mm. Hence, the POP12 is determined by counting the number of quantile-mapped, dressed ensemble members equal to or exceeding 0.254 mm and dividing by the total number of members. With the use of eight surrounding data points, the effective ensemble size is now 9 times larger than the size of the original MME.

f. Location-dependent Savitzky–Golay smoothing

The sixth step, a location-dependent Savitzky–Golay smoothing, is conceptually relatively simple but

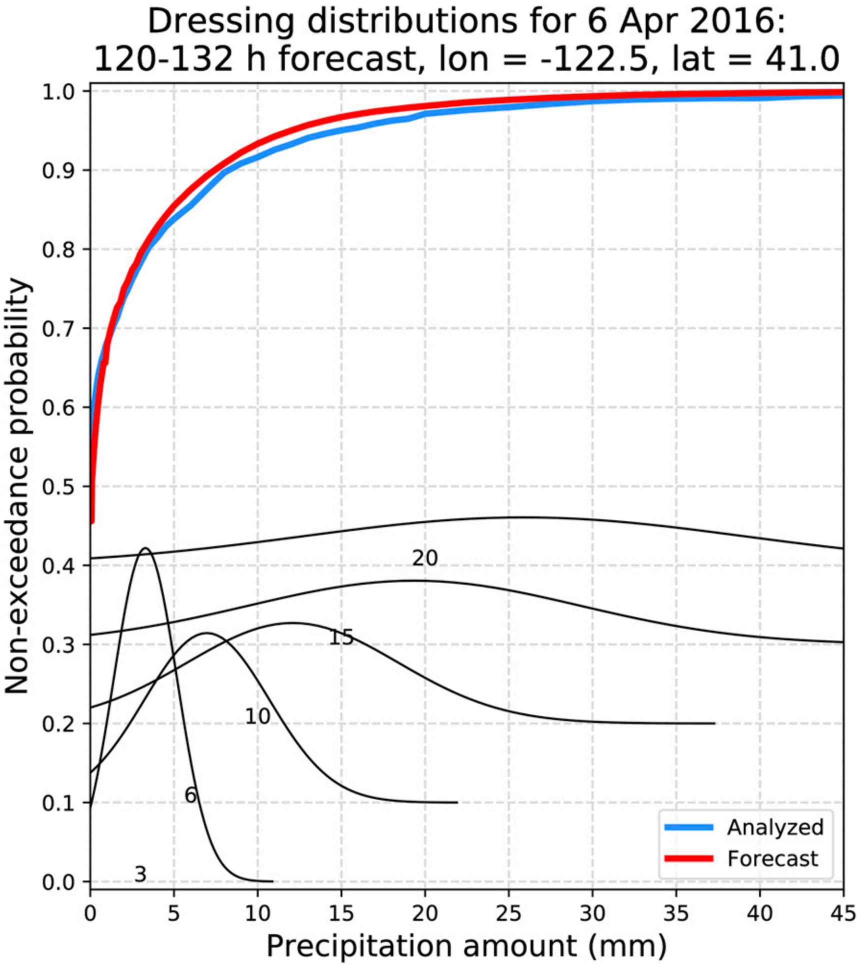

FIG. 5. Illustration of the dressing procedure. CDFs are shown for CCPA analyses (blue) and for NCEP ensemble precipitation forecasts (red). Here, the forecasts were for +120- to +132-h lead for 47.75°N, 123.75°W initialized at 0000 UTC 6 Apr 2016. For initial precipitation amounts of 3, 6, 10, 15, and 20 mm, the underlying pdfs for the dressing are shown in the black curves. The pdfs are normal distributions with means centered on the quantile-mapped values. Per the text, the standard deviations for the stochastic noise are set to 0.25 + 0.5 × quantile-mapped values.

algorithmically more complicated. With a few modifications, the algorithm follows the procedure outlined in Hamill et al. (2015); see also file A in the online supplement. The underlying premise is this: probabilities estimated from the ensemble are subject to sampling error, and this may result in POP12 forecasts that emerge from step 5 above to have small-scale noisiness that is visually distracting and meteorologically meaningless. An exception to this may be in mountainous regions such as the western United States, where there could be realistic and small-scale geographic variations of POP12 related to terrain features. Hence, it would be desirable to provide some smoothing of the POP12 forecasts, with more smoothing applied in the flatter central and eastern United States than in the mountainous western United States. The smoothing should also preserve the character of coherent maxima.

Savitzky–Golay (S-G) smoothing, described and justified in Press et al. (1992), is a suitable algorithm for smoothing. As opposed to boxcar smoothers (taking the arithmetic average of surrounding grid points), the S-G smoothing fits a local polynomial, and if higher-order polynomials are chosen by the user, then the S-G smoother can preserve much of the amplitude of even small-scale coherent features while smoothing incoherent ones. For this application, the S-G smoothing was applied to the 2D array of POP12 forecasts fitting a

