Improving National Blend of Models Probabilistic Precipitation Forecasts Using Long Time Series of Reforecasts and Precipitation Reanalyses. Part I: Methods

THOMAS M. HAMILL,[b,c] DIANA R. STOVERN,[a,b] AND LESLEY L. SMITH[a,b]

[a] *Cooperative Institute for Research in the Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *NOAA/Physical Sciences Laboratory, Boulder, Colorado*
[c] *IBM/Weather Company, Andover, Massachusetts*

(Manuscript received 8 November 2022, in final form 27 February 2023, accepted 1 March 2023)

ABSTRACT: This article describes proposed revised methods for the statistical postprocessing of precipitation amount intended for the NOAA's National Blend of Models using the Global Ensemble Forecast System version 12 data (GEFSv12). The procedure updates the previously established procedure of quantile mapping, weighting of sorted members, and dressing of the ensemble. The revised method leverages the long reforecast training dataset that has become available to improve quantile mapping of GEFSv12 data by eliminating the use of supplemental locations, that is, training data from other grid points. It establishes improved definitions of cumulative distributions through a spline-fitting approach. It provides updated algorithms for the weighting of sorted members based on closest-member histogram statistics, and it establishes an objective method for the dressing of the quantile-mapped, weighted ensemble. Verification statistics and case studies are provided in the accompanying article (Part II).

KEYWORDS: Ensembles; Forecast verification/skill; Forecasting techniques; Numerical weather prediction/forecasting; Probabilistic Quantitative Precipitation Forecasting (PQPF); Statistical forecasting

1. Introduction

On 20 September 2020, the National Oceanic and Atmospheric Administration's (NOAA's) Global Ensemble Forecast System, version 12 (GEFSv12) was implemented. Many improvements were made to this ensemble prediction system with this implementation, including a new atmospheric dynamical core applied at ~1/4° grid spacing, improved stochastic physics, improved ensemble-based data assimilation, and updated physical parameterizations. The GEFSv12 system provides 31-member real-time ensemble forecasts initialized at 0600, 1200, and 1800 UTC to +16-day lead time. Once daily, from 0000 UTC initial conditions, the forecasts are integrated to +35 days. The GEFSv12 system configuration and basic verification is described in more depth in Zhou et al. (2022).

To facilitate a more comprehensive statistical postprocessing and forecast verification, the GEFSv12 implementation was accompanied by the production of a 20-yr set of reanalyses and reforecasts covering 2000–19, using a very similar data assimilation system as is used in real time and using the same GEFSv12 system. The reanalyses are retrospective gridded analyses of the weather suitable for initializing the forecast, and the reforecasts are retrospective forecasts using a statistically consistent version of the forecast model. Previous-generation reforecasts for the GEFS system were described in Hamill et al. (2004, 2006, 2013).

While the real-time medium-range GEFSv12 ensemble is produced four times daily with 31 members, the accompanying reforecasts were produced only once per day, initialized at 0000 UTC,[1] with 5 members generated to +16-day lead. Once per week, an 11-member ensemble forecast was generated to +35-day lead. The reanalysis was described in Hamill et al. (2022), and the reforecasts were described in Guan et al. (2022). Primary applications for these reforecast datasets include the calibration of 6–10- and 8–14-day forecasts for the Climate Prediction Center and the calibration of forcings for NOAA's Hydrologic Ensemble Forecast System (HEFS; Demargne et al. 2014). Many more applications are in development and technology transfer to the National Weather Service (NWS), including week-2 fire-weather potential postprocessed forecasts (Worsnop et al. 2021), subseasonal hurricane potential for the Atlantic basin (Switanek et al. 2023), week-2–4 precipitation forecasting (Scheuerer et al. 2020), and the application discussed in this paper, the statistical postprocessing of precipitation forecasts for the NWS National Blend of Models (NBM).

The intent behind the NBM (Craven et al. 2020) is to produce improved numerical guidance by combining and calibrating guidance from multiple prediction systems, both those from NOAA and when available from other agencies, including the U.S. Navy, the Canadian Meteorological Centre, and the European Centre for Medium-Range Weather Forecasts (ECMWF). The hope is that such centrally produced guidance will be of sufficient quality and detail that forecasters

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-22-0308.s1.

Corresponding author: Thomas M. Hamill, tom.hamill@ibm.com

[1] Forecasts were actually initialized at the end of a 9-h data assimilation window ending at 0300 UTC; one may think of this practically as a 0000 UTC initialization, with the first 3 h of forecasts adjusted to the incoming observations.

DOI: 10.1175/MWR-D-22-0308.1

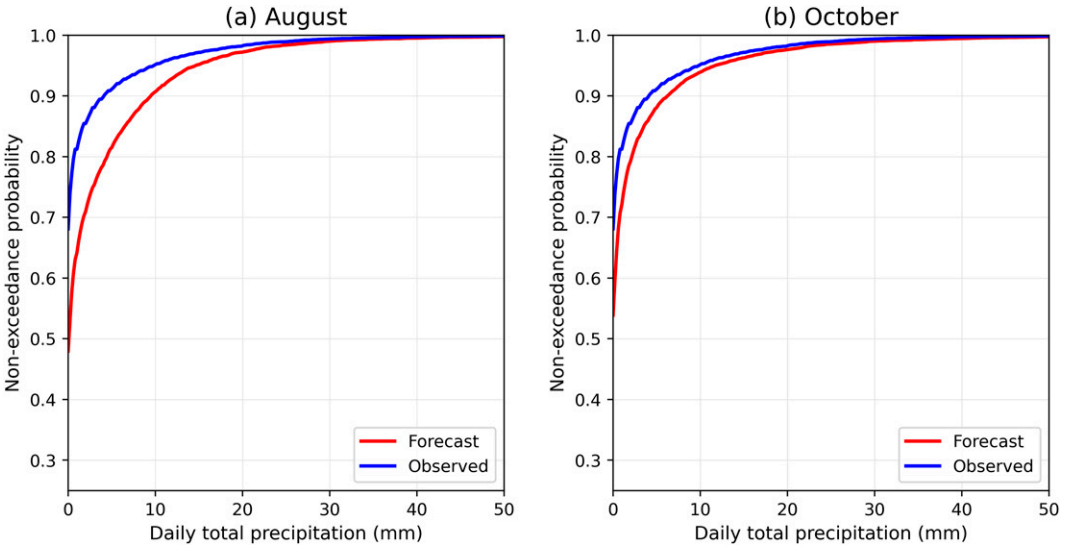

FIG. 1. (a) August and (b) October empirical cumulative distribution functions of 0–1-day accumulated GEFSv12 reforecast data at the grid point nearest to Ithaca, NY, and 24-h observed total precipitation at the Ithaca Game Farm station using 2000–19 data.

will, in most cases, see no need to manually modify the guidance, sparing much of their time to concentrate on more important work activities. In the earlier version of the NBM precipitation algorithm, even if reforecasts were available, the precipitation calibration algorithm utilized a short training dataset comprising the most recent 60 days of forecasts and precipitation analyses. The procedures that were developed for this were built around quantile mapping (Voisin et al. 2010; Hopson and Webster 2010; Maraun 2013), and the NBM application of this was described in Hamill et al. (2017, hereafter H17) and Hamill and Scheuerer (2018, hereafter HS18). The quantile-mapping technique, when incorporating data from "supplemental locations," were able to provide forecasts with improved probabilistic skill and reliability relative to the raw guidance. An additional calibration element, the variable weighting of dressed, sorted members based on closest-member histogram statistics (HS18) further improved forecast skill and reliability, especially for probabilistic forecasts with higher thresholds.

Despite the improvement of the postprocessed forecasts relative to raw guidance, the use of supplemental locations to increase training data sample size tended to reduce the amount of terrain-related precipitation detail in the western United States (J. Craven 2021, personal communication), and the use of the last 60 days of training data reduced NBM forecast skill and reliability in transition from warm to cool seasons, and vice versa. Figure 1 shows an example of why this happens, using data from one particular location, the "Game Farm" weather station at Cornell University, Ithaca, New York. Figure 1a shows the cumulative distribution functions (CDFs) of the 24-h accumulated GEFSv12 reforecasts and observations for August 2000–19, and Fig. 1b shows these for October 2000–19. As reflected in the differences between cumulative distributions, a large over-forecast bias of moderate precipitation was evident in August, where precipitation was more convective in nature. This overforecast bias was more muted in October, when precipitation was more stratiform in character. For example, a 10-mm forecast in August would be statistically adjusted with quantile mapping to a value of ~5 mm. In October, the same 10-mm forecast would be adjusted to ~8 mm. Hence, the existing NBM procedure which leverages August data in the calibration of October forecasts may be making inappropriate statistical modifications.

