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

DIANA R. STOVERN,[a,b] THOMAS M. HAMILL,[b,c] 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 1 March 2023, accepted 21 March 2023)

ABSTRACT: This second part of the series presents results from verifying a precipitation forecast calibration method discussed in the first part, based on quantile mapping (QM), weighting of sorted members, and dressing of the ensemble. NOAA's Global Ensemble Forecast System, version 12 (GEFSv12), reforecasts were used in this study. The method was validated with preoperational GEFSv12 forecasts from December 2017 to November 2019. The method is proposed as an enhancement for GEFSv12 precipitation postprocessing in NOAA's National Blend of Models. The first part described adaptations to the methodology to leverage the ~20-yr GEFSv12 reforecast data. As shown here in this part, when compared with probabilistic quantitative precipitation forecasts from the raw ensemble, the adapted method produced downscaled, high-resolution forecasts that were significantly more reliable and skillful than raw ensemble-derived probabilities, especially at shorter lead times (i.e., <5 days) and for forecasts of events from light precipitation to >10 mm (6 h)$^{-1}$. Cool-season events in the western United States were especially improved when the QM algorithm was applied, providing a statistical downscaling with realistic smaller-scale detail related to terrain features. The method provided less value added for forecasts of longer lead times and for the heaviest precipitation.

KEYWORDS: Downscaling; Statistical techniques; Forecast verification/skill; Probabilistic Quantitative Precipitation Forecasting (PQPF); Ensembles; Postprocessing

1. Introduction

Skillful and reliable probabilistic quantitative precipitation forecasts (PQPF) are necessary for a variety of applications. Forecasters at the National Oceanic Atmospheric Administration (NOAA) use PQPFs to provide impact-based decision support services to water resource managers and emergency personnel, especially for characterizing the uncertainty leading up to a possible extreme-precipitation event (Dahl and Xue 2016). The ensemble precipitation data used to generate PQPFs are used as forcing for hydrologic models, which also needs to be skillful to reduce uncertainty and improve the accuracy of streamflow forecasts (Brown et al. 2012). NOAA's Hydrological Ensemble Forecasting System (HEFS; Demargne et al. 2014) and National Blend of Models (NBM; Hamill et al. 2017; Hamill and Scheuerer 2018; Craven et al. 2020) are operational applications that currently incorporate ensemble data into their production for producing short- and medium-term (i.e., lead times from <1 to 15 days) hydrologic and atmospheric forecasts. One of the primary ensemble systems used to generate the forecasts comes from the U.S. National Weather Service's Global Ensemble Forecasting System (GEFS).

The GEFS has gone under many improvements since it became operational in 1992 (Toth and Kalnay 1993). One of the major improvements with the most recent update to version 12 (GEFSv12) in September 2020 is the replacement of the dynamical core from the legacy Global Spectral Model to the Finite Volume Cubed-Sphere Dynamical Core (Zhou et al. 2022). GEFSv12 was also used to generate a 20-yr global reanalysis dataset from 2000 to 2019 and a 30-yr reforecast dataset from 1989 to 2019, which are described in detail in Hamill et al. (2022) and Guan et al. (2022). The ensemble system was also increased from 21 members to 31, and the horizontal grid spacing is now ~26 km. A detailed description of the upgrades and verification in comparison with the GEFS, version 11 (GEFSv11), is described in Zhou et al. (2022). The verification performed with regard to 24-h accumulated precipitation forecasts had shown that, with these improvements, probabilistic forecasts of 1, 5, 10, and 20 mm were both more reliable and skillful in the GEFSv12 than in the GEFSv11.

Despite the improvements in the GEFSv12 precipitation forecasts, there was still a tendency for the model to provide unreliable forecast probabilities over the contiguous United States (CONUS; Zhou et al. 2022 and results herein). Even with the reduction in grid spacing, the model is still too coarse to resolve fine-scale topographically forced precipitation variability, as has been the case in legacy versions of the GEFS (Lewis et al. 2017). Further, global models like the GEFSv12 that parameterize the effects of deep convection also tend to struggle with heavy precipitation events (Herman and Schumacher 2016). Although not explicitly in reference to the GEFSv12, other precipitation biases may exist that are dependent on model type, lead time, location, and time of day

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-22-0310.s1.

Corresponding author: Diana R. Stovern, diana.stovern@noaa.gov

DOI: 10.1175/MWR-D-22-0310.1

(Yuan et al. 2005; Hamill 2012; Zhu and Luo 2015). These biases can, in turn, produce biased ensemble streamflow predictions when the raw ensemble forecasts are directly used as precipitation forcing in hydrologic prediction systems. Thus, it has become commonplace to postprocess the raw model forecasts to correct for systematic errors and make them more suitable for operational use and chained applications like hydrologic prediction (Hamill and Whitaker 2006; Hamill et al. 2008, 2013, 2015; Vannitsem et al. 2018, 2021).

In Hamill et al. (2017, hereinafter referred to as H17), quantile mapping (QM) was applied to ensemble model guidance from the National Weather Service and Environment Canada, with the intent to demonstrate algorithmic concepts that could be used in the NBM. QM provides an amount-dependent bias correction based on the differences between forecast and analyzed cumulative precipitation distribution functions (Voisin et al. 2010; Hopson and Webster 2010; Maraun 2013). The training dataset in H17 used the previous 60 days of coincident forecasts and analyses along with "supplemental locations" to populate the cumulative distribution functions (CDFs). Supplemental locations were defined as nearby grid points that had similar precipitation climatology and terrain features. H17 included applying the quantile mapping using forecast data from a 3×3 stencil surrounding grid points to synthetically enlarge the ensemble, decreasing sampling variability and ameliorating overconfidence in the placement of precipitation features. A simplified dressing of the ensemble was performed by adding amount-dependent random noise to each member to increase the spread of the ensemble, and Savitzky–Golay smoothing (Savitzky and Golay 1964) was also applied to the POP field.

Additional improvements to the QM technique described in H17 were presented in Hamill and Scheuerer (2018, hereinafter HS18). One of revisions of the procedure included estimates of the forecast and analyzed CDFs with a "fraction zero" (the fraction of samples with zero precipitation) and a Gamma distribution for positive amounts. Adjustments to the quantile mapping were also made when forecasts were exceptionally wet; see section 3b(2) of HS18 for more details. Additional changes included the weighting of sorted members based on "closest-member histogram" statistics and dressing of sorted members with Gaussian-distributed kernels of probability density. With the adjustments made in HS18, additional skill and forecast reliability was added beyond the results presented in H17, particularly at heavier precipitation thresholds.

Although the technique described in H17 and HS18 improved reliability for 12-hourly probability of precipitation (POP) and 6-hourly deterministic QPF relative to raw forecast guidance, the use of a short training dataset (the past 60 days) in conjunction with the use of supplemental locations reduced the amount of terrain-related precipitation detail in the intermountain western United States and provided unrealistic quantile mappings during transitions between seasons with predominantly stratiform and predominantly convective precipitation. Given that the GEFSv12 implementation was accompanied with a 20-yr reforecast dataset, Hamill et al. (2023, hereinafter Part I) of this series described a revised quantile-mapping approach, weighting, and dressing approach to precipitation calibration that leveraged those reforecasts to potentially improve precipitation calibration. The approach as described in Part I abandons the use of supplemental locations, uses multidecadal reforecast data to populate the CDFs, and applies more careful estimations of the associated CDFs using spline-fitting procedures and improved weighting and dressing. The research hypothesis from Part I states 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 (Hamill et al. 2022); with more carefully defined CDFs based only on the data from the grid point in question, the new method should be able to define amount-dependent biases with terrain-related detail, presuming the high-resolution analysis has subgrid-scale detail relative to the coarser-resolution forecast. The more careful estimation of closest-member histograms and dressing statistics may also provide an improvement over the previous-generation algorithm. See Part I) for algorithmic details.

In this part, the characteristics of the GEFSv12 precipitation forecasts with the revised reforecast-based quantile mapping, weighting, and dressing procedure will be evaluated relative to the raw ensemble guidance. The 6-hourly precipitation totals are used, and training and validation data consists of high-resolution precipitation analyses synthesized from the Climatology Calibrated Precipitation Analysis (CCPA; Hou et al. 2014) over the contiguous United States and the Multi-Source Weighted Ensemble Precipitation (MSWEP; Beck et al. 2019) elsewhere. The relative impact of each step of the calibration process will be evaluated, starting with the results from just quantile mapping, then quantile mapping with closest-member histogram weighting, and finally quantile mapping with the weighting and dressing. The verification metrics used in this study are introduced in section 2. The results comparing the performance of the procedure using "retro" forecasts from December 2017 to November 2019 are discussed in section 3. Section 4 concludes with a summary and discussion, including future work.

2. Data and evaluation methodology

The raw and postprocessed GEFSv12 precipitation forecasts were evaluated using the 0000 UTC cycle from the GEFSv12 preproduction retrospective ("retro") forecasts for all days between 1 December 2017 and 30 November 2019. These were more completely described in Part I. The high-resolution (~3 km; 6-hourly) precipitation analyses used for verification were also described in that article. The reforecast ensemble dataset was additionally enlarged using a 5×5 stencil of forecast values from the surrounding grid points. As described in H17 and HS18, the purpose of this was to reduce sampling variability at each grid point, deal with systematic position errors, and create smoother spatial maps of ensemble probabilities. An example in H17 (H17, their Figs. 7a,b) shows the benefit of quantile mapping using the stencil versus without the stencil. Part I contains more information on how the 5×5 stencil was applied in this experiment.

Because of the volume of data that was generated with the quantile-mapping routine, when generating objective statistics of the results, we chose to evaluate data on a thinned grid, every 10th gridpoint of the original ~3-km National Digital Forecast Database (NDFD) output grid. Verification was performed on the set of thinned grids within the CONUS and selected points in the Columbia River basin in Canada and tributaries of the Rio Grande in Mexico. However, for case studies, the full fields including surrounding regions were shown when demonstrating differences between raw and postprocessed guidance.