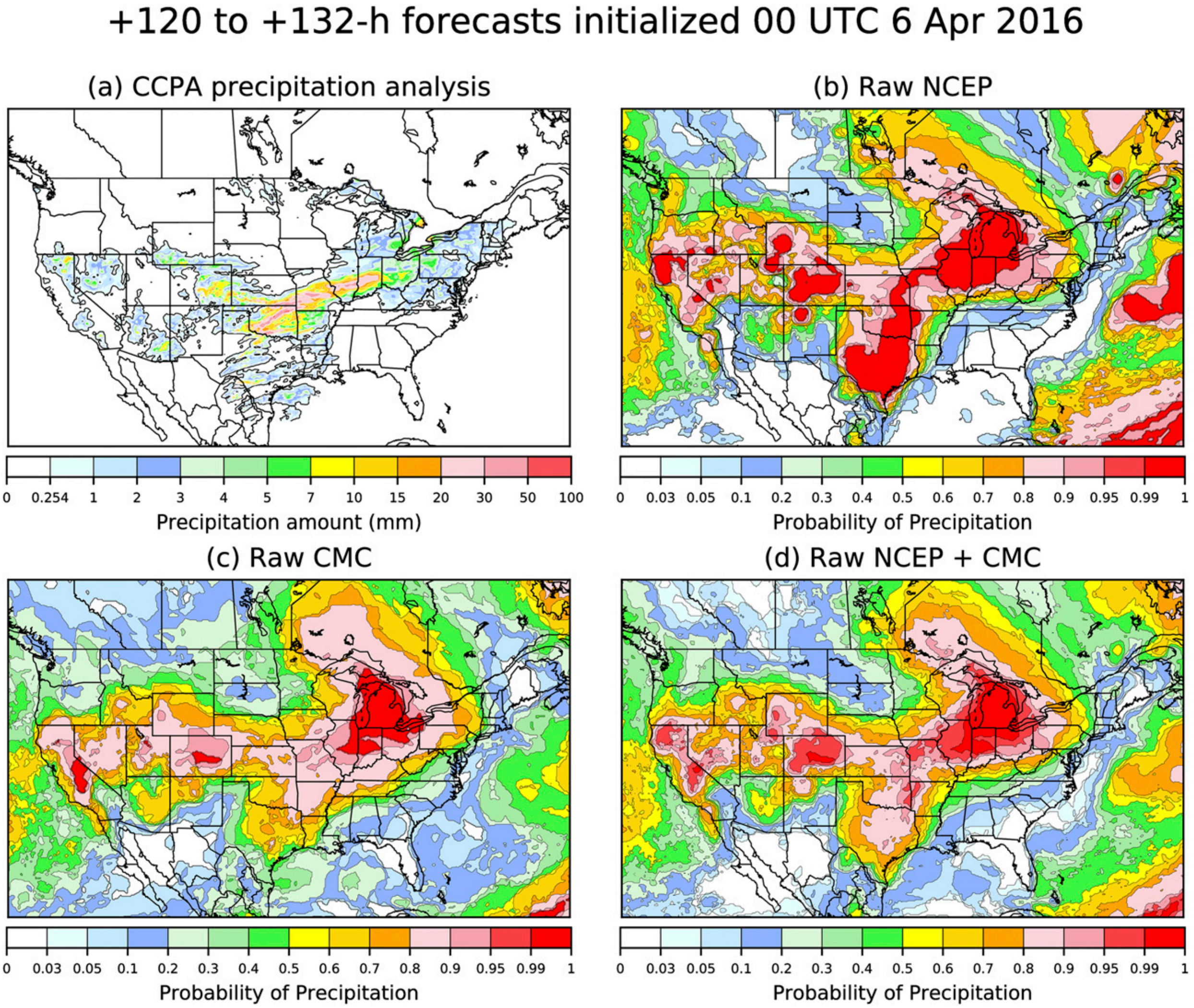
+120 to +132-h forecasts initialized 00 UTC 6 Apr 2016
(a) CCPA precipitation analysis
(b) Raw NCEP
(c) Raw CMC
(d) Raw NCEP + CMC
0 0.254 1 2 3 4 5 7 10 15 20 30 50 100
Precipitation amount (mm)
0 0.03 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95 0.99 1
Probability of Precipitation

FIG. 6. Case study illustrating the steps of the POP12 algorithm, here for +120- to +132-h forecasts initialized at 0000 UTC 6 Apr 2016. (a) The verifying precipitation analysis; areas inside the black contour exceed the 0.254 mm $(12\,\mathrm{h})^{-1}$ POP threshold. The POP12 results derived from ensemble relative frequency from the (b) NCEP and (c) CMC systems. (d) The combined NCEP + CMC raw multimodel ensemble POP.

third-order polynomial using data in a region ± four grid points around the grid point of interest. More details on the algorithmic specifics are provided in file B of the online supplement.

With the raw POP12 field and the S-G smoothed field, the final POP12 is generated from a linear combination of the two, with more weight applied to the S-G smoothed field in regions with flatter terrain. Again, more details are presented in file B of the online supplement.

Figures 6 and 7 present a case study that illustrates each major step of the postprocessing. Data shown are for +120- to +132-h forecasts initialized at 0000 UTC 6 April 2016. First, Fig. 6a shows the analyzed precipitation amounts used for verification. Areas inside the black contour are above 0.254 mm. Figures 6b and 6c show the POP12 forecasts from the NCEP GEFS and CMC systems, respectively, determined from the ensemble relative frequency. Both the NCEP and CMC ensembles, even at this advanced lead time, have high POP12 probabilities covering much of the country, including the mountainous western United States. The deterministic forecasts from each center are not shown. Figure 6d next shows the MME POP12, combining the data from the raw ensembles. The probabilities are not as sharp, for the areas where the NCEP and CMC systems have their highest probabilities differ somewhat. For example, the NCEP system has 100% probabilities over much of Texas while the CMC system does not. As will be shown in

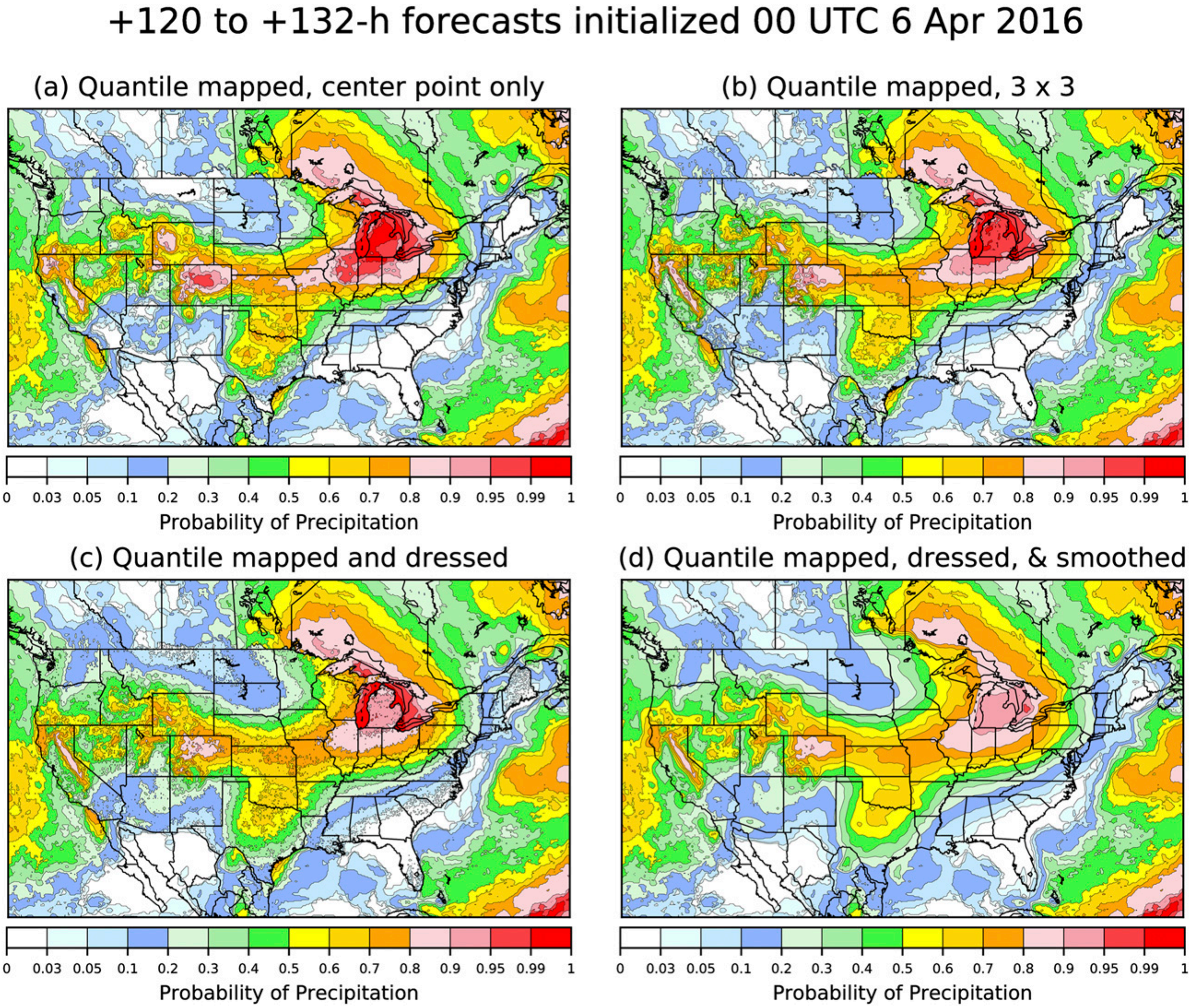
+120 to +132-h forecasts initialized 00 UTC 6 Apr 2016
(a) Quantile mapped, center point only
(b) Quantile mapped, 3 x 3
(c) Quantile mapped and dressed
(d) Quantile mapped, dressed, & smoothed
0 0.03 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95 0.99 1
Probability of Precipitation