This article describes a proposed adaptation to previously described NBM precipitation postprocessing methods. For the sake of software continuity and maintainability, the method is designed to provide tractable modifications to the existing algorithm, and only for GEFSv12 data, where the reforecasts were available. Other conventional postprocessing procedures (e.g., Sloughter et al. 2007; Ben Bouallègue 2013; Scheuerer and Hamill 2015) machine-learning procedures such as discussed in Schumacher et al. (2021) or M. Ghazvinian, Y. Zhang, D. J. Seo, T. M. Hamill, and N. Fernando (2022, unpublished manuscript) may advance the skill beyond that of the algorithm discussed here, though such methods may lack the ease of interpretability of the current algorithm. Still, given their improved skill, especially for heavier precipitation events, one of these might be a candidate for a future incorporation into the NBM.

The proposed adaptation for the GEFSv12 system continues the previous use of quantile mapping, but the CDFs utilized in the quantile mapping are now generated through a more sophisticated spline-fitting procedure that leverages the lengthy reforecasts and high-resolution precipitation analyses. With two decades of daily forecasts potentially available, the use of supplemental locations can be abandoned; there should be adequate sample sizes from a time series of forecasts and precipitation reanalyses at each particular grid point to estimate the precipitation CDFs. Further, there is no need to use training data from a previous season; data from a coincident

season is readily available spanning the decades in question. Additional refinements are also described for the weighting of the sorted ensemble members before determining probabilities and for the dressing of weighted, quantile-mapped members. GEFSv12 reforecasts and coincident precipitation analyses will be used for the 2002–19 period (the precipitation analyses were not available prior to 2002). Raw and quantile-mapped forecasts will be compared using a larger ensemble of preproduction "retro" forecasts spanning December 2017–November 2019, and the probabilistic forecast skill from raw, quantile-mapped, quantile-mapped and weighted, and quantile-mapped, weighted, and dressed (Roulston and Smith 2003) forecasts will be evaluated.

The research hypothesis is that the quantile mapping and rank-dependent weighting using the coarser-resolution reforecast data and finer-resolution precipitation analysis data will improve skill and reliability and provide a statistical downscaling. A more specific research hypothesis underlying the use of quantile mapping applied on a gridpoint basis is that precipitation forecast biases have a strong dependency on the location on the analysis grid, a possibly nonlinear function of precipitation forecast amount, and may vary with forecast lead time, phase of the diurnal cycle, and of the month of the year. This would then permit adjustments for the tendency of some numerical prediction systems to perhaps overforecast light precipitation and underforecast heavier amounts; to adjust for possible diurnally dependent biases due to errors in the timing of convective initiation, or biases that differ between predominantly convective precipitation (warm season) and stratiform (cool season). If analyzed data used in the quantile mapping are at a higher spatial resolution than the forecast, an implicit statistical downscaling is possible. Biases could depend upon other factors such as the forecast wind direction, the thermodynamic stability, the mode of convection, and other factors that are not accounted for in the quantile-mapping procedure, and that is a potential limitation of this approach, one not quantified here.

Below, section 2 will describe the data and evaluation metrics to be used in this study. Section 3 provides a brief review of this previously defined methodology, discusses general algorithmic differences in the new approach, and then provides a detailed explanation of the revised quantile-mapping method using cubic splines as well as revised weighting and dressing. Section 4 provides some brief conclusions. An accompanying article, Stovern et al. (2023, hereafter Part II, will provide the evaluation of forecasts using the revised procedure.

2. Reforecast and precipitation analysis data

a. Reforecasts

Reforecast data from the GEFSv12 system were used in this project for building quantile-mapping relationships. To align with the period where precipitation analysis data were available, reforecasts from 1 January 2002 to 31 December 2019 over the contiguous United States (CONUS) and surrounding areas were used, with five-member reforecasts available each day from 0000 UTC initial conditions; section 3b will describe how calibration of the other cycles uses the 0000 UTC data in a way that facilitates estimation of diurnally dependent biases. For this NBM application, we used the forecasts to +10-day lead time, with accumulated precipitation saved every 6-h period. These precipitation data for each of the five reforecast members were converted to netCDF[2] files, with separate files for each forecast lead time and each calendar month. These data and code are available for others to use at ftp://ftp.cdc.noaa.gov/Projects/GEFSv12/NBM/and https://github.com/ThomasMoreHamill/GEFSv12_NBM, respectively. A given netCDF file for a given lead time and month contained all the ensemble forecasts during the 2002–19 period. As precipitation forecast bias characteristics can vary from one month or one season to the next, saving data separately for each month facilitated a month-by-month estimation of the CDFs used in the quantile-mapping process. Similarly, forecast bias may be diurnally dependent and dependent on the forecast lead, hence saving data for different lead times in different files. The five members from the reforecast included the control and the first four perturbed forecasts. These data were downloaded from the grib data store at Amazon web Services, https://noaa-gefs-retrospective.s3.amazonaws.com/index.html before spatial subsetting and conversion to netCDF format.

The GEFSv12 grid for this project surrounded CONUS, with the domain from longitude 138.5° to 59.25°W every 0.25°, and from latitude 19.25° to 57.25°N every 0.25°. This domain was large enough to include basins in Canada and Mexico that flow into rivers touching the United States, given presumed interest in using this product for hydrologic prediction.

See Zhou et al. (2022) for more documentation on the GEFSv12 system and its performance, and Guan et al. (2022) for information on the new reforecasts.

b. Precipitation analyses

For the precipitation analyses, given that the high-resolution Climatology-Calibrated Precipitation Analyses (CCPA; Hou et al. 2014) only became available in 2002, the precipitation analysis record covered the 2002–19 period. The overall gridded precipitation analysis used in these analyses was a combination of: 1) CCPA on the National Digital Forecast Database (NDFD) grid, a 2345 (x) × 1597 (y) grid surrounding the CONUS on a Lambert-conformal projection with a grid spacing of approximately 3 km, and 2) Multi-Source Weighted-Ensemble Precipitation analyses, version 2.60 (MSWEP; Beck et al. 2019), originally on a 0.05° latitude–longitude grid (~4.2 km at 40° latitude), interpolated to the 3-km NDFD grid. One or the other analysis was used for any given point. The combined analyses used CCPA for grid points inside the CONUS land boundary, interpolated MSWEP outside, and MSWEP for very few selected grid points with missing CCPA data inside the CONUS. CCPA data were also used in the Columbia River basin of Canada and tributaries of the Rio Grande River in Mexico. The combination of precipitation analyses allowed for a relatively seamless product from land to ocean.

[2] Documentation on the netCDF format can be found at https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_documentation.html.

c. Retrospective preproduction forecast data

"Retro" forecast data were preproduction GEFSv12 simulations conducted from 1 December 2017 to 30 November 2019, initialized every day at 0000 UTC, with a few case days missing. These data were used to generate closest-member histogram and dressing statistics and compare probabilistic forecasts of the raw and quantile mapped algorithm, described later. As with the real-time GEFSv12, the retro ensemble consisted of 31 members, larger than the reforecast ensemble. The 6-hourly ensemble accumulated precipitation forecasts were extracted on the same 0.25° grid as was used for the reforecast processing previously described.

Though overlapping in time with the reforecast training data, these data could be considered somewhat independent, in that initial conditions for the reforecast were generated from a lower-resolution reanalysis (Hamill et al. 2022) whereas the retro runs used the operational initial conditions generated at the time with a higher-resolution version of the forecast model. The retro-run data are also described in Hamill et al. (2022, section 4d). Subsequent figures will also show that the calibration statistics, including the spline-fitted CDFs, the closest-member histogram statistics, and dressing statistics had little variation between cross-validation samples, indicating that a more formal separation of training and verification data would likely have little impact.

3. Quantile-mapping, weighting, and dressing methodology

a. Overview of previous methodology and differences with the new method

HS18 described for the National Blend of Models precipitation postprocessing a multistep procedure that could be applied to individual ensemble prediction systems. Each coarser-resolution forecast member nearest to the finer-resolution analyzed point and each forecast at a stencil of surrounding points was quantile mapped to provide estimated values of the analyzed state. CDFs used in the quantile mapping were determined from gamma distributions fitted to the empirical distribution for points with precipitation greater than zero precipitation, with some modification described in HS18 for exceptionally wet forecasts (see section 3.b.2 of HS18). Parametric representations of the CDFs used in the quantile mapping were estimated separately for forecasts and analysis grids, separately for each forecast lead time (forecasts) or time of the day (analyses), separately for each forecast and analysis grid point, and for each day in the test period, with each day using the previous 60 days of training data.