Forecast lead times from +6 to +240 h in 6-hourly accumulation periods were evaluated, but particular attention was paid to the +24-h (+1 day), +72-h (+3 day), +120-h (+5 day), and +240-h (+10 day) lead times. PQPF for 6-hourly amounts exceeding 0.254 mm (POPs) and 1, 5, 10, and 25 mm were analyzed, but the results for the POPs, $P[\text{obs} > 5 \text{ mm } (6 \text{ h})^{-1}]$, $P[\text{obs} > 10 \text{ mm } (6 \text{ h})^{-1}]$, and $P[\text{obs} > 25 \text{ mm } (6 \text{ h})^{-1}]$ thresholds will be the focus. When characterizing these thresholds, POPs will sometimes be referred to as light precipitation, $P[\text{obs} > 5 \text{ mm } (6 \text{ h})^{-1}]$ and $P(\text{obs} > 10 \text{ mm } (6 \text{ h})^{-1}]$ are considered moderate precipitation, and $P[\text{obs} > 25 \text{ mm } (6 \text{ h})^{-1}]$ is heavy precipitation. Results for the "warm season" (April–September) and "cool season" (October–March) will be provided. Results tabulated for the western United States were calculated using all points inside the CONUS that were in the western 3/7ths of the domain, or approximately west of 103°W longitude. Similarly, the region considered the eastern United States covers all grid points including and east of roughly 103°W longitude.

Similar evaluation metrics as those used in H17 and HS18 will be applied here. Brier skill scores (BSS) and reliability diagrams (Wilks 2011) will be the primary methods for evaluating the probabilistic forecasts. The BSS were calculated relative to the climatology from the CCPA/MSWEP gridded data. The climatology was determined separately for each grid point, for each 6-hourly accumulation period during the day, and for each month of the year. Relative to this climatology, a perfect probabilistic forecast will have a BSS value of 1.0, whereas a forecast with a BSS value of 0.0 indicates that the forecast has the same skill as the climatology.

Reliability diagrams were also used to assess the consistency between the forecast probabilities and the observed frequency (Wilks 2011; H17). Each reliability diagram contains an inset histogram that displays the frequency for which forecasts of various probabilities were issued. A reliable forecast will display a 1–1 relationship between the issued forecast probability and the observed relative frequency. A sharp forecast will have probabilities that deviate from climatology, with more forecast probabilities closer to 0 or 1; a desired skillful forecast exhibits both sharpness and reliability.

Confidence intervals for the 5th and 95th percentiles of the resampled distribution are displayed on the reliability diagrams to indicate the uncertainty in the reliability calculations due to sample size. For reliability, the bootstrap procedure generated 100 resamplings of the underlying contingency tables needed to produce the reliability, with replacement. These resampled overall contingency tables were computed from a sum of the random daily contingency tables, the dates of which were selected randomly, with replacement. Those daily contingency tables were populated with data from grid points across the domain. From the 100 contingency tables, 100 reliability curves were estimated, and the 5th and 95th percentiles of the analyzed relative frequency were reported. When fewer than 100 samples were available for a given probability bin in the contingency table, no confidence intervals were plotted.

Several case studies will be shown that qualitatively demonstrate how the postprocessing alters the spatial structure of probabilistic rainfall for different precipitation regimes across the CONUS. The examples show the strengths and weaknesses of the algorithm when the QM is applied, providing insight to how the postprocessed guidance can be used by forecasters in regard to heavy-rainfall prediction.

3. Results

a. *Reliability*

1) POPs

The raw GEFSv12 showed a lack of reliability for POPs at all lead times out to +240 h for both cool- and warm-season events. Shown in Fig. 1 are the reliability diagrams for the +24-, +120-, and +240-h lead times. Raw forecasts were generally overconfident, with forecast probabilities above ~20% being too high (precipitation observed less frequently than predicted) and probabilities below ~20% being correspondingly too low. The overall unreliability for both seasons may have several causes, including insufficient spread in the ensemble from the choice of initial conditions and suboptimal physical parameterizations and "stochastifications" thereof, as well as the different grids of the raw GEFSv12 (~25 km) versus the verification (3 km). There are inevitable errors in precipitation analyses as well (Gehne et al. 2016), and the forecasts have not been dressed with possible precipitation errors (Hamill 2001, Fig. 6) as may be a desirable practice; the magnitude of precipitation errors as a function of precipitation amount and season have not been estimated, to our knowledge. When comparing the evaluation of raw GEFSv12 reliability here relative to previous evaluations against coarser-resolution analyses as in HS18, however, the GEFSv12 appeared to be more reliable than the previous model version. This suggests that there was less error to be ameliorated through statistical postprocessing with the newer GEFSv12 version in comparison with older versions, and the skill improvement from postprocessing may be lessened.

Applying quantile mapping substantially improved the reliability for all lead times and seasons, correcting the tendency for the raw GEFS to overestimate mid- to higher-end probabilities. The tendency for the raw GEFS to underestimate lower-end probabilities was further exacerbated with the quantile mapping; however, this was remediated when the weighting and dressing were applied. The largest benefit of the dressing and weighting is to add further refinement to the quantile mapping, particularly at shorter lead times since not as much improvement is to be had at later lead times after the QM is applied (Figs. 1b,c,e,f)

A regional analysis indicated that the raw GEFSv12 had more of a tendency in the eastern United States to

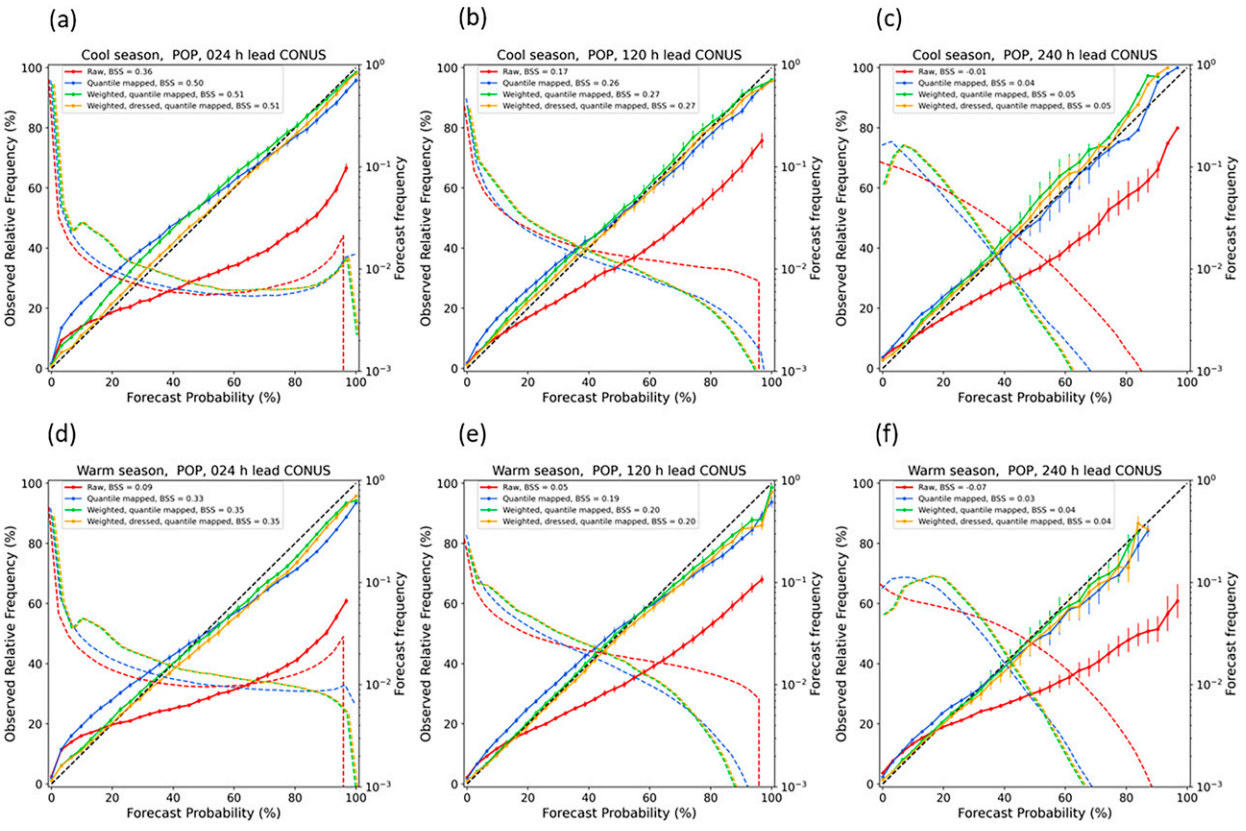

FIG. 1. Reliability diagrams (left axis label) and logarithmic observed relative frequency (right axis label) for probability of exceeding nonzero precipitation (POP) in 6-h accumulation periods for the (a)–(c) cool-season and (d)–(f) warm-season results ending at the (left) +24-, (center) +120-, and (right) +240-h lead times. BSSs are noted in the legend. Error bars represent the 5th and 95th percentiles from a 100-sample bootstrap distribution generated by sampling case days with replacement. Each panel shows the raw GEFS forecast (red curve), quantile-mapped forecast (blue curve), quantile-mapped forecast after adding the closest-member histogram weighting (green curve), and quantile-mapped forecasts after adding the weighting and dressing (orange curve).

overestimate mid- to high-values POPs in comparison with the western United States, especially at early lead times. However, the bias toward underestimating lower-end POPs was coming more from the western United States, especially in the warm season (see Fig. S1 in the online supplemental material). The quantile mapping, dressing and weighting corrected for the regional biases in both the warm and cool seasons.

2) $P[\text{OBS} > 10\ \text{MM}\ (6\ \text{H})^{-1}]$

Like the POPS, the raw GEFSv12 over-forecasted probabilities of 6-h precipitation greater than 10 mm across the CONUS for both warm- and cool-season precipitation events (Fig. 2). Quantile mapping, dressing, and weighting corrected the overconfidence in the model out to a +72-h lead time, more so in the cool season (Figs. 2a–c) versus the warm season (Figs. 2d–f). Interestingly, quantile mapping by itself in the cool season during the first +48 h produced such reliable forecasts that there was not much room for improvement from the dressing and weighting. Forecast probabilities in the warm season were still slightly overestimated after the post-processing steps were applied.