FIG. 7. A continuation of the case study illustrating steps in the POP12 algorithm, here for +120- to +132-h 789+ forecasts initialized at 0000 UTC 6 Apr 2016. The POP12 results after (a) the application of deterministic quantile mapping using only the grid point in question, (b) the application of deterministic quantile mapping using the 3 × 3 stencil of grid points, (c) dressing of each member with Gaussian noise, and (d) the final S-G smoothing.

section 7, however, POP12's from the raw MME are still too sharp. Figure 7a illustrates the POP12 after application of deterministic quantile mapping using only the central grid point, and Fig. 7b shows POP12 when using the full 3 × 3 stencil of points when quantile mapping. The use of the 3 × 3 stencil generally lowers high probabilities slightly and provides somewhat greater accentuation of the terrain-related variability of POP12 in the western United States. Figure 7c shows POP12 after dressing with the random normal noise. Small-scale noise is introduced, along with a further reduction of the high-end POPs and some elevation of low-end POPs. Finally, Fig. 7d shows the end product, after application of the location-dependent S-G smoothing. Much of the small-scale variability outside of the mountainous regions of the western United States has been reduced.

6. The 6-hourly quantitative precipitation forecast algorithm

Before providing a description of the details of QPF06, it is worth considering the changes in precipitation characteristics we can expect from ensemble averaging, as the ensemble mean might be considered as a surrogate deterministic forecast. Figure 8a, inspired by a similar figure in Ravela et al. (2007), shows a synthetic ensemble of precipitation forecasts with different east–west positions and slightly different amplitudes. Presumably, this bears some resemblance to

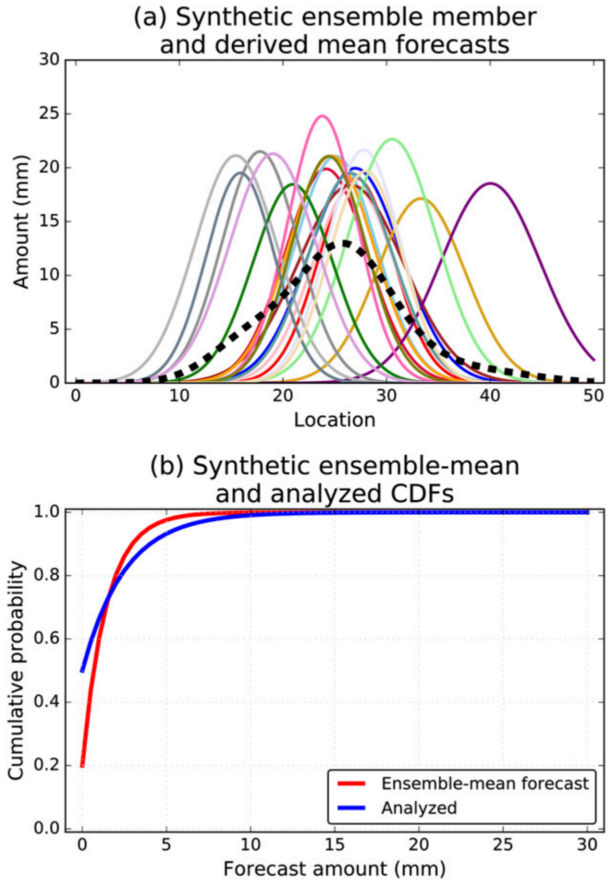

FIG. 8. (a) Illustration of an ensemble of synthetic precipitation amounts along a segment of a latitude circle that spans a forecast heavy-precipitation event. Ensemble members (solid colored curves) differ somewhat in position and amplitude. The ensemble-mean amount is also shown by the heavier dashed black curve. (b) Illustration of the shapes typical of CDFs from ensemble-mean forecasts and analyzed states.

what forecasters might see in a typical medium-range ensemble forecast. Let us assume the truth could be any one of these ensemble members; that is, the ensemble and truth are assumed to be exchangeable. Because of the diversity of ensemble positions, the mean forecast in Fig. 8a underestimates the amplitudes of the maximum relative to individual members and the potential truth, and the span of the east–west position with light precipitation in the ensemble mean is broader in scale than the members or the truth. Hence, while in ideal situations the ensemble-mean forecast will minimize the root-mean-square (RMS) error, tallied over many cases it does not provide the forecaster with a reasonable estimate of the potential magnitude of the heaviest precipitation, and it forecasts too wide a region with nonzero precipitation.

Underlying the QPF methodology developed here is the assumption that a forecaster cares less about the minimization of error and instead seeks from deterministic guidance the following: (i) an estimate of the most likely position of a precipitation maximum and (ii) an accurate estimate of the maximum precipitation possible. Further, we assume that a forecaster implicitly understands the potential for some position error and is able to convey the spatial uncertainty in associated worded discussions. In a statistical sense, the forecaster is assumed to prefer deterministic guidance that is unbiased (Wilks 2011, Eq. 8.10); that is, for any particular precipitation amount, the expected areal coverage of the forecast exceeding that amount is equal to the areal coverage of the analyzed exceeding that amount.

A relatively simple method is now described that leverages the POP12 technology, in particular the deterministic quantile mapping of Eq. (3). Rather than quantile mapping an individual ensemble member using CDFs populated with member forecast data and analyzed data, we simply quantile map from the *ensemble-mean state* to the analyzed state. Figure 8b illustrates what might occur with such a quantile mapping approach, giving the characteristics of the ensemble mean noted in Fig. 8a. Notice that there are fewer zero-precipitation events in the ensemble mean and also fewer heavier-precipitation events. Consequently, a relatively light ensemble-mean amount may be quantile mapped to a zero precipitation, and a moderate ensemble-mean amount may be mapped to a much heavier amount.