As 60 previous days at one location would not have provided a sufficient sample to accurately estimate the CDFs, "supplemental locations" were identified for each grid point and each lead time/time of day and each month. These supplemental locations defined the locations of other grid points with similar analyzed precipitation climatologies and terrain characteristics, and the training sample used to estimate CDFs used data not only from the grid point in question but also from its defined supplemental locations.

After the quantile mapping and before determination of probabilities, for each case day and each grid point the quantile-mapped members were sorted and weighted using "closest-member histogram" statistics (HS18 and section 3d, below) and then dressed with a Gaussian probability distribution whose standard deviation depended on forecast amount. HS18 also described the synthesis into a multisystem blended forecast product.

The general equation for determining probability distributions from the weighted, quantile-mapped, dressed ensemble in the revised algorithm is as follows:

$$\Phi(x) = \sum_{(i)=1}^{n\times 25} h_{(i)} \times \Phi_N\left\{\frac{\left(x - \overline{R}[\tilde{x}^f_{(i)}]\right)^2}{2\sigma^2_{(i)}}\right\}, \quad (1)$$

where $\Phi(x)$ is the cumulative probability associated with the precipitation amount x, (i) is the sorted ensemble member in the n-member ensemble including quantile-mapped members enlarge from a surrounding array of 5×5 stencil points. The application of and rationale for the stencil are described in section 3c, below. The term $h_{(i)}$ is the closest-member histogram weight for this sorted member; $\Phi_N\{\cdot\}$ is a Gaussian-distributed kernel of probability density, where $\tilde{x}^f_{(i)}$ is the quantile-mapped value and $\overline{R}[\,\cdot\,]$ is a regression correction to the quantile-mapped value based on objectively defined dressing statistics. The term $\sigma^2_{(i)}$ is the variance of the dressing kernel. Details on all the constituent terms of this equation are described in depth later.

Table 1 provides a summary of the algorithmic changes made in processing the new GEFSv12 data relative to the procedure in HS18. All aspects of the HS18 algorithm were re-evaluated in light of the larger training sample sizes made possible with reforecasts and precipitation analyses from 2002 to the present. The long time series of reforecasts and high-resolution precipitation analyses permitted making several improvements to the quantile-mapping procedure of HS18. First and foremost, the use of supplemental locations was discarded with the availability of long time reforecast and precipitation analysis time series. CDFs instead were estimated only using the data at the forecast or analysis grid point in question, estimated with a spline-fitting procedure. Sample sizes to populate CDFs were further enlarged by using a 3-month period centered on the month of interest. Changes were also made to refine the closest-member histogram weighting and dressing components. We now discuss in more detail the algorithmic changes.

b. Revised quantile mapping

The implementation of quantile mapping discussed here requires the estimation of CDFs of precipitation identified separately for each grid point, for each month, for each forecast lead time (forecast), or for the time of day (analysis). Reading in a long time series of past forecasts and observations/analyses on high-resolution grids every time we seek to apply quantile mapping would be slow and computationally burdensome. Hence, this procedure had a separate training phase, where a compact representation of cumulative distributions was estimated, followed by a prediction phase, where the real-time

TABLE 1. Changes between the postprocessing methodology discussed in Hamill and Scheuerer (2018) and those developed for GEFSv12 with its reforecasts.

System characteristic	Method in Hamill and Scheuerer (2018)	Method in this paper
Forecast training data	GEFS version 10 and other models, last 60 days of forecasts	GEFS version 12 reforecasts, 2002–19; data for a given month uses that month and surrounding months; weighting and dressing statistics used Dec 2017–Nov 2019 retro forecast data
Analyzed training data	CCPA over CONUS, 1/8°	Merged CCPA and MSWEP precipitation analyses on the ~3-km NDFD grid surrounding the CONUS
CDFs used in quantile mapping	Fraction zero and gamma distribution for positive values	Fraction zero and (i) a spline fit for positive precipitation values when there were a large number of positive precipitation samples, or (ii) fitting with a gamma distribution (smaller number of positive samples)
Treatment of extreme precipitation	Quantile mapping to the 90th percentile, regression correction between the 90th and 99th percentile added to quantile mapped value; no additional correction beyond the 99th percentile	Quantile mapping up to the 99th percentile in distribution of samples with positive precipitation, raw ensemble plus quantile mapping at the 99th percentile for ensemble forecasts greater than he 99th
Stencil size	5 × 5, spacing increasing between points with lead time	5 × 5, spacing increasing between points with lead time
Closest-member histogram weighting	Stratified into three categories of precipitation amount	Stratified into seven categories of precipitation amount
Dressing of quantile-mapped member forecasts with Gaussian error distribution	Dressing distribution standard deviation is a linear function of the quantile-mapped member amount	Different dressing statistics for lowest, intermediate, and highest members, all of which vary as a function of quantile-mapped member amount

forecasts were corrected using those compact representations of the cumulative distributions.

Define the CDF of the ith member of an ensemble forecast at a particular grid point, lead time, and time of year as

$$F^f(x_i^f) = p_i^f = P(X^f \le x_i^f), \tag{2}$$

where x_i^f is the ith member's forecast raw precipitation amount and X^f is a random variable drawn from the climatological distribution of forecast precipitation amount appropriate to this location, lead time, and time of year. The term $F^f(x_i^f)$ defines the percentile p_i^f in this cumulative distribution associated with this forecast precipitation amount. The CDF for the analyzed amount $F^a(x^a)$ is defined similarly. There is an inverse distribution function, the quantile function, which maps percentile back to a precipitation amount. For example, with the precipitation forecast, this quantile function is

$$x_i^f = F^{f-1}(p_i^f), \quad p_i^f \in [0, 1]. \tag{3}$$

The quantile-mapping procedure generates a bias-corrected (and potentially downscaled) precipitation amount $\tilde{x}_i^f$ conditional on the forecast precipitation amount:

$$\tilde{x}_i^f = F^{a-1}[F^f(x_i^f)]. \tag{4}$$

That is, the percentile of the CDF for this grid point's ensemble member forecast amount is determined, and then the estimated analyzed amount is determined from the analyzed amount associated with the same percentile.

We envision the operational implementation of this algorithm for 0000, 0600, 1200, and 1800 UTC cycles using only 0000 UTC cycle reforecasts. To do this, when quantile mapping the forecasts from cycles other than 0000 UTC initialization, the forecast CDF data are shifted. For example, a 6–12-h forecast from the 1200 UTC cycle will span the 1800–0000 UTC period, the same as the 18–24-h forecast from the 0000 UTC cycle. In this case, the fitted CDFs of the 0000 UTC cycle, 18–24-h forecasts are substituted in the quantile mapping.

Fitted relationships between precipitation and a function of cumulative probability were determined separately for each grid point, lead time (forecast) or 6-h accumulation period in the diurnal cycle (analysis), and separately for each month. For a given month, the data from the two surrounding months were also used, and reforecast and precipitation analysis data from all years from 2002 to 2019 were used. Following Scheuerer et al. (2020), we chose to apply a cubic-spline fit to the cumulative hazard function (CHF) of positive precipitation amounts for locations providing a sufficient sample size of positive precipitation values. This CHF provided a more linear relationship between precipitation and (transformed) cumulative probability, making CDF function-fitting more accurate after back transformation from the CHF to the CDF This procedure also ensured that spline-fitted CDF values were bounded to the interval [0, 1]. For a CDF $F_f(x_i^f)$, the CHF was defined as

$$H(x_i^f) = -\log[1 - F_f(x_i^f)]. \tag{5}$$

For the forecast, cubic splines were used to estimate the CHF value given the precipitation amount. The implicit forecast

percentile associated with this precipitation amount was then estimated through a back transformation of Eq. (5). With the analyzed data, the spline function was used to estimate the precipitation amount given the value of cumulative hazard associated with the particular forecast percentile.