A regional analysis for this threshold (see Fig. S2 in the online supplemental material) showed that the unreliability in both the warm and cool seasons were coming more from the western United States than the eastern United States. For the warm season at the +48-h lead time specifically, the climatologically drier western United States had fewer 6-h rainfall totals greater than 10 mm (Fig. S2c). When heavier rain did occur, it was likely more primarily driven by smaller-scale convective processes that intermediate-resolution models like the GEFSv12 tended to be unable to resolve and correctly propagate (Herman and Schumacher 2016). Quantile mapping, weighting, and dressing still improved reliability over the raw GEFSv12 in the west, thus correcting gross biases during warm-season events but likely could not address the limitations of model resolution and the use of parameterized deep convection. The reliability was much better during cool-season events (Fig. S2a) in the western United States but was greatest overall for cool-season events in the eastern United States (Fig. S2b).

3) $P[\text{OBS} > 25\ \text{MM}\ (6\ \text{H})^{-1}]$

For heavy precipitation thresholds {i.e., P[obs > 25 mm $(6\ \text{h})^{-1}$]}, reliability from the raw GEFSv12 was better in the

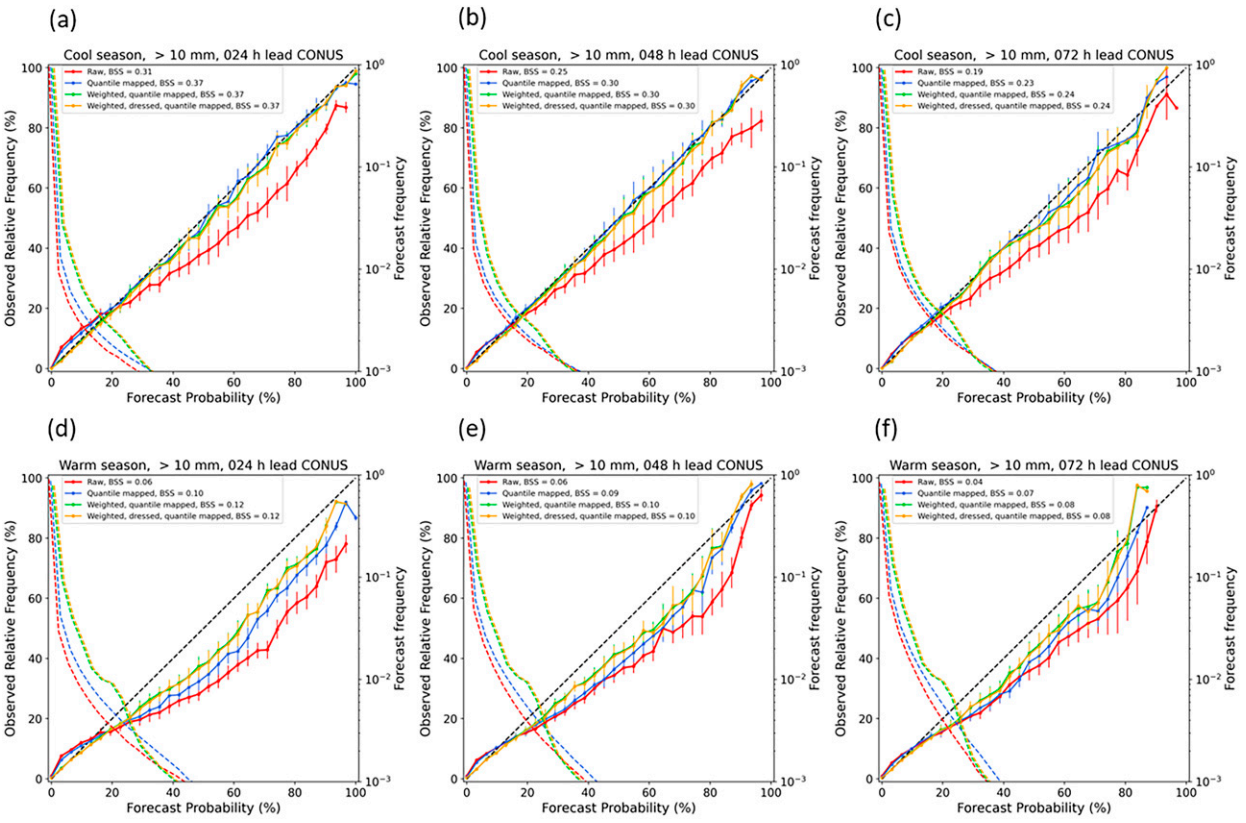

FIG. 2. As in Fig. 1, but for the probability of 6-h rainfall exceeding 10 mm for (a)–(c) cool-season and (d)–(f) warm-season events at lead times of (left) +24, (center) +48, and (right) +72 h.

cool season than in the warm season. During the convective warm season, the analyses are not very accurate as there will be substantial subgrid variability and hence representativity errors affecting both the forecasts and the verification. Reliability did not improve for either warm or cool-season events when the quantile mapping, dressing, and weighting were applied (Fig. S3 in the online supplemental material). This was the case for all lead times, regions, and seasons that showed a systematic bias toward overestimating forecast probabilities. This may have been due to a limitation on how finely the closest-member histograms were binned for heavier/extreme amounts. In results (not shown), a simplified representation of closest-member histograms was generated by fitting beta distributions to the sorted closest-member ranks, using the mean and variance of these ranks and the method of moments estimator (Wilks 2011, his section 4.4.4); these results showed that the shape of the closest-member histogram continues to change as mean precipitation amount increases. This indicates that the forecasts may be unreliable at >25 mm $(6\text{ h})^{-1}$ due in large part to suboptimal weighting.

b. *Brier skill scores*

Figures 3–5 show the Brier skill score of the GEFSv12 6-h accumulated probabilistic precipitation forecasts both before and after each step of the postprocessing algorithm was applied, separated by region (i.e., eastern United States and western United States), and season (i.e., warm and cool). Results were plotted every 12 h for lead times out to +240 h. Starting with POPs, the quantile mapping added substantial skill over the raw GEFSv12 for all seasons and regions (Fig. 3). The closest-member histogram weighting and dressing added only slight skill to the quantile mapping, with the largest benefit for cool-season events in the eastern United States (Fig. 3b). This suggests that the primary benefit from the additional steps of weighting and dressing was to improve reliability, but with a somewhat corresponding decrease in resolution (see Brier score decomposition in Wilks 2011, section 8.4.3). Note that the postprocessed GEFSv12 had higher forecast skill during the cool season than in the warm season, especially in the eastern United States prior to the +120-h lead time. This may be because winter precipitation tended to be larger in scale and the model was better able to resolve it. However, postprocessing added the greatest improvement to the raw GEFSv12 during the eastern U.S. warm season (Fig. 3d). For example, at the +24-h lead time, the raw BSS was ~0.1 and increased to ~0.37 after the weighting and dressing was applied. In comparison, for the eastern U.S. cool season, the skill at +24 h increased from 0.4 to 0.54. With increasing lead time, the amount of improvement in BSS that the quantile mapping added to the raw GEFSv12 generally decreased for both seasons and regions; likely this was because the spread deficiency and bias are particularly large at the shortest lead times, affected by model spinup.

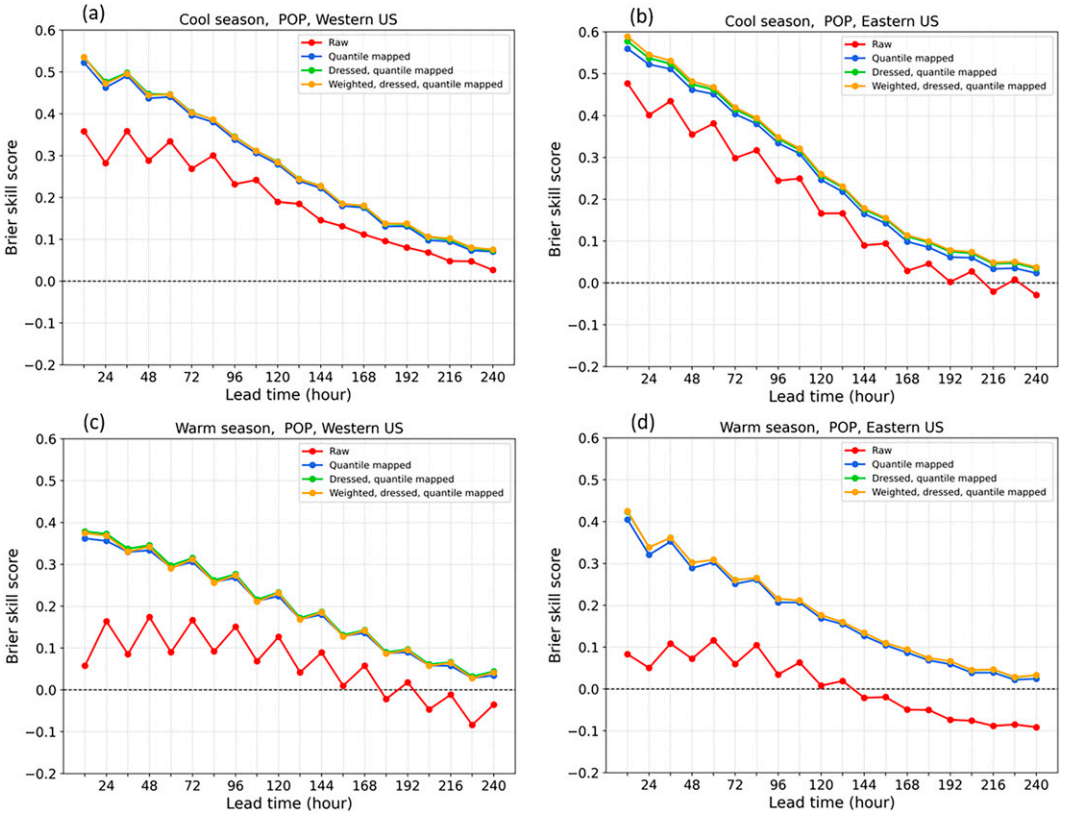

FIG. 3. Brier skill scores for POPs during the (a),(b) cool and (c),(d) warm seasons in the (left) western and (right) eastern United States.