The QPF06 algorithm is thus as follows. 1) Populate the CDFs for the ensemble-mean forecast and analyzed data using the past 60 days and the supplemental location data. 2) At each grid point, quantile map the current ensemble-mean forecast values using the forecast and analyzed CDFs generated in step 1. No 3 × 3 stencil of surrounding points is used for the QPF06 application. 3) Apply the same S-G smoothing procedure to the ensemble-mean forecast as was applied to the POP12s, providing more smoothing in areas of flatter terrain.

An example of the process is shown in a case study (Fig. 9). Figure 9a shows the ensemble-mean forecast, here for +126- to +132-h forecasts initialized at 0000 UTC 6 April 2016. Larger ensemble-mean precipitation amounts are forecast in north-central Colorado and southeastern Kansas, but these mean forecasts are 7–10 mm. The northwestern United States is covered with a broad shield of lighter ensemble-mean precipitation amounts. After quantile mapping of the mean forecast, the maximum in northern Colorado is increased to 20–30 mm, and the maximum in southeast Kansas is increased to 10–15 mm. The area with nonzero precipitation in the northwest United States is substantially decreased. The subsequent smoothing of the features in Fig. 9c does not greatly change the look of the forecast product. Finally, comparing against the

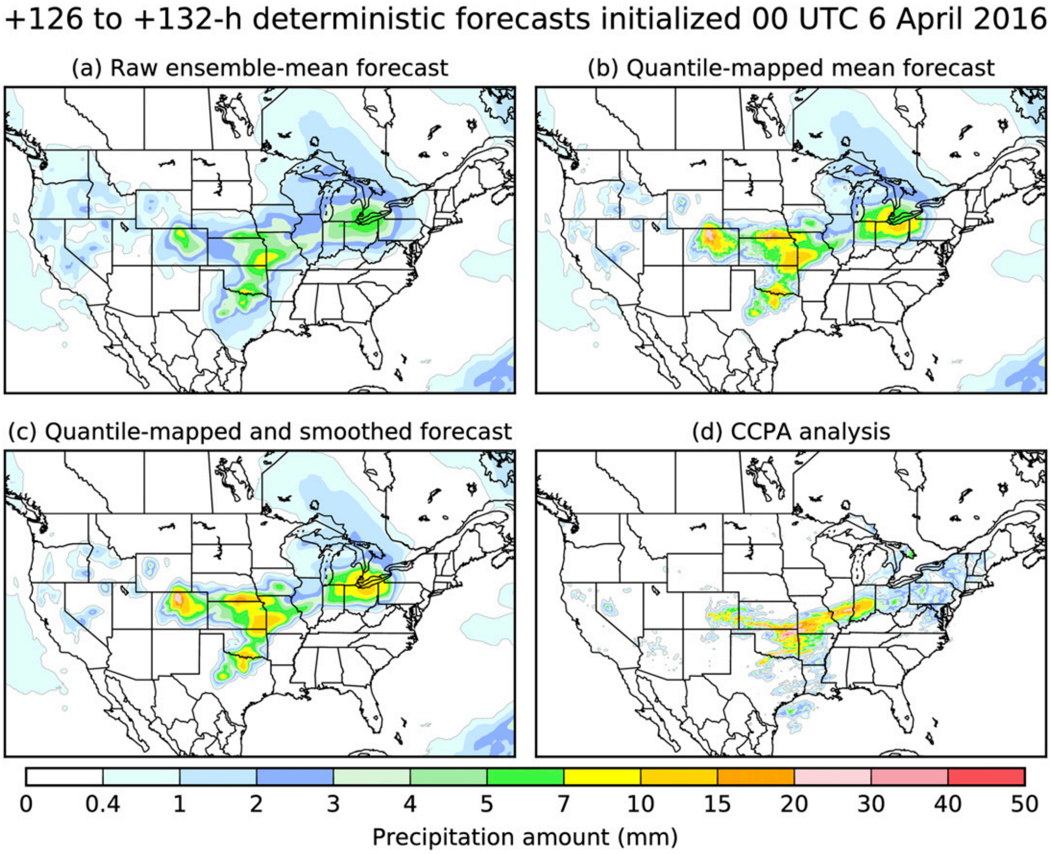
+126 to +132-h deterministic forecasts initialized 00 UTC 6 April 2016
(a) Raw ensemble-mean forecast
(b) Quantile-mapped mean forecast
(c) Quantile-mapped and smoothed forecast
(d) CCPA analysis
0
0.4
1
2
3
4
5
7
10
15
20
30
40
50
Precipitation amount (mm)

FIG. 9. An example illustrating the steps in the production of a quantile-mapped deterministic +126- to +132-h- forecast of the 6-hourly quantitative precipitation. (a) Raw, multimodel ensemble-mean forecast. (b) Quantile mapping of the ensemble mean. (c) After spatial smoothing of the quantile-mapped forecast, and (d) the verifying analysis, for comparison.

verification in Fig. 9d, we see that the quantile-mapped forecasts incorrectly placed the maximum in northern Colorado; the closest associated analyzed maximum was ~20 mm in southeast Colorado. Precipitation in excess of 50 mm was analyzed in northeast Oklahoma, close to the quantile-mapped maximum in southeast Kansas. Quantile mapping increased amounts by 50% or more in the region, yet this was not enough. The less widespread precipitation in the northwest United States produced a better degree of correspondence with the analyses. Overall, there appears to be a greater similarity of the quantile-mapped and smoothed forecasts to the analyzed data than for the ensemble mean. We also note that deterministic precipitation forecasting for such long leads is notoriously difficult; probabilistic methods at these advanced leads is preferable, given the substantial growth of chaotic errors in numerical precipitation forecasts.

This methodology was inspired by a related method, the ''probability-matched mean'' described by Ebert (2001). There are differences, though. In the current approach the ensemble mean is quantile mapped to analyzed data, whereas in Ebert's approach, the mean is mapped to resemble the original forecast-member data.