To estimate the forecast CHF spline function for a particular tuple of (location, lead time, and month of the year), the climatological time series of reforecast data for a particular location was first sorted. The proportion of values in this sample with zero precipitation was denoted as the fraction zero (FZ). An empirical CDF was generated for the sorted positive samples using the Hazen plotting position (Wilks 2011, Table 3.2). From this, the empirical CHF of positive precipitation amounts was then generated using Eq. (5). If there was a "sufficient" number of positive precipitation samples, a cubic-spline function was then created to estimate CHF given precipitation amount. With reforecast data and its 5 members, sufficient was somewhat arbitrarily defined as more than or equal to 100 positive samples. With less than 100 samples but more than 10 samples, a gamma distribution was fit instead, following HS18. With 10 or fewer samples, no quantile mapping was applied to the raw data for this point. After this training phase, the spline coefficients were stored to disk for later use in quantile mapping the real-time forecast. The final estimated CDF consisted of a point mass FZ at zero precipitation and a CDF of positive values starting at FZ and weighted by (1 − FZ):

$$F^f(x_i^f) = \begin{cases} \text{FZ}, & \text{if } x_i^f = 0 \\ \text{FZ} + (1 - \text{FZ}) \times x_i^f, & \text{if } x_i^f > 0 \end{cases}. \quad (6)$$

For the analysis data, data collection was similar, involving determination of an analyzed fraction zero, the sorting of positive data, and then determination of an empirical CDF and CHF. However, as opposed to the forecast, where the spline function was used to predict cumulative hazard given precipitation amount, the spline function fitting for the analyzed data predicted the precipitation amount given the hazard function. The spline fitting procedure for analyzed data also defaulted use of simpler approaches with few samples. For more than 10 but fewer than 50 analysis samples, gamma distributions were fit. For 10 or fewer samples, no quantile mapping was applied. The fitted spline coefficients were again stored to disk.

More interior knots used in the spline fitting permits a closer function fit to the empirical data, but more knots may be unhelpful and the resulting spline fit may replicate distributional noise on top of a signal when fewer precipitation samples are available. The procedure thus defined a variable number of knots based on the number of samples with positive precipitation. For the forecast, if there were more than 100 nonzero precipitation values (nz), the integer number of interior knots was defined as min (9, nz/30). For the analyzed data, which had 1/5 the number of samples without the 5-member ensemble of forecast values, the number of interior knots was defined as max [min (9, nz/30), 3]. That is, the number of knots was at least 3 and no more than 9, with values in between determined by the sample size. The upper limit on the number of knots constrained possible overfitting to the data.

The knot locations of the cubic spline were chosen to provide more closely spaced knots toward higher precipitation amounts, where large quantile mappings are possible and function precision is desired. Since both cumulative probability and beta distributions (Wilks 2011, section 4.4.4) are defined on the interval [0, 1], given the flexibility of the beta distribution to assume characteristics from U shaped to flat to skewed to near-Gaussian, we chose to define a beta distribution so that, when divided into intervals with equal probability, the interval boundaries would be closer together near 1.0 than near 0.0. Through trial and error, a beta distribution was chosen with parameters $\alpha = 3.5$ and $\beta = 1.0$. The pdf $f(z)$ of a beta distribution is

$$f(z) = \left[\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)}\right] z^{\alpha-1}(1 - z)^{\beta-1}; 0 \le z \le 1; \alpha, \beta > 0. \quad (7)$$

As an example of how this helps objectively set more knots for higher precipitation values, let us presume that the algorithm described above determined that for this sample size, eight interior knots were appropriate. Figure 2 illustrates this beta-distribution pdf and how this pdf was divided into 8 + 1 = 9 separate bins with equal probability. The choice of beta parameters with $\alpha > \beta$ allowed us to specify values of z closer together at high values. For a forecast, these values of z then define the cumulative probability values whose associated precipitation amount were used to define the interior knots.

An example of empirical and fitted distributions of positive precipitation values are shown in Figs. 3 and 4, using July–August–September forecast and analyzed data at Ithaca, New York, for +18–24-h lead forecasts. These data would be used for quantile mapping of August forecasts. For the forecasts in Fig. 3, based on the number of samples with positive precipitation, nine interior knots were chosen following the process illustrated in Fig. 2. The interior knots are denoted by the locations of the black dots in Fig. 3b, which are closer together at the higher percentiles. The empirical CDF was transformed to the empirical CHF shown with the dashed red line in Fig. 3a. A spline was fit to provide coefficients to predict a simpler fitted CHF, which is shown with the thick blue curve using the full training dataset. Finally, the fitted CHF can be inversely transformed back to a CDF; the fitted blue curve in Fig. 3b represents this. The fit to the empirical data is so close that the underlying empirical distribution in red is nearly totally hidden in the plot. Figure 4 shows related data for coincident analyses at this location.

Figure 3 also plots 18 cross-validated replications of the forecast CHF and CDF, leaving one year of data out of the training sample in each replication. The cross-validated curves are difficult to see and generally overlap each other to a great extent, with more detail shown in the inset boxes. While there appears to be some spread in the cross validations of CHF for large precipitation amount in Fig. 3a, after transformation back to CDF in Fig. 3b, these differences are minor. The similarity of CDFs after cross validation illustrates that the simplification to overlap the reforecast training data and the retro prediction data, while undesirable, should have little effect on the results.

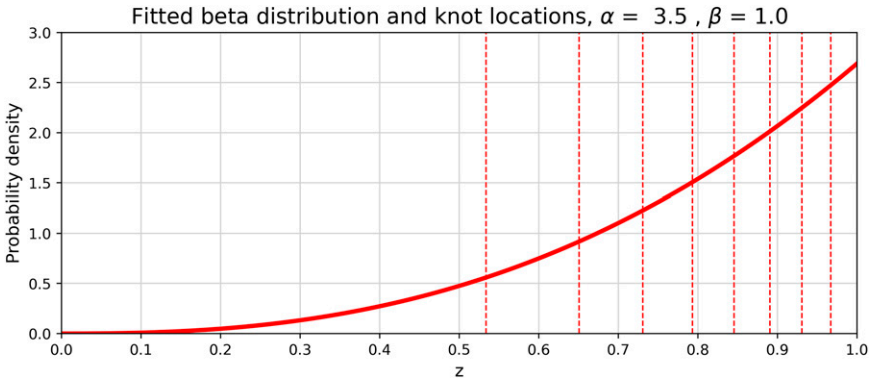

FIG. 2. Illustration of a beta distribution on the interval [0, 1] with parameters $\alpha = 3.5$ and $\beta = 1.0$. Dashed red lines denote the definition of eight values that divide the integrated beta distribution probability density into nine equal partitions. They also define the percentiles of the CDF used as knot locations.

Figure 5 shows the final fitted CDFs of the forecast and analyses when the fractions zero were included, which differ between forecast and analyses. With the inclusion of fraction zeros, which was smaller for the forecast, the forecast was conditionally biased to be wetter than analyzed for light to moderate precipitation amounts. For example, a 5-mm forecast was associated with approximately the 91st percentile of the fitted forecast distribution. This 91st percentile was at approximately 2.5 mm in the analyzed distribution. Hence, the quantile-mapping process adjusted the 5-mm forecast to ~2.5 mm in the analyzed data.

Typically the empirical and fitted CDFs are very flat at high precipitation amounts, and the quantile mappings can be very sensitive to a small mis-estimation of the slope. Accordingly, we adapted the conservative methodology of HS18 to limit possibly unphysical mappings at the highest amounts. Let $q_{0.99}^{f+}$ represent the quantile, the precipitation amount associated with the 99th percentile in the spline-fitted distribution of positive forecast precipitation amounts. We limit the possible quantile mappings of very heavy events by constraining the quantile mapping beyond the 99th percentile to the sum of the quantile mapping at 99th percentile plus the amount the forecast is larger than the 99th percentile:

$$\tilde{x}_i^f = \begin{cases} F^{\alpha-1}[F^f(x_i^f)] & \text{if } x_i^f \le q_{0.99}^{f+} \\ F^{\alpha-1}[F^f(q_{0.99}^{f+})] + x_i^f - q_{0.99}^{f+} & \text{if } x_i^f > q_{0.99}^{f+} \end{cases}. \quad (8)$$