Although the skill in the eastern U.S. cool season started out highest of the four categories, the skill dropped off so quickly that by 120+ h, the skill in the eastern U.S. cool season was lower than both the western U.S. cool and warm seasons. By +240 h, the skill in the eastern U.S. cool season was the lowest of the four categories.

We note that forecast skill is typically a function of spatial and temporal averaging, with more skill when averaging over larger areas and longer periods of time (e.g., Islam et al. 1993; Hamill 2014). While the skills shown here may appear small, especially at short leads, the short 6-h time window and verification on the ~3-km NDFD grid should be considered when interpreting results. Verification is performed on this fine grid and with this accumulation period, as this reflects the eventual operational product's grid.

Another interesting characteristic in the POPs was that the BSS of the raw GEFSv12 displayed a diurnal "sawtooth" pattern, particularly in the western U.S. warm season (Fig. 3c) and eastern U.S. cool season (Fig. 3b). In the western U.S. warm season, the skill tended to be lower between 0600 and 1200 UTC (overnight hours) and higher between 1800 and 0000 UTC (afternoon hours). In the western U.S. warm season, particularly during the North American monsoon, occurrences of precipitation are much higher in the afternoon hours than overnight (Gochis et al. 2003). The smaller sample size of precipitation during the overnight hours likely affects the scores to produce the diurnal variation seen in Fig. 3c.

In contrast, in the eastern U.S. cool season, skill was lower between 1800 and 0000 UTC (afternoon hours) and higher between 0600 and 1200 UTC (overnight hours). This diurnal pattern was generally dampened when the postprocessing algorithm was applied.

The forecast BSSs for $P[\text{obs} > 10 \text{ mm } (6 \text{ h})^{-1}]$ are shown in Fig. 4. Similar to the POPs, the raw and quantile-mapped GEFS were more skillful at early lead times in the eastern U.S. cool season (Fig. 4b), but the greatest improvement from applying the quantile mapping occurred in the western United States during both the cool and warm seasons (Figs. 4a,c). Note here that applying the weighting and dressing did not add much skill over the improvements made from the quantile mapping alone; while reliability was improved (previous figures), again, apparently the forecast resolution was degraded by nearly the same amount. Forecast skill in the cool season dropped to BSS values less than 0.1 after +144 h, even when the quantile mapping was applied (Figs. 4a,b). In the warm season, there was generally no skill in the raw GEFSv12 in the western United States, but some skill was added back into the forecast at early lead times when quantile mapping was applied (Fig. 4c). However, skill generally remained low overall, especially after +168 h when the BSS was near zero. In the eastern U.S. warm season, skill in the raw GEFS started out much higher relative to the western U.S. warm season, but improvements from quantile mapping were not as great and fell to near zero after the +192-h lead time

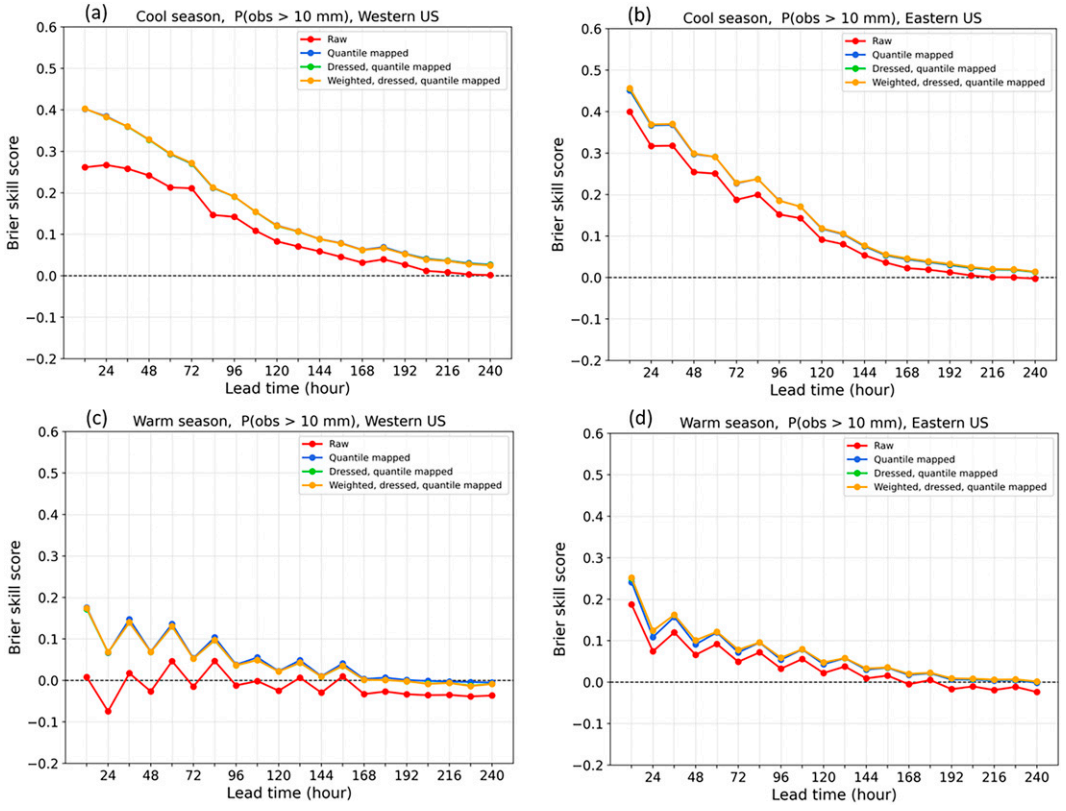

FIG. 4. Brier skill scores for $P[\text{obs} > 10 \text{ mm } (6 \text{ h})^{-1}]$ for the (a),(b) cool and (c),(d) warm seasons in the (left) western and (right) eastern United States.

(Fig. 4d). Note that, for $P[\text{obs} > 10 \text{ mm } (6 \text{ h})^{-1}]$, the diurnal variation of skill was not as pronounced as for the POPs.

For the heavy-rain threshold, $P[\text{obs} > 25 \text{ mm } (6 \text{ h})^{-1}]$, the raw GEFSv12 displayed marginal skill in the cool season up to a +72-h lead time (Figs. 5a,b) but displayed little to no skill in the warm season (Figs. 5c,d). Quantile mapping improved the skill in the western United States more than the eastern United States, but weighting and dressing did not provide much benefit over the QM alone. At this threshold, the largest impact that the quantile mapping algorithm had was on forecasts in the western United States, especially during the cool season at short lead times (Fig. 5a). Although the quantile-mapped GEFSv12 was more skillful than the raw GEFSv12 after +72 h, it was only marginally better, since BSS values fell to about less than 0.1 thereafter. Furthermore, like the lower-end precipitation thresholds, there did seem to be a diurnal pattern associated with the skill of the quantile-mapped forecast (Figs. 5a–c). In particular, the quantile-mapped forecast during the western U.S. cool season was more skillful in the 1800–0000 UTC time frame (afternoon) and less skillful in the 0600–1200 UTC time frame (overnight; Fig. 5a).

c. *Case studies*

1) WARM-SEASON EVENTS

The previous results demonstrated that quantile mapping of the GEFSv12 POPs significantly improved reliability and skill out to a +240-h lead time for warm-season events over the CONUS domain. After postprocessing, skill was highest with the most improvement shown at early lead times (e.g., Figs. 3c,d). The reliability diagrams demonstrated that an advantage of the quantile mapping routine was that it corrected the tendency that the raw GEFSv12 had to overestimate mid- to high-end probabilities across the entire United States. It also corrected the tendency that it had to underestimate low-end probabilities, which occurred primarily in the western United States at early lead times. Several cases from July 2019 will be shown that spatially demonstrate how the postprocessing routine alters the raw forecast. The cases should provide context to the objective verification previously shown and give a sense of some of the limitations to the routine.

The 6 h ending at 0000 UTC 16 July 2019 was an active precipitation period in both the eastern and western United States. In the eastern United States, Tropical Storm Barry had weakened into a tropical depression (TD), bringing bands of moderate to heavy precipitation to the Lower Mississippi Valley and Ohio River Valley (Fig. 6a). The upper Midwest from Minnesota through Wisconsin and Michigan received some of the heaviest precipitation during this period, with some areas of convection exceeding 50 mm $(6 \text{ h})^{-1}$. In the western United States, a monsoonal moisture plume had extended as far north as western Montana. This brought elevated convective precipitation through almost the entire north–south extent of the

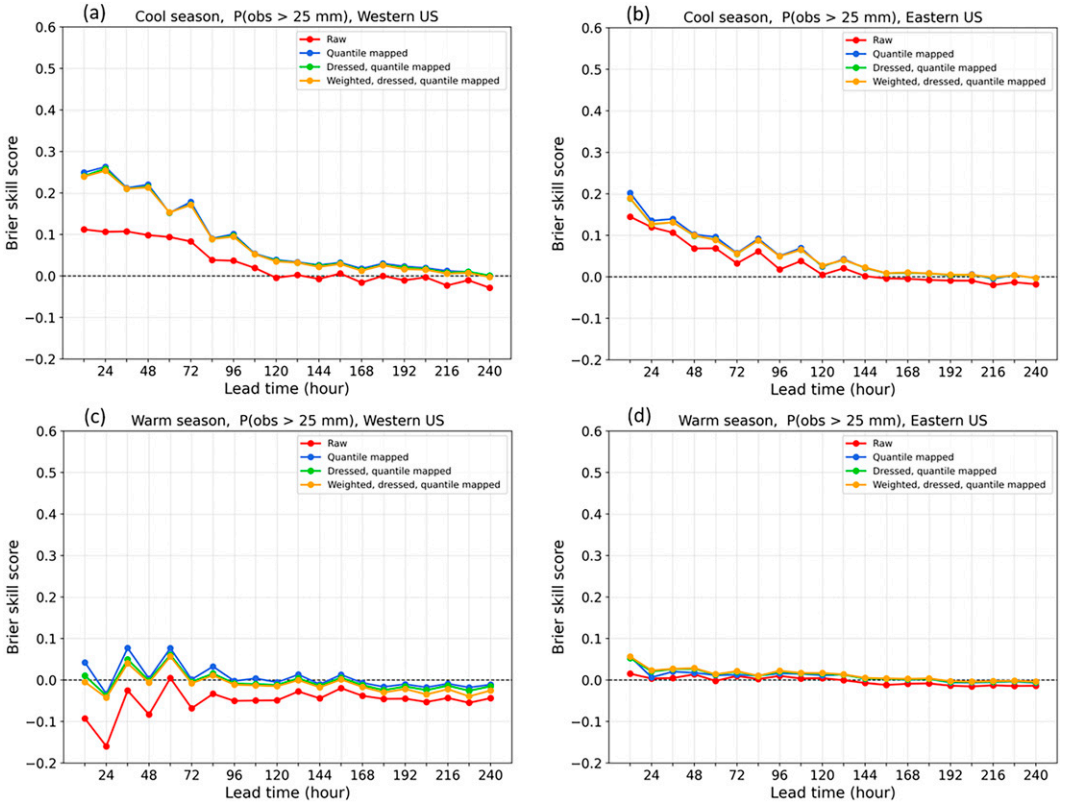

FIG. 5. As in Fig. 4, but for P[obs > 25 mm (6 h)$^{-1}$].