7. Objective forecast verification

We first perform a basic verification of the POP12 forecasts using reliability diagrams and BSS*s*. Figures 10–12 provide reliability diagrams and BSSs for forecasts of leads +12 to +24, +84 to +96, and +156 to +168 h, respectively. The top panels in each figure show the reliability of raw individual models and the multimodel guidance verified against the 1/8° analyses. The associated BSSs are also noted in the diagrams. The bottom panels show reliability after postprocessing, with center-point (panel d) and 3 × 3 (panel e) quantile mapping, then with subsequent dressing (panel f) and smoothing (panel g). The general unreliability of the raw guidance is quite evident, and the multimodel combination only provides an improvement over the better of the two systems at the longer leads. After center-point quantile mapping, both the reliability and

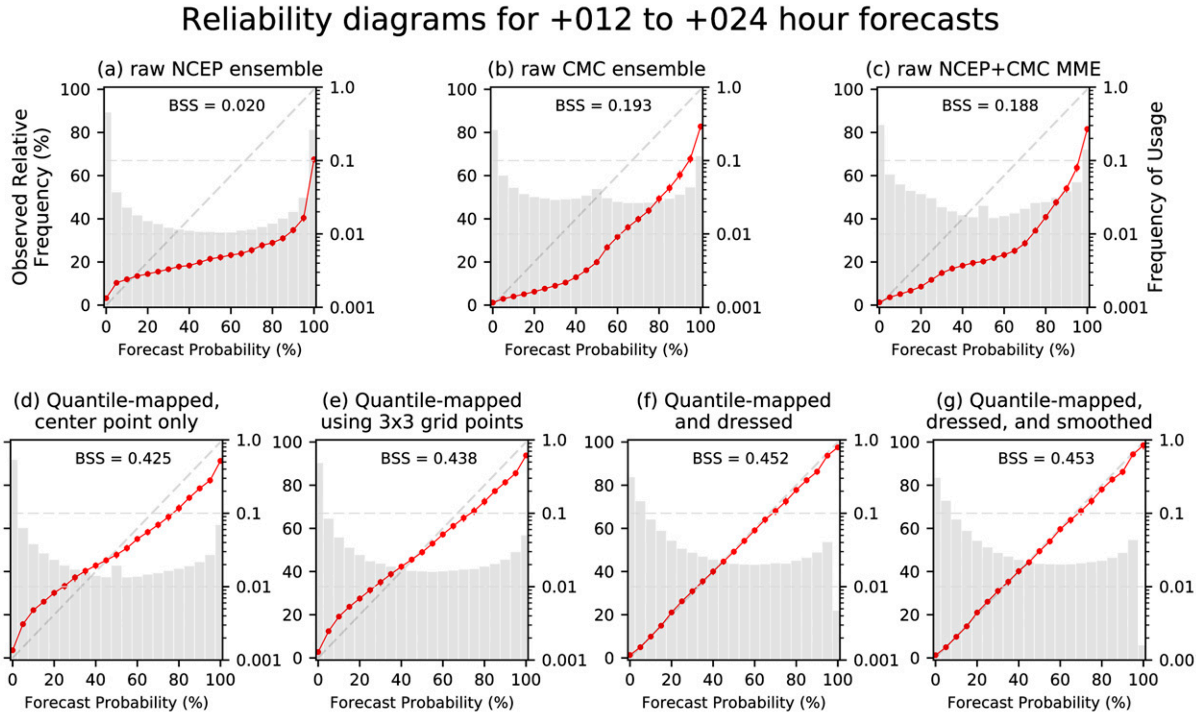

FIG. 10. Reliability diagrams (red curves) for +12- to +24-h POP12 forecasts over the CONUS, providing the observed relative frequency for a given forecast probability. Error bars represent the 5th and 95th percentiles from a 1000-sample bootstrap distribution generated by sampling case days with replacement. Histograms in gray show the overall frequency with which forecasts of a given probability are issued (scale on the right), and BSSs are noted. (a) Raw NCEP ensemble forecasts, (b) raw CMC ensemble forecasts, (c) raw multimodel ensemble forecasts, (d) postprocessed guidance after stochastic quantile mapping using the center point only, (e) after stochastic quantile mapping using 3 × 3 stencil of points, (f) after dressing, and (g) after smoothing.

skill are much improved, though there is some underforecasting of the POPs at low probabilities and some overforecast tendency at higher probabilities. Application of quantile mapping using the 3 × 3 stencil further improves the reliability and skill slightly. Dressing improves the reliability substantially and the skill only slightly, as much of the increased reliability has come about through a corresponding decrease in forecast sharpness. Smoothing has minimal effect on reliability and skill, though as shown in Fig. 7d, the unrealistic small-scale spatial detail in nonmountainous regions is dramatically reduced.

We note that these results represent the verification against the ⅛° CCPA analyses. Results of verification against station observations were also performed (not shown), and there was slightly less reliability against these point measurements.

Why was there a slight lack of reliability at high forecast probabilities? We hypothesize but cannot confirm that this may be a consequence of the quantile mapping using the last 60 days of training data in the presence of seasonally dependent model biases; that is, the quantile mappings were somewhat inaccurate in their adjustment for amount-dependent bias. Also, the structure of the stochastic noise added was determined more through trial and error than through objective verification. We expect to addressed this in subsequent research.

We have not developed another algorithm in the literature to use as a basis for comparison. However, one can get a sense of the skill improvement of these forecasts by comparing against previously published results using GEFS reforecasts, albeit for a different period. In Hamill et al. (2015), GEFS reforecasts during the 2002–13 period were postprocessed (cross validated) with an analog approach and using the same ⅛° CCPA data. In that study, the GEFS BSSs for exceeding $> 1\,\mathrm{mm}\,(12\,\mathrm{h})^{-1}$ averaged over April–June[2] were ~0.38 at 12–24 h, 0.21 at 84–96 h, and 0.07 at 156–168 h. These are compared to the current 0.453, 0.286, and 0.129 values, respectively (Figs. 8–10). The improvements from leveraging multimodel ensemble data here, plus a methodology tailored to exploit the most from small sample sizes, appeared to result in an improvement

[2] Previous experiments (not shown here) have indicated that BSSs for the >0.254 and the >1.0 mm events were quite similar for postprocessed guidance from GEFS reforecasts.