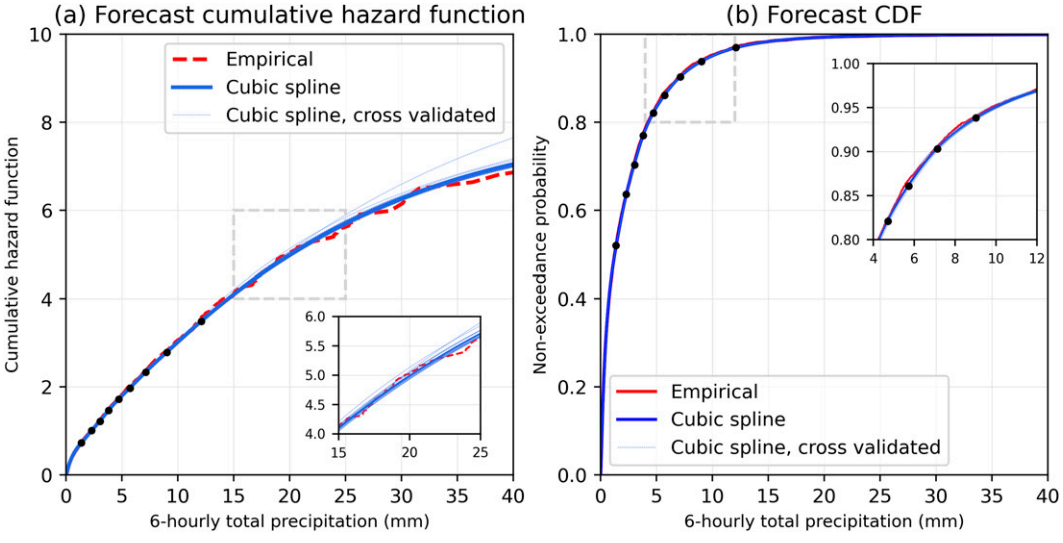

FIG. 3. (a) Forecast cumulative hazard functions, with empirical (red), spline fitted (blue), and 18 different spline fitted, leave-one-year-out cross validated (dashed light blue). Data presented here are July–August–September Ithaca, NY, data for 18–24-h precipitation forecasts with positive amounts. Dots indicate the precipitation amounts of the interior spline knots. (b) Corresponding empirical and fitted forecast CDFs as a function of the 6-hourly positive precipitation amount. Dots again represent spline interior knot locations. Inset boxes for dashed gray region show more detail on the cross-validated spline fits.

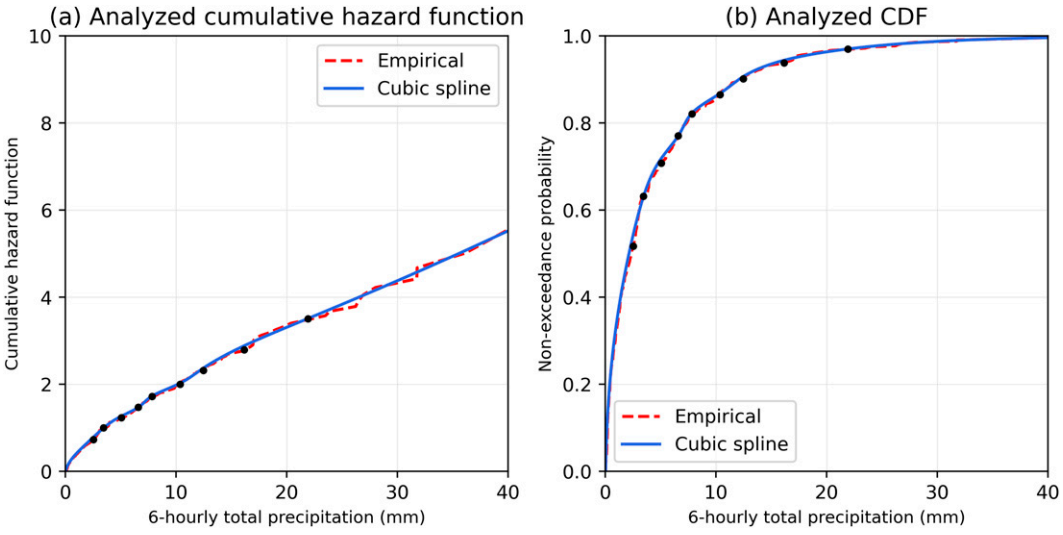

FIG. 4. As in Fig. 3, but for analyzed July–August–September positive precipitation samples of 1800–0000 UTC accumulated precipitation amounts at Ithaca, NY. No cross validations were plotted.

c. *Use of stencil of surrounding points*

Following H17 and HS18, interpolated forecasts at a stencil of 5 × 5 grid points surrounding the location where a quantile mapped forecast is desired are used. This procedure will both enlarge the quantile-mapped ensemble sample size and help to account for the potential of systematic overconfidence of the position of precipitation features in the ensemble prediction system, considering forecast information from nearby points. The potential benefit of using the stencil in regions of complex terrain is explained graphically in Fig. 3 of H17 using a smaller, 3 × 3 stencil.

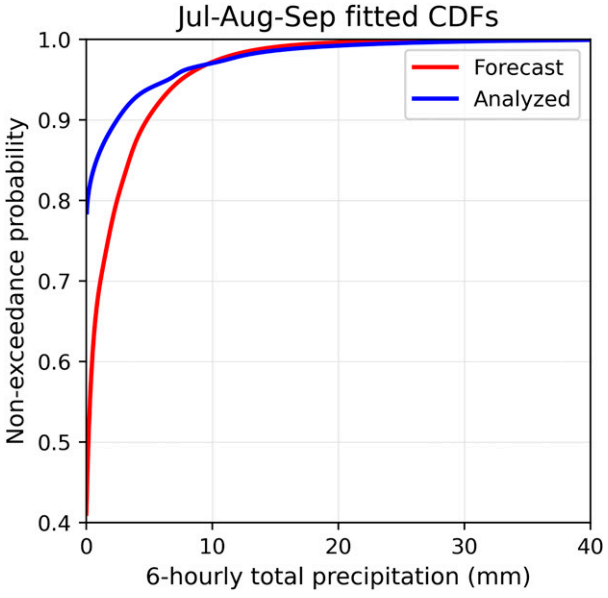

FIG. 5. Fitted CDFs from Figs. 3b and 4b, now including fraction zero, that would implicitly be used in quantile mapping for 0000 UTC initialized 18–24-h forecasts near Ithaca, NY.

Consider the production of +24-h quantile-mapped ensemble member forecasts at the red point in Fig. 6a. For each forecast member and each point in the surrounding stencil, both black and red points, the percentile in the cumulative forecast distribution is determined relative to its local forecast climatology using the previously described spline-fitting procedure. For any given ensemble member's forecast, this then produces 25 estimates of forecast percentiles.

Since we desire a predictive distribution at the red center point (with possibly an analyzed climatology that is distinct from those at the surrounding points), in the quantile-mapping procedure, the 25 forecast percentiles collected across the stencil points become 25 estimates of the analyzed percentile at the central location, the red dot. As implied in Fig. 4, there is a back transformation from percentile to the quantile-mapped precipitation amount using the fitted analyzed distribution *at that central point*. In this way, forecasts at each stencil point are rendered consistent with the analyzed climatology at the red point.

This same basic procedure for quantile mapping using a stencil of surrounding points is applied independently for each ensemble member interpolated to each output analysis grid point in the domain. The result is a 25-fold increased ensemble of quantile mapped forecast values with which to estimate probabilities; hence the summation over $n \times 25$ in Eq. (1).

The spacing between stencil grid points on the ~3-km analysis grid is defined as a function of lead time in hours:

$$s = 12 + \text{int}\left(\frac{L}{18}\right), \tag{9}$$

where int refers to the integer value of the quotient, rounded down, and L is the lead time in hours. Units are in terms of grid points, which are 0.25° apart. An example of how the stencil size increases with lead time is shown in Figs. 6a and 6b. The separation between points is larger for

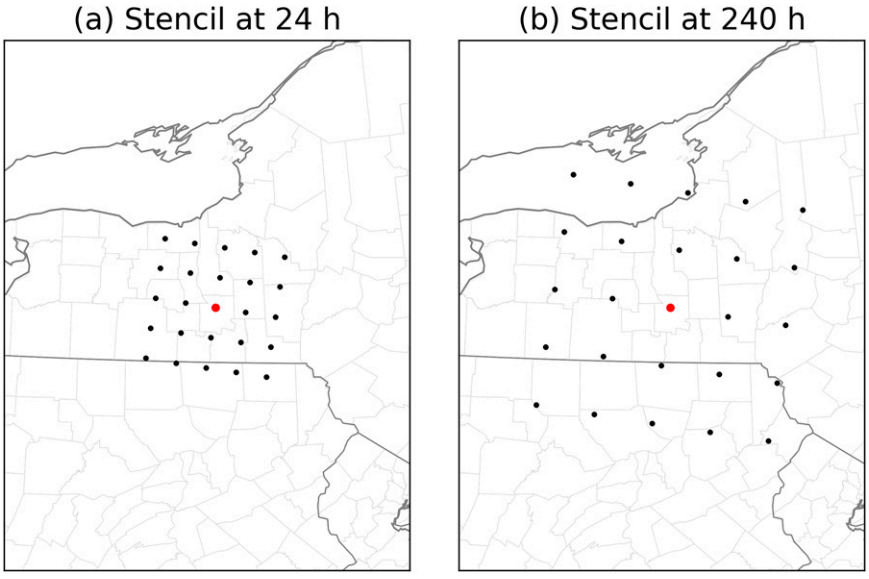

FIG. 6. Illustration of the stencil of surrounding grid points used to enlarge quantile mapping ensemble size, and how this varies as a function of forecast lead time. Quantile-mapped values are sought at the analysis point (red). Percentiles of the forecast are determined at each of the 25 points in the stencil and quantile mapped to the analysis point. Stencil for the (a) 24- and (b) 240-h forecast.