United States Rockies, spanning as far south as southeastern Arizona.

There were several areas worth noting from the raw GEFSv12 POPs that were either underestimated or overestimated at the +24-h lead time (Fig. 6b). The blue arrows in the images point to areas in the western United States where at least 0.254 mm (6 h)$^{-1}$ precipitation was analyzed by the CCPA/MSWEP (contoured, stippled) but where the POPs were zero. Central Colorado (arrow 1) and southern Montana and western Wyoming (arrow 2) were some of these areas. The blue ovals highlight the approximate areas in the eastern United States where there was no observed precipitation but where the raw GEFSv12 POPs were over 50%. In particular, the POPs were overestimated along the western and northern edges of the verifying precipitation for TD Barry (ovals 1 and 2). Rainfall on the periphery of the system in the upper Midwest was also overestimated, especially along the southwestern and northwestern edges located over Minnesota (ovals 3 and 4).

When the quantile mapping was applied, higher-end probabilities were reduced across the entire United States (Fig. 6c). In the eastern United States, low- and midrange probabilities were reduced along the edges of the precipitation systems that were over-forecasted by the raw GEFSv12 (e.g., ovals in Fig. 6b). The POPs on the northern periphery of TD Barry where the raw GEFSv12 had values over 50% were reduced to values between 0% and 30% after quantile mapping. Probabilities also substantially reduced along the system going through Minnesota, Wisconsin, and Michigan, constraining the higher-end POPs to areas that actually verified as having rainfall. In contrast, the low- and midrange probabilities in the western United States had increased in spatial coverage, effectively capturing more of the analyzed footprint of precipitation over Colorado and Wyoming that the raw GEFSv12 had missed. With each additional step of the postprocessing, the low- and midrange probabilities slightly increased, which corrected the tendency that the raw and QM GEFSv12 had to under forecast the low-end probabilities in the Rocky mountain region of the western United States (Fig. 6d).

The POPs at the +24-h lead time for precipitation analyzed between 1800 and 0000 UTC 18 July 2019 (Fig. 7a) displayed similar results to the previous case. Notably, quantile mapping dampened higher-end probabilities across most of the United States and increased the spatial coverage of lower-end probabilities in the west and decreased coverage of low- and midrange probabilities in the east (Figs. 7b,c). Adding the additional steps of weighting and dressing increased low- and midrange probabilities across the entire United States (Fig. 7d). One thing notable about this particular period is that there was a mesoscale convective system (MCS) that produced 75 mm (6 h)$^{-1}$ of rainfall over eastern Missouri and Iowa that was inadequately captured by both the raw and postprocessed GEFS (arrow 1). The POPs from the raw

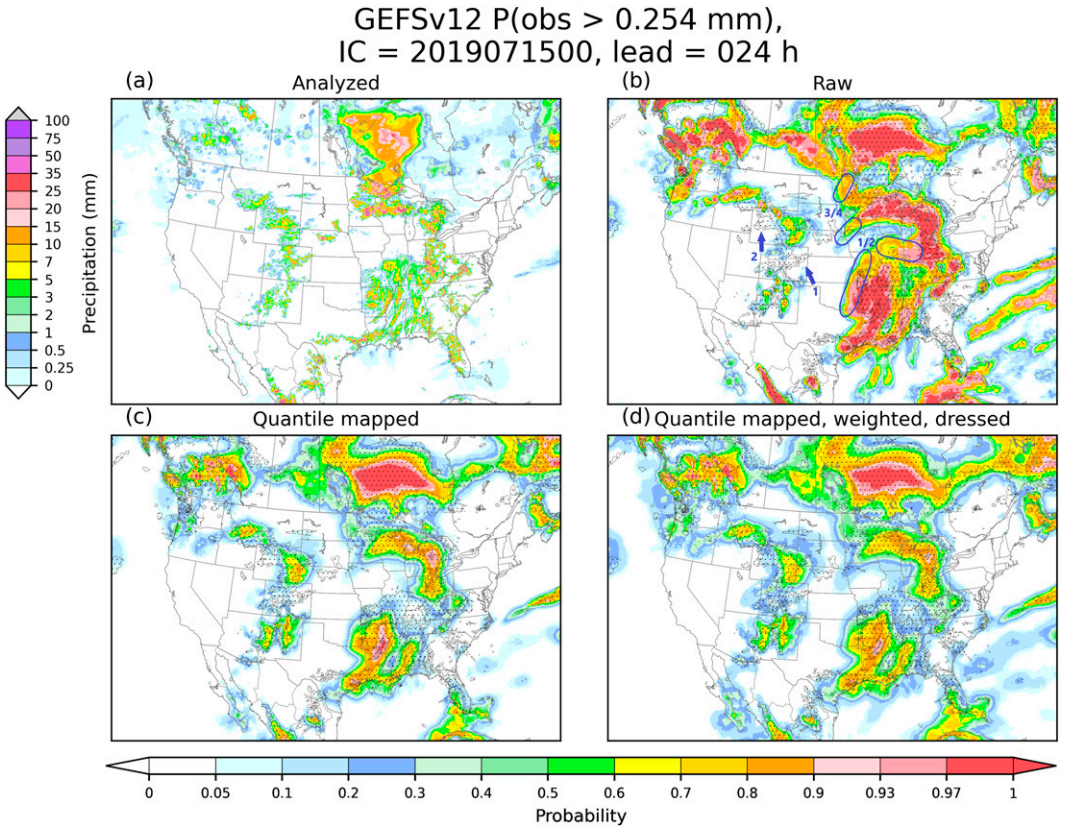

FIG. 6. (a) Analyzed precipitation from the CCPA/MSWEP for the 6-h period ending at 0000 UTC 16 Jul 2019. The GEFS probability of precipitation exceeding 0.254 mm in 6 h for the +24-h lead time is shown for the corresponding (b) raw; (c) quantile-mapped; and (d) quantile-mapped, weighted, and dressed forecasts (initialized at 1800–0000 UTC 15 Jul 2019). The 0.254-mm contour (black line) from the CCPA/MSWEP is overlaid (stippled, with black outline) on (b)–(d) for comparison.

GEFS had shown probabilities around 80% near the area where heavy rainfall occurred on the western side of the MCS, but postprocessing reduced the probabilities to around 10% (Figs. 7b–d); this was likely due to the application of quantile mapping using the 5 × 5 stencil that incorporated surrounding dry forecast points.

2) COOL-SEASON EVENTS

In both the raw and postprocessed GEFSv12, skill and reliability were generally better for cool-season events than for warm-season events, especially for POPs in the first 120 h of the forecast. For cool-season precipitation, the regional analysis of the BSS indicated that although skill was generally higher for eastern U.S. events, quantile mapping added the greatest improvement to the western United States (Fig. 3a). To demonstrate what quantile mapping does spatially for cool-season events, the raw and postprocessed POPs will be compared for the 6-h observed precipitation ending at 0000 UTC 2 March 2018 (Fig. 8a). During this period, there was a strong nor'easter impacting the eastern United States. The heaviest precipitation from this system brought values exceeding 25 mm $(6\ \text{h})^{-1}$ to an area centered near the border of Ohio and Pennsylvania, with another weaker precipitation maxima spanning laterally through North Carolina. In the western United States, a strong low pressure system brought heavy precipitation to areas of high terrain, with 6-h amounts exceeding 25 mm in the Sierra Nevada range and along the coastal ranges of California and Oregon.

The +24-h lead time from the raw and quantile-mapped GEFS had POPs that showed some similar characteristics to the warm-season case analyses (Figs. 8b,c). Specifically, quantile mapping reduced higher-end probabilities across the entire United States, dampened lower-end probabilities in the east, and expanded the area of lower-end probabilities in the west. Note that in the western United States, the quantile mapping reduced the spatial extent of high probabilities to areas of higher terrain in California's coastal mountain ranges, the Sierra Nevada, the Cascade Range through Oregon, and the Rocky Mountains from Utah up through Idaho and northern Washington (Fig. 9). Probabilities were also reduced in areas of lower elevation through the northwestern United States, effectively downscaling to delineate some of the dominant terrain features. The improved representation of the POPS over the high terrain can be seen out to at least the +120-h lead time (Figs. 9d,f).