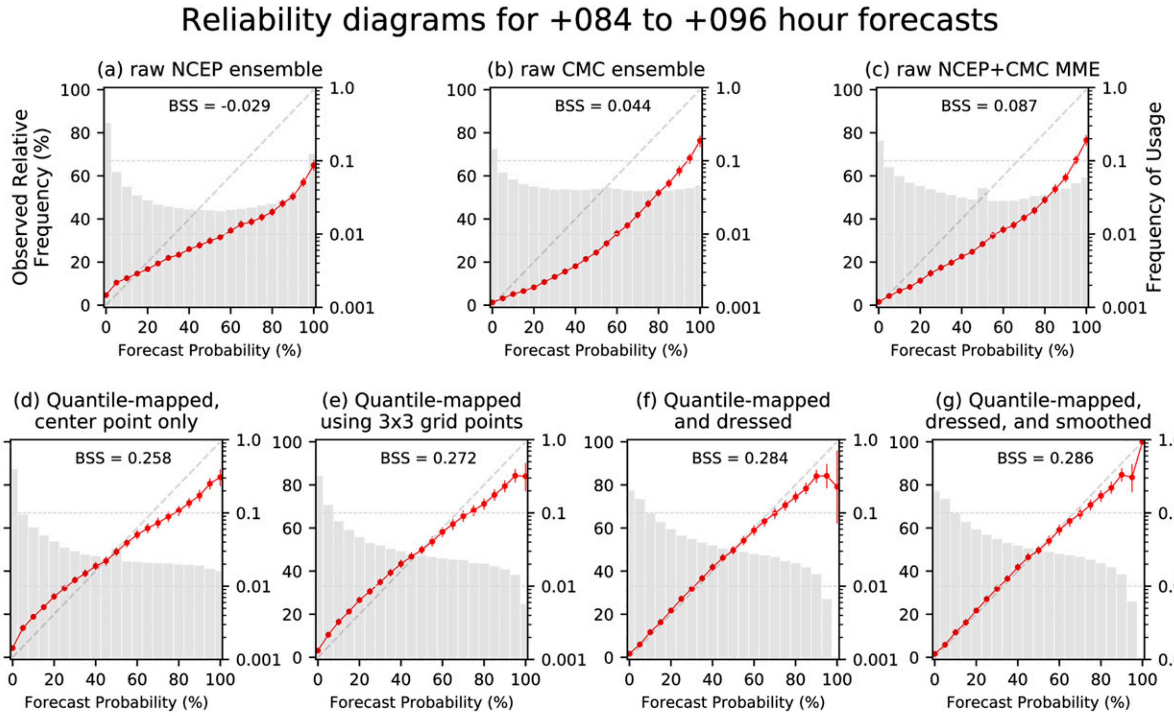

FIG. 11. As in Fig. 10, but for +84- to +96-h forecasts.

with respect to reforecast-based, single-model postprocessing, at least for the POP12 event analyzed here.

Now consider the verification of deterministic forecasts (Fig. 13). Equitable threat scores are improved slightly with the postprocessing for higher precipitation amounts, though there is a slight diminishment, for example, with lighter precipitation at moderate leads (Fig. 13b). We note that the ETS has a tendency to reward skill for the overforecasting of events, presuming there is some relationship of the observed to the forecast position (Hamill 1999). In some sense, the postprocessed guidance represents a good-faith effort to adjust the deterministic ensemble-mean forecast so the bias of the adjusted precipitation is near 1.0 regardless of the event. As can be seen, this appears to increase its ETS slightly for the higher precipitation amounts. Biases are not exactly 1.0 for all events, in part because of the smoothing of the forecasts and in part because of the regression quantile mapping at extreme high terciles of the forecast, as described in section 5c.

8. Discussion and conclusions

The NOAA/NWS National Blend of Models project is intended to provide objective, nationally consistent postprocessed guidance for key weather elements in the NWS National Digital Forecast Database. This article described the postprocessing methodologies for 12-hourly probability of precipitation (POP12) and 6-hourly deterministic quantitative precipitation (QPF06). Model guidance from global deterministic and ensemble prediction systems from the National Weather Service and Environment Canada (now known as Environment and Climate Change Canada) were used. The forecasts from these systems were postprocessed through a procedure known as quantile mapping, a procedure that permits amount-dependent bias corrections based on the differences between forecast and analyzed cumulative precipitation distributions. Because of the limited training data available due to possibly frequent model changes (training data were limited to the previous 60 days in this application), the underlying cumulative distribution functions (CDFs) would be highly noisy if the CDFs were populated independently for each grid point. Accordingly, the approach demonstrated here also used forecasts and analyses from supplemental locations to populate the CDFs. Supplemental locations are other grid points that were expected to have similar forecast bias characteristics as a result of the similarity of precipitation climatology and terrain features. Other somewhat novel features of the POP12 algorithm included (i) the synthetic enlargement of ensemble data by quantile mapping data from a 3 × 3 stencil of surrounding grid points, (ii) dressing of the ensemble with amount-dependent random noise to increase the spread of the ensemble, and

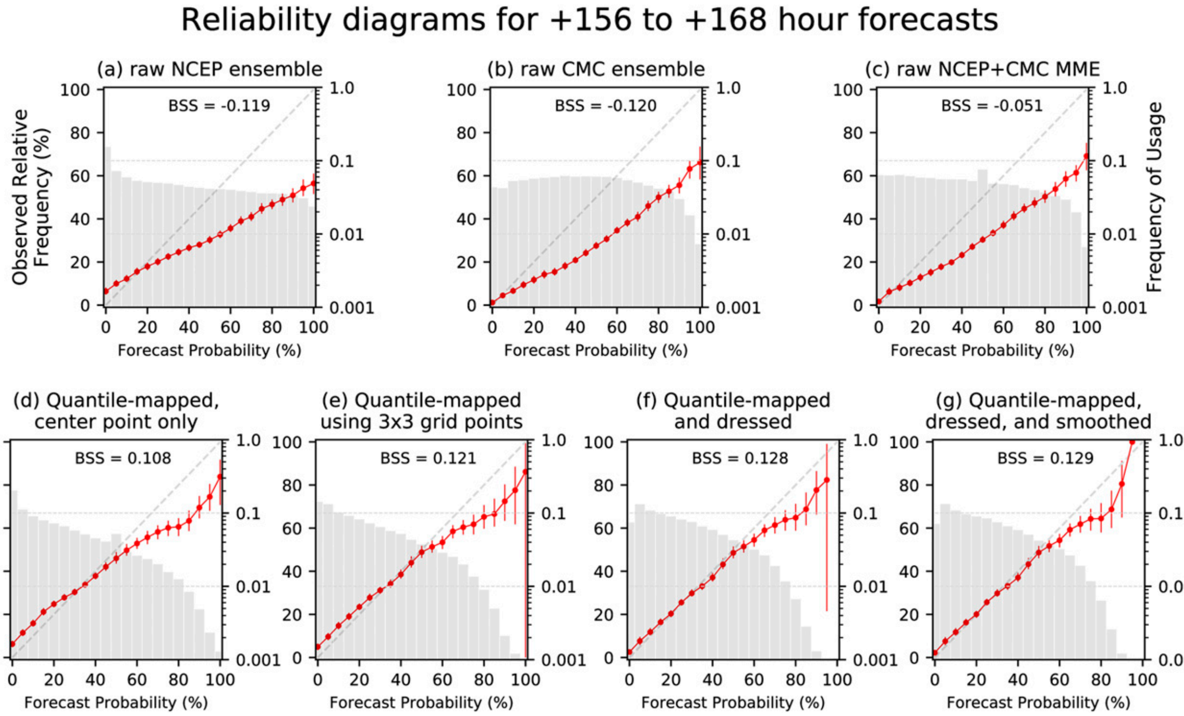

FIG. 12. As in Fig. 10, but for +156- to +168-h forecasts.

(iii) a location-dependent smoothing of the POP12 field, with more smoothing applied in regions of flat terrain.

The QPF06 procedure also leveraged the supplemental locations and the location-dependent smoothing, but it omitted dressing after the quantile mapping and the use of the 3 × 3 stencil.

Case studies and objective verification results were presented for both POP12 and QPF06. These showed that the postprocessing generates much more skillful POP12 guidance and much improved reliability. The postprocessed forecasts also provided a statistical downscaling, accentuating forecast POP12 and QPF06 in the high terrain of the western United States. QPF06 forecasts were much less biased with respect to the analyzed data, with similar or slightly improved threat scores.