longer lead forecasts, presuming that larger positional errors are more likely at longer lead times. Several experiments were conducted with stencils of various separations; the choice reflected in Eq. (9) represents the approximate optimal from these tests in terms of probabilistic skill and compromises between forecast reliability and resolution (Wilks 2011, section 8.4.3).

d. *Closest-member histograms and estimations of event probability*

As described in HS18, closest-member histogram statistics objectively reweight the sorted, quantile-mapped ensemble of forecasts before determining the probabilities, ameliorating amount-dependent spread deficiencies and remaining biases and thereby improving reliability. Closest-member histograms are very similar in character and interpretation to rank histograms (Hamill 2001); the former populates a histogram based on the rank of the sorted member closest to the observed, the latter determines the rank of the observed relative to a sorted ensemble. Closest-member histograms are more relevant to the weighting and dressing of sorted members. These histograms were generated from the December 2017–November 2019 31 × 25 quantile-mapped members from the ensemble retro run. Statistics were pooled over the CONUS under the assumption that the prior quantile mapping corrected location-dependent biases. Histograms for a given month were populated with data from the calendar month of interest and the surrounding two calendar months. The histograms were generated by determining which member of the sorted ensemble was nearest to the analyzed state. In the case of ties such as zero analyzed precipitation and multiple forecast members with zero, the closest-member rank was randomly assigned among the ties.

With the lengthier reforecast and precipitation reanalysis data, it was possible to stratify the closest-member histograms into a larger number of categories based on the ensemble-mean amount. This was helpful because the histogram characteristics varied significantly with ensemble-mean amount (Fig. 7). Closest-member histograms were populated for categories 1) $0.01\,\text{mm} < \overline{\tilde{x}}^f \le 0.1\,\text{mm}$, 2) $0.01\,\text{mm} < \overline{\tilde{x}}^f \le 0.5\,\text{mm}$, 3) $0.5\,\text{mm} < \overline{\tilde{x}}^f \le 2.0\,\text{mm}$, 4) $2\,\text{mm} < \overline{\tilde{x}}^f \le 6\,\text{mm}$, 5) $6\,\text{mm} < \overline{\tilde{x}}^f \le 15\,\text{mm}$, and 6) $15\,\text{mm} < \overline{\tilde{x}}^f$, where $\overline{\tilde{x}}^f$ denotes the quantile-mapped ensemble mean.

The sample closest-member histograms in Fig. 7 had a U shape indicating a lack of spread and with other shape characteristics that vary with ensemble-mean amount, much akin to rank histograms (Hamill 2001); the U shape is somewhat disguised by the use of a logarithmic y axis. For larger ensemble-mean amounts such as Fig. 7f, there was an oscillation in the histograms at higher sorted ranks, with peaks at fractional rank 28/31, 29/31, 30/31, and 31/31 and relative minima in between. Likely this was caused by an interaction between the use of a stencil of surrounding grid points to enlarge the original ensemble and the characteristic of precipitation forecast features to be larger in scale than the span of the stencil. If, say, 30 of the 31 original members were forecasting a quantile-mapped mean at the center point of >15 mm, but the remaining member forecast was 5 mm at the stencil center point, it is unlikely that any other quantile-mapped member in the 5 × 5 stencil would have a forecast > 15 mm, thereby contributing to closest-member histogram ranks between 30/31 and 31/31. The oscillations are thus a real characteristic of the use of a stencil in the postprocessing algorithm and not

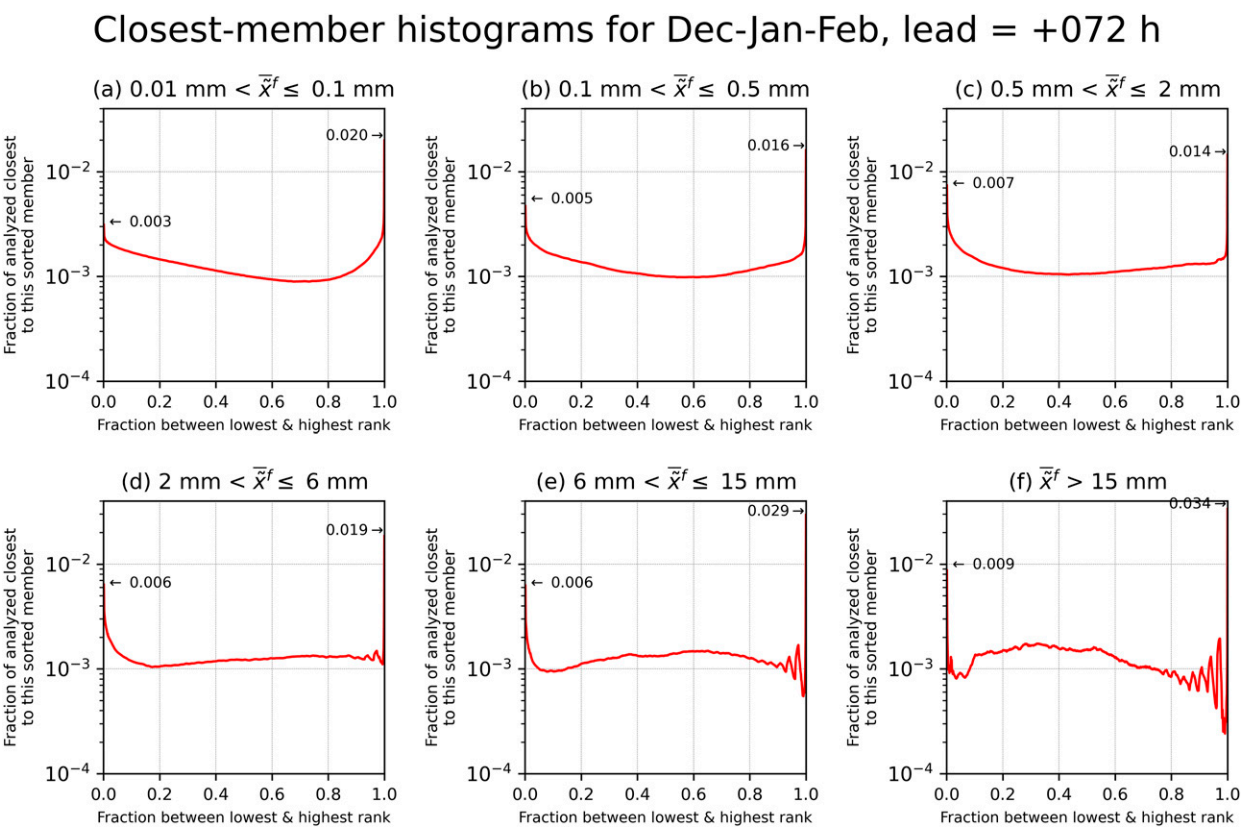

FIG. 7. Sample closest-member histograms for 6-h accumulated precipitation ending at +72-h lead time accumulated over retro forecast samples in December–January–February. Histograms are presented for different categories of quantile-mapped ensemble-mean precipitation amount. Closest-member histograms for ensemble-mean precipitation < 0.01 mm are not shown (and assumed uniform).

a coding bug. The online supplemental material for this article provides closest-member histograms for many more lead times and months of the year for comparison.

While Fig. 7 does not show confidence intervals, these were computed using statistics from a leave-one-day-out cross validation. For each day of the month in question across the two years in the retro period, a cross-validated estimate of the closest-member histogram was generated without data from the day in question. The 5th and 95th percentiles of the cross validated were very small for this lead time, so small they would not show distinctly on the plot, and hence they were omitted. They are shown in plots in the online appendix, where at long leads the confidence intervals are larger. In general, the small confidence intervals suggest that a formal separation of training and validation data would have little practical effect on the result.