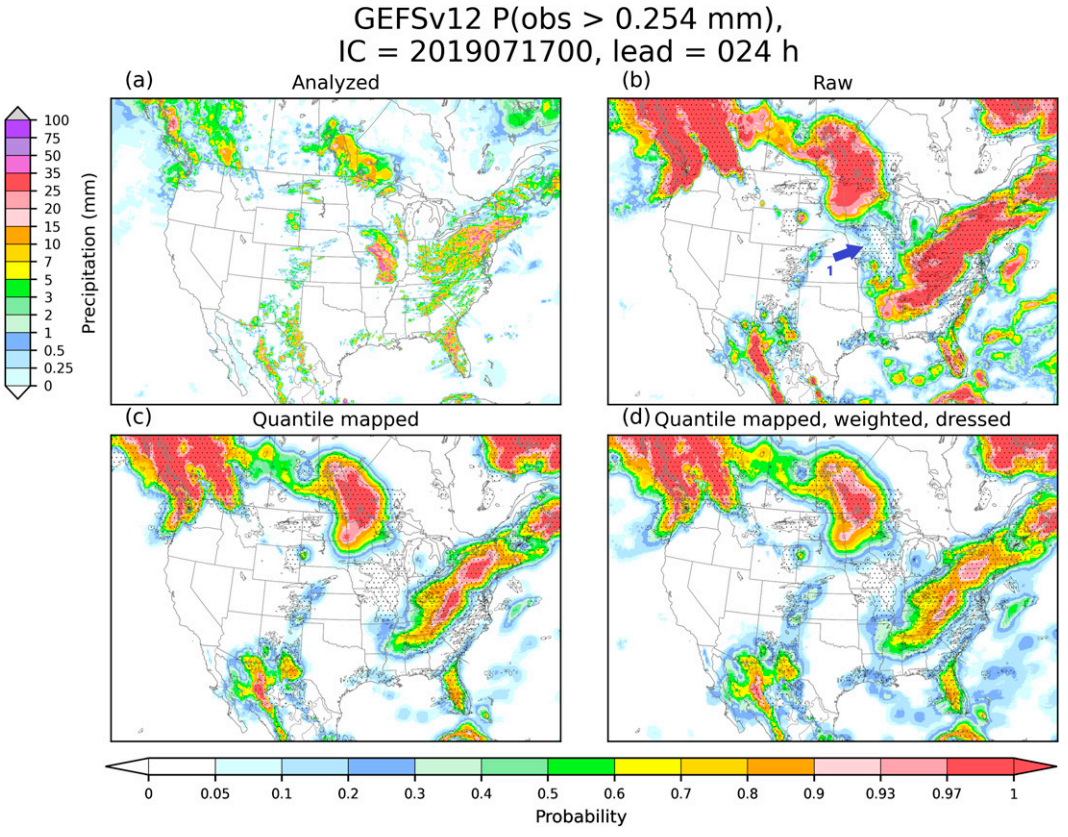

FIG. 7. As in Fig. 6, but (a) for the 6-h period ending at 0000 UTC 18 Jul 2019 and (b)–(d) initialized at 0000 UTC 17 Jul 2019.

Another case study from the 6-h period ending at 0000 UTC 3 February 2019 demonstrates some of the strengths that the quantile mapping routine had on cool-season events in the western United States. During this period, an atmospheric river made landfall and brought heavy precipitation to southern California and the Sierra Nevada range. There was an area that received over 75 mm (6 h)$^{-1}$ over the San Gabriel mountain range. Figure 10 shows P[obs $>$ 5 mm (6 h)$^{-1}$] for the +24- (Figs. 10a,b), +72- (Figs. 10d,e), and +120-h (Figs. 10g,h) lead times along with a black contour that encompasses the observed precipitation equal to or exceeding 5 mm. The analyzed precipitation from the CCPA/MSWEP (Figs. 10c,f,i) is also shown for comparison. The quantile mapping improved the spatial representation of the higher-end probabilities for this threshold over the terrain in southern California. In some areas, quantile mapping caused the lower-end probabilities to increase in spatial coverage. This is especially noticeable at all shown lead times over the northern California coastal mountain ranges and in northwestern Arizona at the +72- and +120-h lead times.

Also, recall that although reliability did not improve much for P[obs $>$ 25 mm (6 h)$^{-1}$], the BSS significantly increased for western United States, cool-season events with quantile mapping (Fig. 5a). Figure 11 shows that, relative to the raw GEFS at the +24-, +72-, and +120-h lead times, the quantile mapped forecast for P[obs $>$ 25 mm (6 h)$^{-1}$] was spatially more accurate, with enhanced probabilities over the San Gabriel mountain range. The quantile-mapped forecast indicated a chance for heavy precipitation in the area that received 75 mm of 6-h rainfall, even out to a +120-h lead time. This result suggests that the P[obs $>$ 25 mm (6 h)$^{-1}$] from the quantile-mapped GEFS could be a useful metric for determining locations of heavy rainfall during cool-season events in the western United States.

4. Summary and discussion

In this second part of this series, results were presented demonstrating the effect that a new quantile mapping routine proposed for the NBM had when applied to retrospective forecasts from the 0000 UTC cycle of the GEFSv12 between December 2017 and November 2019. The proposed new method leveraged multidecadal reforecasts and would supersede the current NBM method for GEFSv12 data.

The existing method generates forecast CDFs used in the quantile mapping from the previous 60 days of forecast and analyzed data. Small sample sizes were addressed by supplementing training data with forecasts and analyses from other locations with similar climatologies and terrain features ("supplemental locations"). This previous method had suboptimal skill and reliability during transition seasons (e.g., going from climatologically wet summer to dry autumn) and sometimes misrepresented the intensity and spatial distribution of

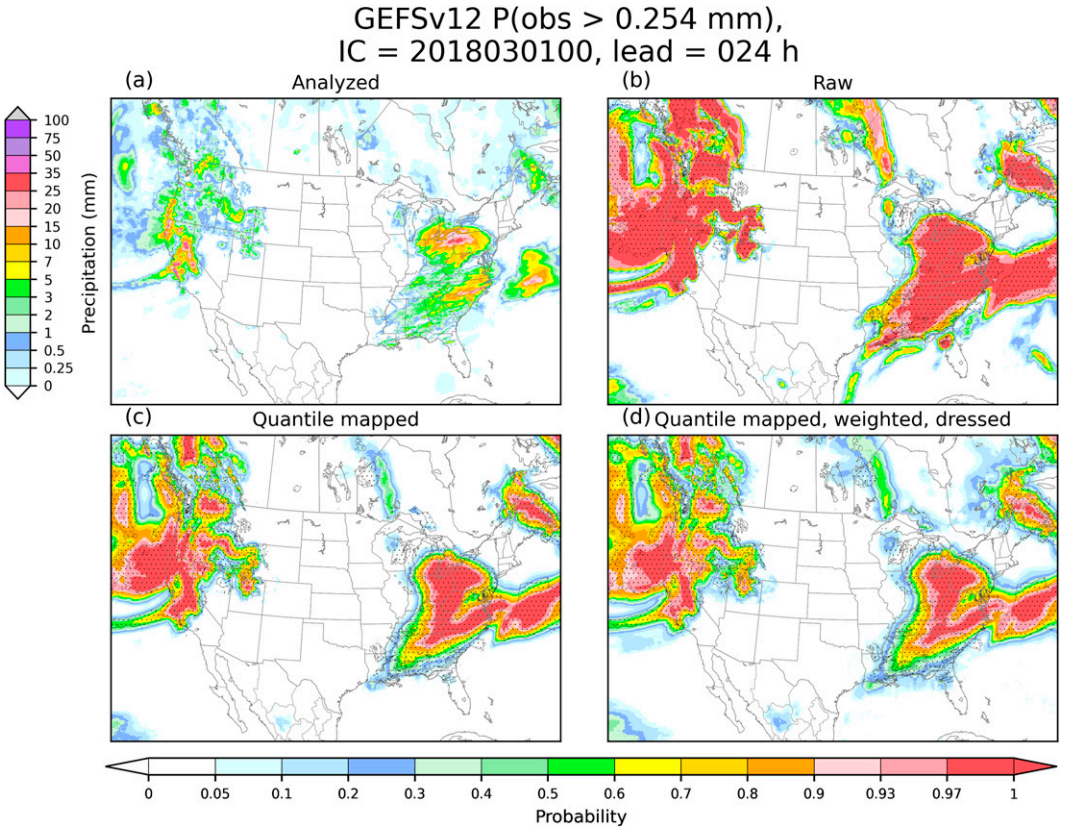

FIG. 8. As in Fig. 6, but (a) for the 6-h period ending at 0000 UTC 2 Mar 2018 and (b)–(d) initialized at 0000 UTC 1 Mar 2018.

precipitation events in high terrain of the western United States.

Part I describes the proposed new method, which utilized nearly two decades of reforecast and analyzed precipitation data and an advanced spline-fitting procedure to generate the forecast CDFs. The forecasts were also reweighted using closest-member histograms and subsequently dressed with a Gaussian probability distribution whose standard deviation depended on the forecast amount. The changes to quantile mapping and also to details of the closest-member reweighting and dressing resulted in downscaled, high-resolution PQPF forecasts that were shown to be significantly more reliable and skillful than the raw forecasts, especially at shorter lead times (i.e., <5 days) and for light to moderate precipitation events. The algorithm is also computationally fast, as the spline-fitting relationships and the weighting and dressing statistics have been precalculated for all lead times. With modest effort, the processing of GEFSv12 precipitation data in the NBM can be adjusted to use the new method. Because other models in the NBM do not have their own daily reforecasts, they would continue to rely on the 60 days of previous forecasts and the use of supplemental locations.

Reliability diagrams showed significant improvement in forecast reliability for all seasons and regions for all but the heaviest rainfall amounts. Forecasts of 6-h POPs were substantially improved across the CONUS out to +240-h lead time after the quantile mapping, weighting, and dressing were applied. For probabilities greater than 20%, the raw GEFSv12 overestimated POPs, but each step of the algorithm sequentially improved the guidance to produce high reliability for both warm and cool-season events. For values generally less than 20%, the raw GEFSv12 underestimated POPs out to at least a +72-h lead time that further degraded with quantile mapping. However, this was corrected when the dressing and weighting was applied. The raw GEFS also had a tendency to overestimate mid- to upper-range forecast values for $P[\text{obs} > 10\ \text{mm}\ (6\ \text{h})^{-1}]$ that was corrected with the quantile mapping, dressing, and weighting, though the improvement in reliability was most notable during the first 72 h of the forecast.