National Blend developers intend to continue to attempt to improve upon the POP12 and QPF06 guidance. For example, the algorithm described here did not yet leverage shorter-range, higher-resolution forecasts from prediction systems such as NOAA's High-Resolution Rapid Refresh (Benjamin et al. 2016). In the future, inclusion of this data is anticipated, either using the methodologies described here or indirectly (postprocessed in some other way, and then combined with precipitation estimates through this system). It is also anticipated that guidance will expand to cover other U.S. areas of interest such as Alaska, Hawaii, and Puerto Rico in the next year or so. Temporal resolution may increase over the CONUS as well.

We are cognizant of the lack of a rigorous experimental control here, though (with some caveats) the skill of this method appeared to beat those from the postprocessing of GEFS reforecasts. The lack of a more rigorous control is in part because of the relatively unique nature of this work, which intended to produce national guidance at high spatial resolution using only a limited amount of training data from coarser-resolution multimodel ensembles. In part, this is a result of the lack of publicly available algorithms in our software language (Fortran). Pending future funding, we do hope to advance this current methodology and to adapt other advanced methodologies such as the one described in Scheuerer and Hamill (2015) and compare forecasts from these systems.

Another shortcoming is that, for brevity, probabilistic forecasts were not generated nor verified for events other than POP12. We presented no evidence here that this methodology is suitable for, say, predicting events such as ≥25 mm $(12\,\mathrm{h})^{-1}$. Past experience has shown that it is much more challenging to postprocess the more extreme events with small training sample sizes (ibid).

Finally, we acknowledge that there are adjustable parameters in this postprocessing method that were set

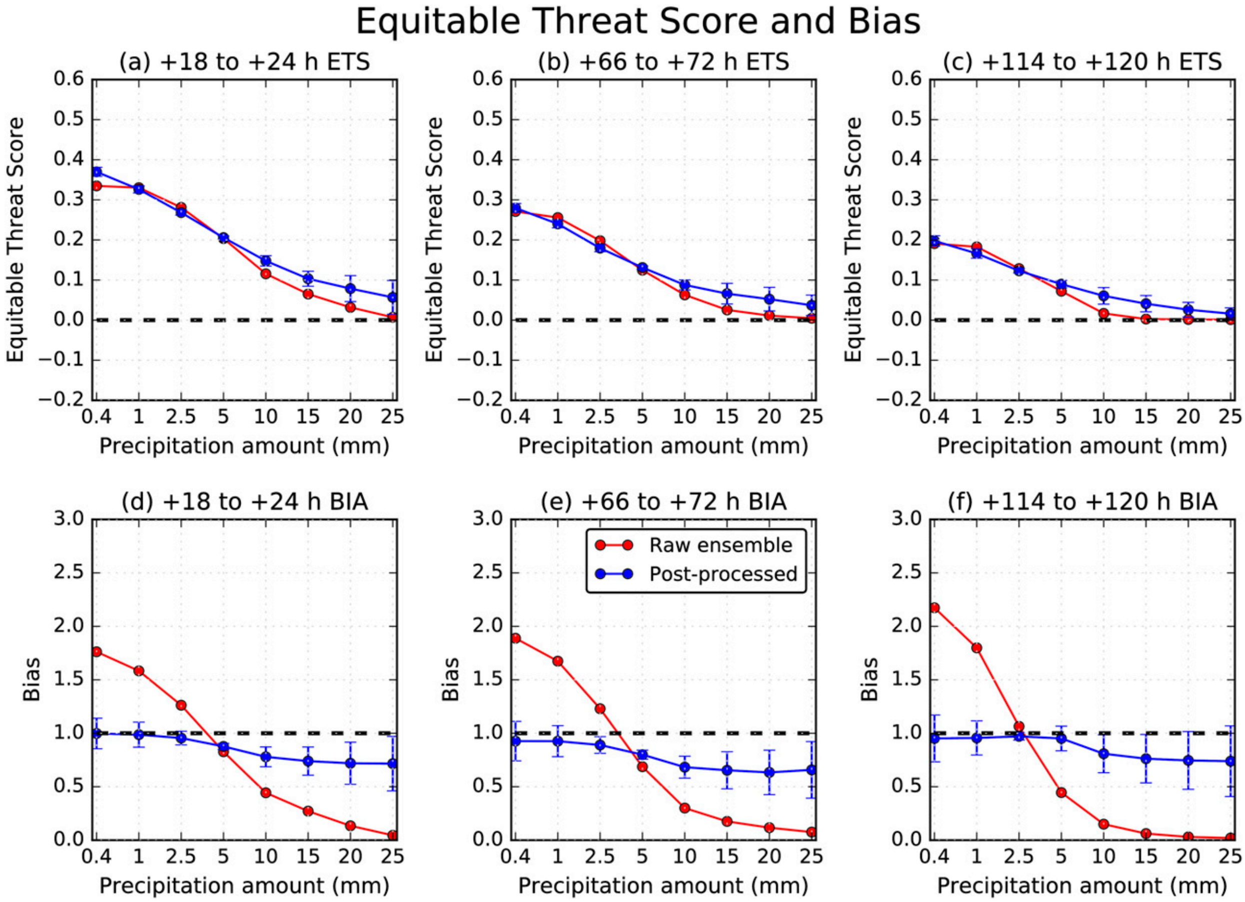

FIG. 13. (top) ETSs and (bottom) biases (BIA) from the raw multimodel ensemble (red) and quantile-mapped and smoothed forecasts (blue). ETSs for the (a) +18 to +24-, (b) +66- to +72-, and (c) +114- to +120-h forecasts. (d)– (f) As in (a)–(c), but for BIAs. Error bars represent the 5th and 95th percentiles of a 1000-replication paired block bootstrap (blocking on daily data), following Hamill (1999).

in a trial-and-error approach. These include the spacing between grid points of the 3 × 3 array of data used to augment the sample size (see Fig. 3 and section 5c) and the magnitude of the stochastic noise added (Fig. 5). In the future, it may be possible to use technologies such as feature calibration and alignment (Nehrkorn et al. 2014) to estimate the typical magnitude of the displacement of the forecast features, and how they vary with forecast lead time. This could be used to set the spacing parameter in this methodology.

In the longer term, NOAA intends to regularly produce reforecast datasets for its global ensemble prediction system, and National Blend developers hope to leverage similar datasets from international partners. With these datasets, we expect to be able to improve upon features of this algorithm, such as defining the CDFs more precisely, or to leverage or design more sophisticated postprocessing algorithms that exploit the rich reforecast data to improve skill and reliability.