For many lead times past the first several days of the forecast, there were not always stable closest-member histogram statistics for ensemble means above 15 mm. This was a limitation of the algorithm, though the verification results presented in Part II show that the technique still improves probability forecasts of exceeding 25 mm. We have also explored simpler alternative fits for these high precipitation amounts using beta distributions and have informally found improved precipitation forecasts of extreme amounts using the closest-member histograms derived from fitted beta distributions. However, we determined it was beyond the reasonable scope of this article to present these results, so they are saved for possible future research.

e. *Objective dressing statistics*

Equation (1) showed that the postprocessing algorithm described here provides the final probabilistic forecast based on a sum of weighted probability distributions that are a function of the quantile-mapped forecast amounts. This is known as a "dressed" ensemble. Roulston and Smith (2003; see their appendix) provide a theoretical justification for including a dressing of exchangeable ensemble members as part of a postprocessing algorithm. Dressing statistics are ideally developed from objective statistics comparing the closest member to the analyzed value.

The proposed algorithm here expands upon the dressing concept by objectively developing dressing statistics for sorted, quantile-mapped members, with separately estimated dressing statistics for the lowest, intermediate, and highest sorted quantile-mapped members. Intermediate members are all those except the lowest and highest sorted members. The assumption underlying this step is that location-dependent amount biases have already been ameliorated with quantile

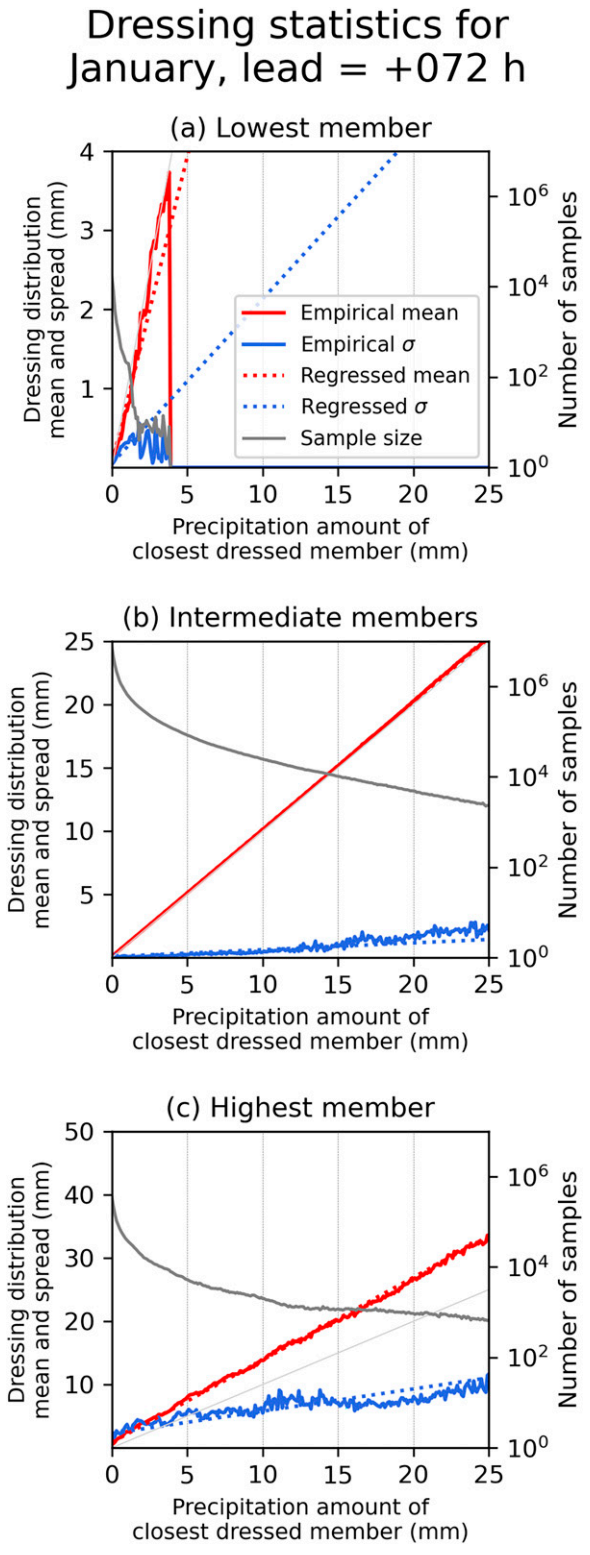

FIG. 8. Empirical and fitted objective dressing statistics for January forecasts and the +72-h lead time using quantile-mapped retro forecast data from December–January–February, classified by (a) lowest sorted member, (b) intermediate-ranked member, and (c) highest sorted member. Solid lines red (dressing mean) and blue (standard deviation) are empirical dressing statistics as a function of the best-member precipitation amount. Dashed lines are the fitted weighted least squares regression relationship to the empirical data. The diagonal light-gray line is the 1-to-1 line for comparison. The heavier, darker gray line is the sample size per bin, with bins every 0.2 mm.

mapping, and the dressing statistics can be calculated based on data pooled across the CONUS.

Dressing statistics for a given lead time were calculated by comparing a set of sorted, quantile-mapped members to the contemporaneous analyzed values across the CONUS grid. Ensemble forecasts from the retro forecast period were used. Using pooled domain statistics gathered by comparing the analyzed data to quantile-mapped, 25-fold enlarged retro forecasts at each forecast lead time, we determined an empirical mean and standard deviation of the dressing statistics for lowest, intermediate, and highest-ranked ensemble members, binned into categories based on the closest-member precipitation amount.

Suppose the lowest sorted quantile-mapped member was closest to the analysis for that point; the precipitation amount was then noted. Over many samples where the lowest sorted member was the closest, statistics were thus gathered for estimating the dressing mean and standard deviation as a function of difference between the quantile-mapped precipitation amount. Similarly, statistics were populated for samples where the intermediate-ranked sorted members were the closest and where the highest sorted member was closest. Figure 8 illustrates these dressing statistics for January and the +72-h lead time. These were actually populated with December 2017, January–February 2018, December 2018, and January–February 2019 31-member (×25) retro forecast data.

Of particular interest are the dressing statistics of the lowest and highest sorted ensemble members. It is possible that the mean of the distribution of analyzed points when the lowest member was closest to the analysis would not be centered on the analyzed value but instead be somewhat lower. Similarly, it is possible that the distribution of analyzed points when the largest member was closest was higher than even that largest member.

Consider first the dressing statistics for the subset of samples where the lowest sorted forecast member was determined to be the closest to the analyzed value. The mean of the analyzed values given lowest member was closest was indeed on average smaller than the value of that lowest sorted quantile-mapped ensemble member itself. As the raw data were noisy, we then fitted a weighted-least squares regression to the data to predict a simpler linear model for dressed mean and standard deviation as a function of closest-member precipitation amount. The weighted-least squares regression line for the mean shows that the regressed mean for the dressing distribution of the lowest sorted member is roughly 75% of the magnitude of the quantile mapped value. For example, when the quantile mapped value of the lowest member is 4 mm, the