Brier skill scores also showed significant improvement for all seasons and regions, particularly for lead times less than +120 h and for light- to moderate-precipitation events. After postprocessing, the absolute skill of the POPs was highest for cool-season events in the eastern United States. However, the reforecast-based procedure produced its greatest skill improvement relative to the raw guidance with warm-season events in both the eastern and western United States. For $P[\text{obs} > 10\ \text{mm}\ (6\ \text{h})^{-1}]$, skill was again highest for cool-season events in the eastern United States, but the greatest improvement from postprocessing occurred for cool-season events in the western United States. In addition, although reliability did not show substantial improvement for $P[\text{obs} > 25\ \text{mm}\ (6\ \text{h})^{-1}]$, the skill was notably improved for western U.S. cool-season events out to at least a +72-h lead

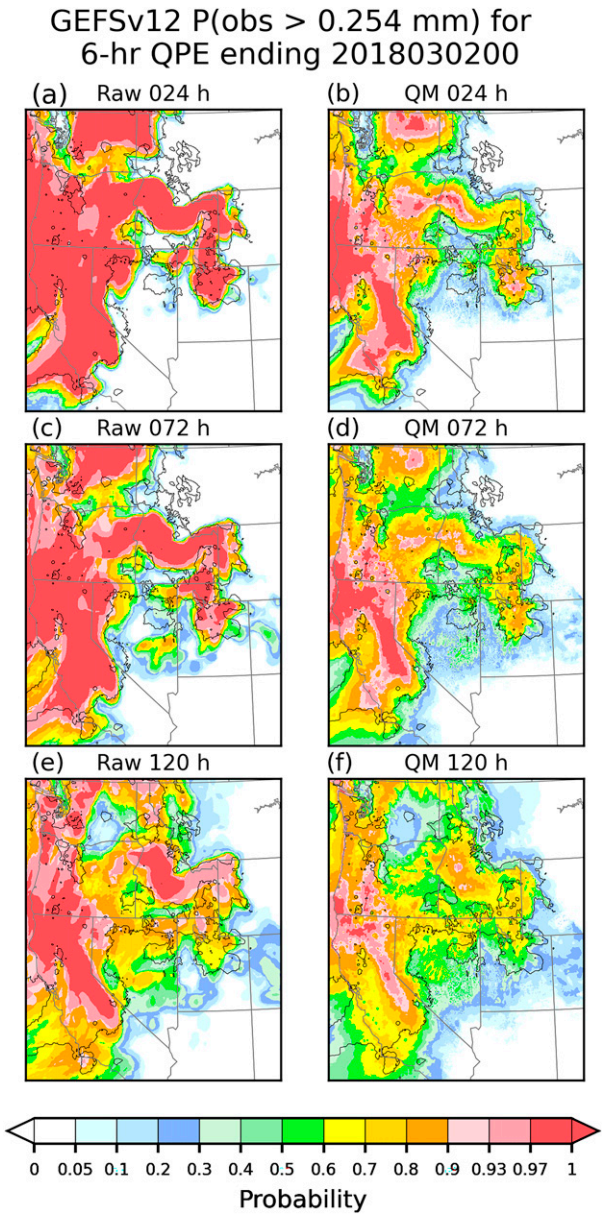

FIG. 9. Probabilities of 6-h rainfall exceeding 0.254 mm (POPs) for the (left) raw and (right) quantile-mapped, weighted, and dressed GEFSv12 at the (a),(b) +24-; (c),(d) +72-; and (e),(f) +120-h lead times. The verifying precipitation equal to or exceeding 0.254 mm from the CCPA/MSWEP for the 6-h period ending at 0000 UTC 2 Mar 2018 is outlined with a black contour for reference.

time. The lack of improvement in reliability of the heaviest forecasts may be due to the combining closest-member histograms for both moderate and heavy precipitation events. This might be addressed in future work with other methods for fitting closest-member histograms for very heavy events, a challenging problem given relatively few training samples.

The case studies presented visually demonstrated the strengths of the routine and provided insight to results from the reliability diagrams and Brier skill scores. For both warm and cool-season events, quantile mapping reduced higher-end POP values, generally dampening probability maxima for all system types (e.g., MCSs, cool-season synoptically driven events, tropical cyclones) across the United States. Low- and mid-end probabilities in the eastern United States were also reduced, but this generally helped to constrain the precipitation systems to areas that actually verified with 0.254 mm $(6\ \text{h})^{-1}$ of precipitation. In the western United States, the quantile mapping, weighting, and dressing effectively downscaled the probabilities to give a better representation of precipitation over the high terrain. The tendency for the raw and quantile-mapped GEFS to under-forecast low-end POPs in the western United States was visually demonstrated for both a warm- and cool-season event. When the dressing and weighting were applied, the lower-end probabilities expanded spatially to capture more of the analyzed precipitation from the CCPA. For moderate precipitation amounts {i.e., $P[\text{obs} > 5\ \text{mm}\ (6\ \text{h})^{-1}]$}, although the reliability diagrams demonstrated that quantile mapping corrected the tendency that the raw GEFS had to over-forecast probabilities by dampening the values, there were some instances where quantile mapping enhanced the probabilities, particularly in areas of the highest terrain in the western United States This often occurred for strong atmospheric river events during the cool-season in the western United States along the coastal mountain ranges and the Sierra Nevada range in California. Probabilities for heavy precipitation {i.e., $P[\text{obs} > 25\ \text{mm}\ (6\ \text{h})^{-1}]$} were generally found to be unreliable both before and after postprocessing. However, with the atmospheric river events analyzed, the downscaling in the western United States enhanced probabilities in higher-terrain to effectively improve the spatial representation of precipitation and provide a better signal to where the heaviest amounts actually occurred, even at a 5-day lead time.

There are some caveats and limitations that exist with the updated algorithm. Besides the few scenarios where the post-processed GEFS correctly enhanced probabilities for heavy rainfall (i.e., western United States, cool season, and atmospheric river events), reliability and skill for heavy rainfall {i.e., $P[\text{obs} > 25\ \text{mm}\ (6\ \text{h})^{-1}]$} generally did not improve when the quantile mapping, dressing, and weighting were applied. In addition to possible changes to closest-member histograms, it is also possible that probabilistic forecasts of heavy rainfall might be produced with more skill using more sophisticated machine-learning algorithms, such as neural networks discussed in Ghazvinian et al. (2022) or random-forest decision tree learning as was used in Herman and Schumacher (2018). Furthermore, bias correction of the location and timing of precipitation events are outside the capabilities of the algorithm. These biases can still arise for landfalling TCs, timing, and initiation of MCSs, and convective events that occur east of the Rocky Mountains in areas of flatter terrain. The quantile mapping routine cannot provide a realistic statistical downscaling of warm-season convective precipitation whose location inside the 25-km GEFSv12 grid box is mostly random. The authors also acknowledge that because of resource limitations, a direct comparison was not made with respect to the legacy NBM method that used 60 days of prior forecasts.

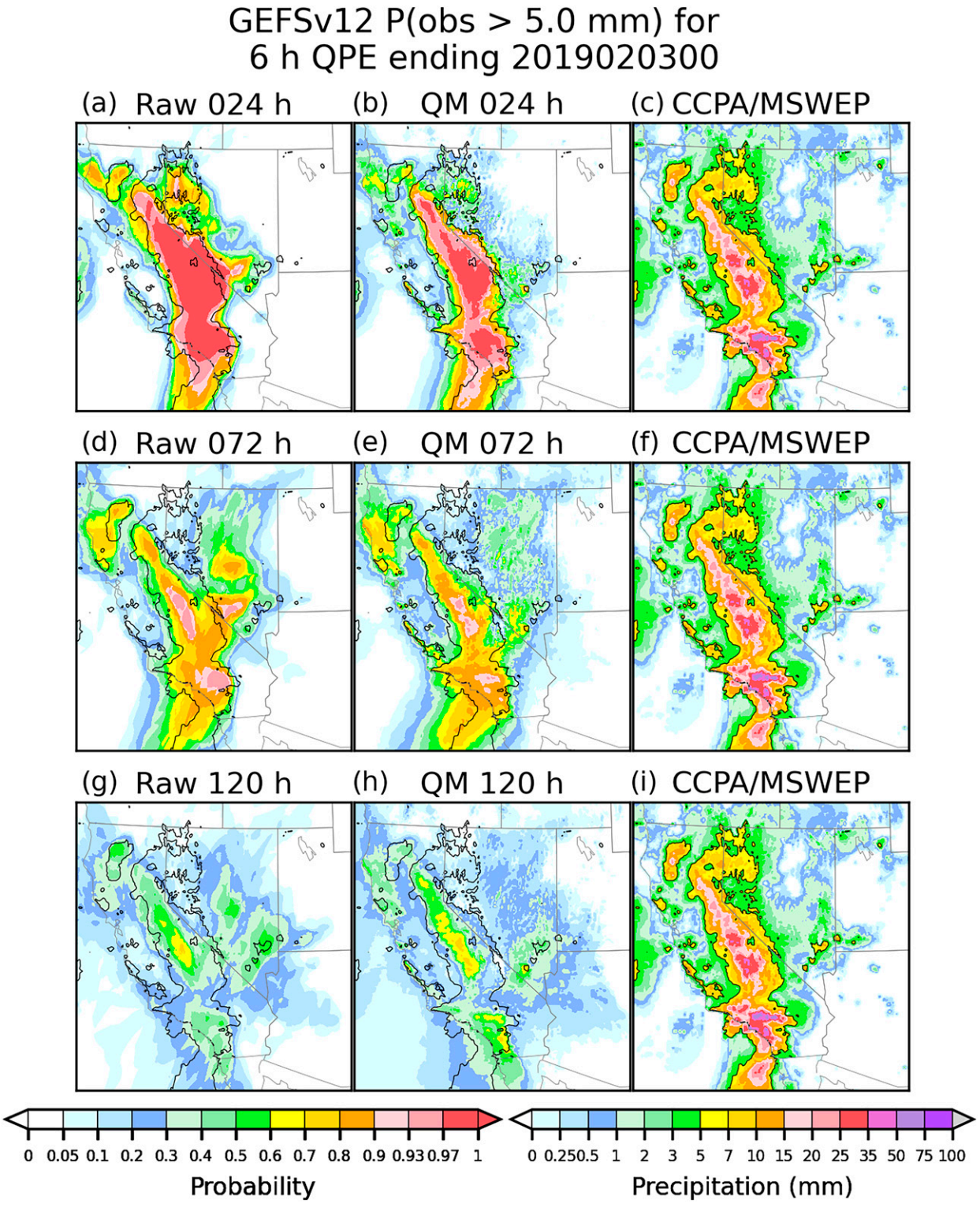

FIG. 10. Probabilities of 6-h accumulated rainfall exceeding 5 mm for the (left) raw and (center) quantile-mapped, weighted, and dressed GEFSv12 at the (a),(b) +24-; (d),(e) +72-; and (g),(h) +120-h lead times corresponding to the corresponding (c),(f),(i) verifying 6-h period ending at 0000 UTC 3 Feb 2019 from the CCPA/MSWEP. The area that verified with precipitation equal to or exceeding 5 mm is outlined with a black contour for reference.