Acknowledgments. The authors thank Chris Daly of Oregon State University for his advice on aspects such as the use of terrain facets. Jeff Craven (NWS Central Region) provided extensive case studies of early versions of this algorithm that were helpful in detecting and correcting problems. This work was supported by funding provided to NOAA/OAR via the High-Impact Weather Prediction Project and to NOAA/NWS via the Sandy Supplemental, both under the Disaster Relief Appropriations Act of 2013 (the latter with Grant NA14NWS4830005). Michael Scheuerer's participation was supported by NOAA Next-Generation Global Prediction System Grant N8MWQNG.

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, and A. Lavagnin, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, doi:10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

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

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, doi:10.1175/MWR-D-15-0242.1.

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.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.*, **137**, 1655–1665, doi:10.1175/2008MWR2682.1.

Charba, J. P., and F. G. Samplatsky, 2011a: Regionalization in fine-grid GFS MOS 6-h quantitative precipitation forecasts. *Mon. Wea. Rev.*, **139**, 24–38, doi:10.1175/2010MWR2926.1.

——, and ——, 2011b: High-resolution GFS-based MOS quantitative precipitation forecasts on a 4-km grid. *Mon. Wea. Rev.*, **139**, 39–68, doi:10.1175/2010MWR3224.1.

Coté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998a: The operational CMC–MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation. *Mon. Wea. Rev.*, **126**, 1373–1395, doi:10.1175/1520-0493(1998)126<1373:TOCMGE>2.0.CO;2.

——, J.-G. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998b: The operational CMC–MRB Global Environmental Multiscale (GEM) model. Part II: Results. *Mon. Wea. Rev.*, **126**, 1397–1418, doi:10.1175/1520-0493(1998)126<1397:TOCMGE>2.0.CO;2.

Craven, J. P., J. Wiedenfeld, J. Gagan, P. Browning, A. Just, and C. Grief, 2013: The NWS Central Region extended forecast process. Preprints, *38th NWA Annual Meeting*, Charleston, SC, National Weather Association, P2.38.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, doi:10.1002/joc.1688.

Ebert, E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, doi:10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, doi:10.1175/WAF843.1.

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, doi:10.1256/qj.05.167.

Gagnon, N., X. Deng, P. L. Houtekamer, S. Beauregard, A. Erfani, M. Charron, R. Lahlo, and J. Marcoux, 2014: Improvements to the Global Ensemble Prediction System (GEPS) from version 3.1.0 to version 4.0.0. Environment Canada Tech. Note, 49 pp. [Available online at http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/technote_geps-400_20141118_e.pdf.]

——, and Coauthors, 2015: Improvements to the Global Ensemble Prediction System from version 4.0.1 to version 4.1.1. Environment Canada Tech. Note, 29 pp. [Available online at http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/technote_geps-411_20151215_e.pdf.]

Glahn, H. R., and D. P. Ruth, 2003: The New Digital Forecast Database of the National Weather Service. *Bull. Amer. Meteor. Soc.*, **84**, 195–201, doi:10.1175/BAMS-84-2-195.

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

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

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

——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, doi:10.1256/qj.06.25.

——, 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.

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

Hopson, T. M., and P. J. Webster, 2010: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *J. Hydrometeor.*, **11**, 618–641, doi:10.1175/2009JHM1006.1.

Hou, D., Z. Toth, Y. Zhu, and W. Yang, 2008: Impact of a stochastic perturbation scheme on global ensemble forecast. *19th Conf. on Probability and Statistics*, New Orleans, LA, Amer. Meteor. Soc., 1.1. [Available online at https://ams.confex.com/ams/88Annual/techprogram/paper_134165.htm.]

——, 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., B. He, and H. L. Mitchell, 2014: Parallel implementation of an ensemble Kalman filter. *Mon. Wea. Rev.*, **142**, 1163–1182, doi:10.1175/MWR-D-13-00011.1.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass, and E. Grimit, 2011: Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Wea. Rev.*, **139**, 2630–2649, doi:10.1175/2010MWR3511.1.

Kleist, D. T., and K. Ide, 2015a: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Mon. Wea. Rev.*, **143**, 433–451, doi:10.1175/MWR-D-13-00351.1.

——, and ——, 2015b: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS.

Part II: 4D EnVar and hybrid variants. *Mon. Wea. Rev.*, **143**, 452–470, doi:10.1175/MWR-D-13-00350.1.

Lerch, S. and S. Baran, S., 2017: Similarity-based semilocal estimation of post-processing models. *J. Roy. Stat. Soc.*, **66C**, 29–51, doi:10.1111/rssc.12153.

Liu, J., and Z. Xie, 2014: BMA probabilistic quantitative precipitation forecasting over the Huaihe basin using TIGGE multimodel ensemble forecasts. *Mon. Wea. Rev.*, **142**, 1542–1555, doi:10.1175/MWR-D-13-00031.1.

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *J. Climate*, **26**, 2137–2143, doi:10.1175/JCLI-D-12-00821.1.

Mass, C. F., J. Baars, G. Wedam, E. Grimit, and R. Steed, 2008: Removal of systematic model bias on a model grid. *Wea. Forecasting*, **23**, 438–459, doi:10.1175/2007WAF2006117.1.

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

Moore, B., K. Mahoney, E. Sukovich, R. Cifelli, and T. Hamill, 2015: Climatology and environmental characteristics of extreme precipitation events in the southeastern United States. *Mon. Wea. Rev.*, **143**, 718–741, doi:10.1175/MWR-D-14-00065.1.

Nehrkorn, T., B. Woods, T. Auligné, and R. N. Hoffman, 2014: Application of feature calibration and alignment to high-resolution analysis: Examples using observations sensitive to cloud and water vapor. *Mon. Wea. Rev.*, **142**, 686–702, doi:10.1175/MWR-D-13-00164.1.

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.

Ravela, S., K. Emanuel, and D. McLaughlin, 2007: Data assimilation by field alignment. *Physica D*, **230**, 127–145, doi:10.1016/j.physd.2006.09.035.

Roulin, E., and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Wea. Rev.*, **140**, 874–888, doi:10.1175/MWR-D-11-00062.1.

Roulston, M., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, doi:10.3402/tellusa.v55i1.12082.

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.

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

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 forecasts. *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.

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

Verkade, J. S., J. D. Brown, P. Reggiani, and A. H. Weerts, 2013: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *J. Hydrol.*, **501**, 73–91, doi:10.1016/j.jhydrol.2013.07.039.

Vislocky, R. L., and J. M. Fritsch, 1997: Performance of an advanced MOS system in the 1996–97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.*, **78**, 2851–2857, doi:10.1175/1520-0477(1997)078<2851:POAAMS>2.0.CO;2.

Voisin, N., J. C. Schaake, and D. P. Lettenmaier, 2010: Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 1603–1627, doi:10.1175/2010WAF2222367.1.

Wilks, D. S., 2006: Comparison of ensemble–MOS methods in the Lorenz ’96 setting. *Meteor. Appl.*, **13**, 246–256, doi:10.1017/S1350482706002192.

——, 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. Elsevier, 676 pp.