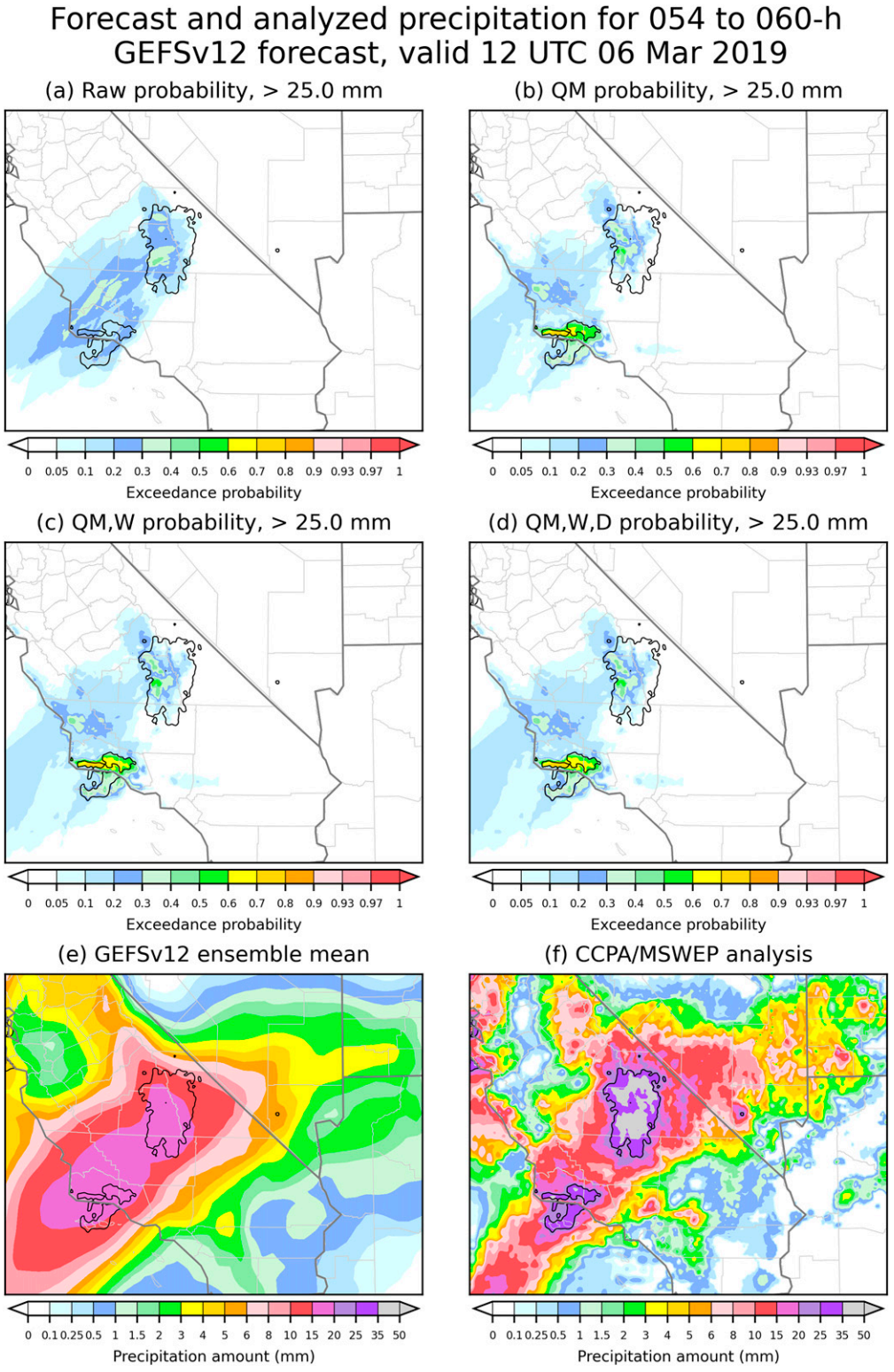

FIG. 9. Retro GEFSv12 precipitation forecast and analyzed data. The GEFSv12 forecasts were initialized at 0000 UTC 4 Mar 2019, and +54- to +60-h forecasts are shown, aligning with the analyzed precipitation during the verification period, 0600–1200 UTC 6 Mar 2019. (a) Raw ensemble probability of exceeding 25 mm 6 h^{-1}. The black contour is the analyzed 25-mm contour, reproduced from (f). (b) As in (a), but for quantile-mapped forecasts. (c) As in (b), but with closest-member histogram weighting included. (d) As in (c), but with dressing included. (e) Raw ensemble-mean precipitation analysis. (f) Combined CCPA/MSWEP precipitation analysis.

mean of the dressed distribution is close to 3 mm. Similarly, in Fig. 8c where the largest sorted member was determined to be the best member, the mean of the dressing distribution was larger than that largest member. For example, suppose the highest sorted member of the ensemble was 15 mm. The fitted dressing relationship for this member indicates the mean of the dressed distribution should be closer to 20 mm.

Taken together, the dressing of these two outlying sorted members together with closest-member histogram weighting slightly increased the spread of the predictive distribution, ameliorating ensemble overconfidence, as will be shown in Part II.

When the closest member was an intermediate-ranked member (Fig. 8b), which is any member other than largest or smallest, the objective dressing statistics predicted a mean nearly centered on the quantile-mapped amount, perhaps 1% larger. In these cases, this best member was likely closely bracketed by other forecast members with nearly similar amounts, and hence the predicted mean of the dressing distribution for that selected best member is very close to the quantile-mapped amount itself, and the dressing distribution has a small standard deviation. If one examines Fig. 8b closely, the mean of the dressed distribution is slightly larger than predicted. While not confirmed, we strongly suspect that this is because of the right-skewed nature of precipitation and even precipitation dressing statistics. It might be slightly preferable to fit, say, a gamma distribution rather than Gaussian. This was not attempted here.

4. A sample forecast

Figure 9 provides an example forecast and the impact of calibration for an atmospheric river landfall in southern California in March 2019. The multipanel figure shows raw, quantile mapped (QM), quantile-mapped and closest-histogram weighted (QM, *W*), and quantile-mapped, weighted, and dressed (QM, *W*, *D*) forecast probabilities for 25 mm 6 h^{-1}. While the forecasts were not improved in the maxima over the southern Sierra Nevada range, the coastal maximum east of Santa Barbara, California, is enhanced up through the quantile mapping; apparently local accentuation of precipitation is common along the coast. The probabilities were also slightly increased with closest-member histogram weighting along the coast, but little change is visible as a result of the last step, dressing.

5. Discussion and conclusions

This article describes adaptations to previously described NBM precipitation postprocessing methods that are intended to transfer to operational use in the National Weather Service under its National Blend of Models program. The intent is to provide GEFSv12 probability forecasts with more appropriate spatial detail, reduced error, and improved reliability, forecasts that improve enough to lessen the necessity of forecasters to make manual modifications to the objectively generated guidance. The second part of this series of articles, Part II, provides results that will show that these objectives were mostly accomplished, providing comprehensive verification and more case studies.

Acknowledgments. This research was supported under a NOAA WPO Precipitation Grand Challenge Grant NA17OAR4320101. Nachiketa Acharya of NOAA and CIRES is thanked for his informal review, as well as two anonymous reviewers. IBM/Weather Company is thanked for allowing us to finish the article after the lead author changed affiliation.

Data availability statement. These data and algorithms are available for others to use at ftp://ftp.cdc.noaa.gov/Projects/GEFSv12/NBM/ and https://github.com/ThomasMoreHamill/GEFSv12_NBM.

REFERENCES

Beck, H. E., E. F. Wood, M. Pan, C. K. Fisher, D. G. Miralles, A. I. J. M. van Dijk, T. R. McVicar, and R. F. Adler, 2019: MSWEP V2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bull. Amer. Meteor. Soc.*, **100**, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1.

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

Craven, J. P., D. E. Rudack, and P. E. Shafer, 2020: National Blend of Models: A statistically post-processed multi-model ensemble. *J. Oper. Meteor.*, **8**, 1–14, https://doi.org/10.15191/nwajom.2020.0801.

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

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hamill, T. M., 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.

——, and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079–4098, https://doi.org/10.1175/MWR-D-18-0147.1.

——, J. S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

——, ——, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46, https://doi.org/10.1175/BAMS-87-1-33.

——, 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, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, 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.

——, and Coauthors, 2022: The reanalysis for the Global Ensemble Forecast System, version 12. *Mon. Wea. Rev.*, **150**, 59–79, https://doi.org/10.1175/MWR-D-21-0023.1.

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, https://doi.org/10.1175/2009JHM1006.1.

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.

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

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

Scheuerer, M., and T. M. Hamill, 2015: Statistical post-processing 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.

——, M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Schumacher, R. S., A. J. Hill, M. Klein, J. Nelson, M. Erickson, S. M. Trojniak, and G. R. Herman, 2021: From random forests to flood forecasts: A research to operations success story. *Bull. Amer. Meteor. Soc.*, **102**, E1742–E1755, https://doi.org/10.1175/BAMS-D-20-0186.1.

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, https://doi.org/10.1175/MWR3441.1.

Stovern, D. R., T. M. Hamill, and L. L. Smith, 2023: Improving National Blend of Models probabilistic precipitation forecasts using long time series of reforecasts and precipitation reanalyses. Part II: Results. *Mon. Wea. Rev.*, **151**, 1535–1550, https://doi.org/10.1175/MWR-D-22-0310.1.

Switanek, M. B, T. M. Hamill, L. N. Long, and M. Scheuerer, 2023: Predicting subseasonal tropical cyclone activity using NOAA and ECMWF reforecasts. *Wea. Forecasting*, **38**, 357–370, https://doi.org/10.1175/WAF-D-22-0124.1.

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

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

Worsnop, R. P., M. Scheuerer, F. Di Giuseppe, C. Barnard, T. M. Hamill, and C. Vitolo, 2021: Probabilistic fire-danger forecasting: A framework for week-two forecasts using statistical post-processing techniques and the Global ECMWF Fire Forecast System (GEFF). *Wea. Forecasting*, **36**, 2113–2125, https://doi.org/10.1175/WAF-D-21-0075.1.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System, version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.