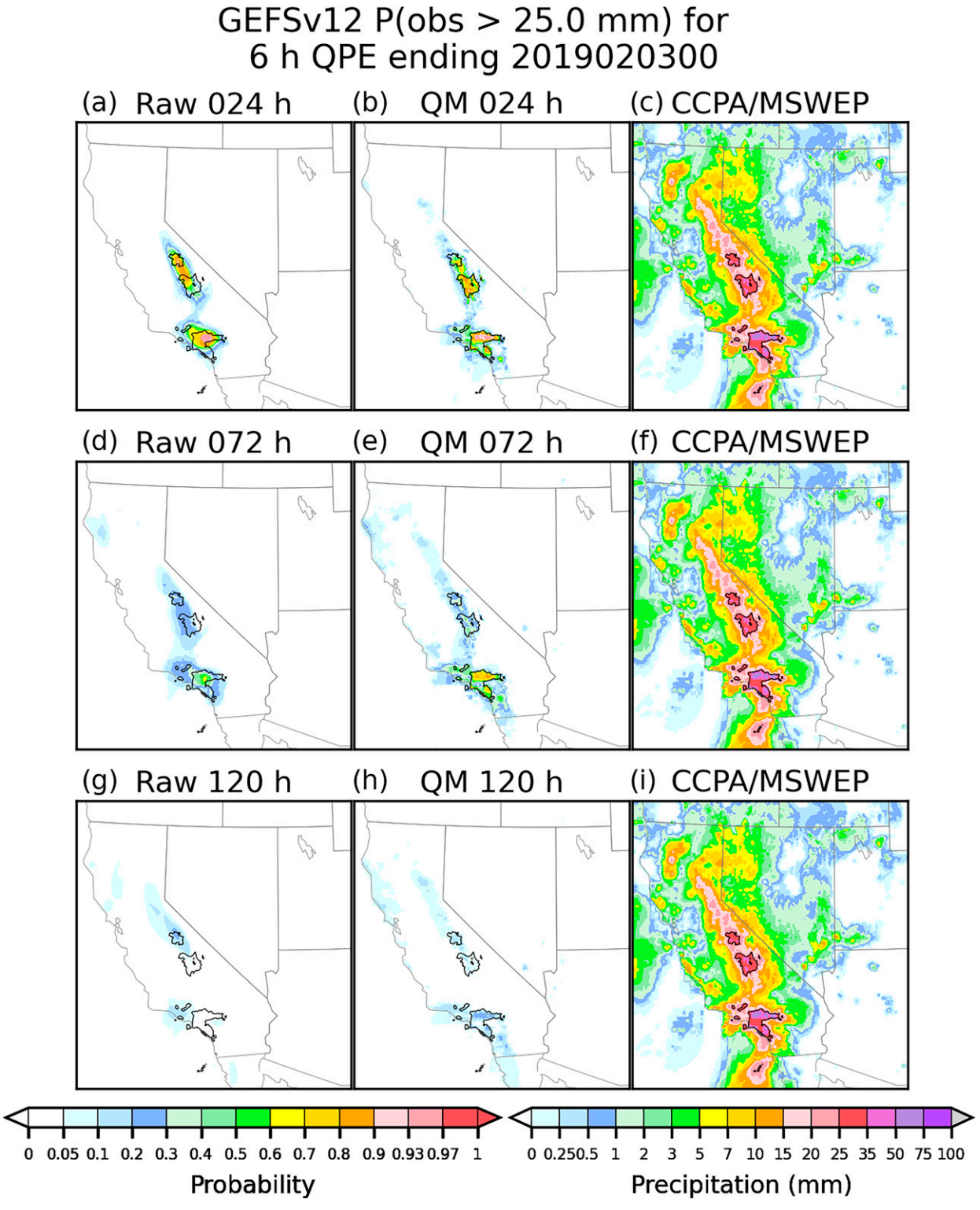

FIG. 11. As in Fig. 10, but for 6-h accumulated rainfall exceeding 25 mm for GEFSv12 and for areas that verified with precipitation equal to or exceeding 25 mm.

A likely next step for this research will be to examine the use of quantile-mapped ensemble members (without the 5 × 5 stencil) as precipitation forcing for hydrologic models such as the Hydrologic Ensemble Forecast System (Demargne et al. 2014), or the National Water Model (Gochis et al. 2020). The quality of the resulting streamflow forecasts will be evaluated against streamflow forecasts using the raw GEFSv12. Collaborative work with hydrologic partners from NWS River Forecast Centers and the Office of Water Prediction is ongoing to see how these forecasts can be used to produce skillful probabilistic forecasts and reduce uncertainty in streamflow forecasts. Given that the new QM procedure produces 31 ensemble members of detailed, downscaled precipitation with improved reliability and skill over the raw model, the authors hypothesize that ensemble streamflow forecasts should be improved as well.

Through NOAA's Earth System Research Laboratory/Physical Sciences Laboratory, 6-h precipitation forecasts of the quantile-mapped deterministic mean and exceedance probabilities are generated and displayed on a public website, which is updated daily using the 0000 UTC run from the GEFSv12 and displayed out to a +240-h lead time (https://www.psl.noaa.gov/forecasts/GQM/).

Acknowledgments. The authors thank Nachiketa Acharya for his review of the paper. This research was supported under a NOAA WPO Precipitation Grand Challenge Grant: NA17OAR4320101.

Data availability statement. The reforecast dataset analyzed in the current study was downloaded from the grib data store at Amazon Web Services (https://noaa-gefs-retrospective.s3.amazonaws.com/index.html). Links and references to additional datasets are included in Part I.

REFERENCES

Beck, H. E., E. F. Wood, M. Pan, C. K. Fisher, D. M. 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.

Brown, J. D., D.-J. Seo, and J. Du, 2012: Verification of precipitation forecasts from NCEP's short-range ensemble forecast (SREF) system with reference to ensemble streamflow prediction using lumped hydrologic models. *J. Hydrometeor.*, **13**, 808–836, https://doi.org/10.1175/JHM-D-11-036.1.

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.

Dahl, N., and M. Xue, 2016: Prediction of the 14 June 2010 Oklahoma City extreme precipitation and flooding event in a multiphysics multi-initial-conditions storm-scale ensemble forecasting system. *Wea. Forecasting*, **31**, 1215–1246, https://doi.org/10.1175/WAF-D-15-0116.1.

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.

Gehne, M., T. M. Hamill, G. N. Kiladis, and K. E. Trenberth, 2016: Comparison of global precipitation estimates across a range of temporal and spatial scales. *J. Climate*, **29**, 7773–7795, https://doi.org/10.1175/JCLI-D-15-0618.1.

Ghazvinian, M., Y. Zhang, T. M. Hamill, D.-J. Seo, and N. Fernando, 2022: Improved probabilistic quantitative precipitation forecasts using short training data through artificial neural networks. *J. Hydrometeor.*, **23**, 1365–1382, https://doi.org/10.1175/JHM-D-22-0021.1.

Gochis, D. J., J.-C. Leal, W. J. Shuttleworth, C. J. Watts, and J. Garatuza-Payan, 2003: Preliminary diagnostics from a new event-based precipitation monitoring system in support of the North American Monsoon Experiment. *J. Hydrometeor.*, **4**, 974–981, https://doi.org/10.1175/1525-7541(2003)004<0974:PDFANE>2.0.CO;2.

——, and Coauthors, 2020: The NCAR WRF-Hydro modeling system technical description, version 5.1.1. NCAR Tech. Note, 107 pp., https://ral.ucar.edu/sites/default/files/public/projects/wrf_hydro/technical-description-user-guide/wrf-hydro-v5.1.1-technical-description.pdf.

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.

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

——, 2014: Performance of operational model precipitation forecast guidance during the 2013 Colorado Front range floods. *Mon. Wea. Rev.*, **142**, 2609–2618, https://doi.org/10.1175/MWR-D-14-00007.1.

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

——, 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.

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

——, 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.

——, 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, https://doi.org/10.1175/MWR-D-15-0004.1.

——, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The United States 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.

——, D. R. Stovern, and L. L. Smith, 2023: Improving National Blend of Models probabilistic precipitation forecasts using long time series of reforecasts and precipitation reanalyses. Part I: Methods. *Mon. Wea. Rev.*, **151**, 1521–1534, https://doi.org/10.1175/MWR-D-22-0308.1.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, https://doi.org/10.1175/WAF-D-16-0093.1.

——, and ——, 2018: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

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.

Islam, S., R. L. Bras, and K. A. Emanuel, 1993: Predictability of mesoscale rainfall in the tropics. *J. Appl. Meteor.*, **32**, 297–310, https://doi.org/10.1175/1520-0450(1993)032<0297:POMRIT>2.0.CO;2.

Lewis, W. R., W. J. Steenburgh, T. I. Alcott, and J. J. Rutz, 2017: GEFS precipitation forecasts and the implications of statistical downscaling over the western United States. *Wea. Forecasting*, **32**, 1007–1028, https://doi.org/10.1175/WAF-D-16-0179.1.

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.

Savitzky, A., and M. J. E. Golay, 1964: Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, **36**, 1627–1639, https://doi.org/10.1021/ac60214a047.

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

Vannitsem, S., D. S. Wilks, and J. W. Messner, Eds., 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier Press, 362 pp.

——, and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

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.

Yuan, H., S. L. Mullen, X. Gao, S. Sorooshian, J. Du, and H.-M. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the Southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Wea. Rev.*, **133**, 279–294, https://doi.org/10.1175/MWR-2858.1.

Zhou, X., and Couthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **32**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Zhu, Y., and Y. Luo, 2015: Precipitation calibration based on frequency-matching method. *Wea. Forecasting*, **30**, 1109–1124, https://doi.org/10.1175/WAF-D-13-00049.1.