

Improving Probabilistic Quantitative Precipitation Forecasts Using Short Training Data through Artificial Neural Networks

MOHAMMADVAGHEF GHAZVINIAN,^a YU ZHANG,^a THOMAS M. HAMILL,^b DONG-JUN SEO,^a AND NELUN FERNANDO^c

^a *Department of Civil Engineering, The University of Texas at Arlington, Arlington, Texas*

^b *NOAA/Physical Science Laboratory, Boulder, Colorado*

^c *Texas Water Development Board, Austin, Texas*

(Manuscript received 14 February 2022, in final form 30 April 2022)

ABSTRACT: Conventional statistical postprocessing techniques offer limited ability to improve the skills of probabilistic guidance for heavy precipitation. This paper introduces two artificial neural network (ANN)-based, geographically aware, and computationally efficient postprocessing schemes, namely, the ANN-multiclass (ANN-Mclass) and the ANN-censored, shifted gamma distribution (ANN-CSGD). Both schemes are implemented to postprocess Global Ensemble Forecast System (GEFS) forecasts to produce probabilistic quantitative precipitation forecasts (PQPFs) over the contiguous United States (CONUS) using a short (60 days), rolling training window. The performances of these schemes are assessed through a set of hindcast experiments, wherein postprocessed 24-h PQPFs from the two ANN schemes were compared against those produced using the benchmark quantile mapping algorithm for lead times ranging from 1 to 8 days. Outcomes of the hindcast experiments show that ANN schemes overall outperform the benchmark as well as the raw forecast over the CONUS in predicting probability of precipitation over a range of thresholds. The relative performance varies among geographic regions, with the two ANN schemes broadly improving upon quantile mapping over the central, south, and southeast, and slightly underperforming along the Pacific coast where skills of raw forecasts are the highest. Between the two schemes, the hybrid ANN-CSGD outperforms at higher rainfall thresholds (i.e., $>50 \text{ mm day}^{-1}$), though the outperformance comes at a slight expense of sharpness and spatial specificity. Collectively, these results confirm the ability of the ANN algorithms to produce skillful PQPFs with a limited training window and point to the prowess of the hybrid scheme for calibrating PQPFs for rare-to-extreme rainfall events.

KEYWORDS: Ensembles; Forecast verification/skill; Hindcasts; Numerical weather prediction/forecasting; Operational forecasting; Probabilistic Quantitative Precipitation Forecasting (PQPF); Probability forecasts/models/distribution; Statistical forecasting; Neural networks

1. Introduction

Accurate, spatially detailed quantitative precipitation forecasts (QPFs) are of paramount importance for applications ranging from flash flood forecasting to reservoir management (Cloke and Pappenberger 2009; Pappenberger and Buizza 2009; Brown et al. 2014; Scheuerer et al. 2017). Despite continuing improvements in the accuracy of QPFs from numerical weather prediction (NWP) models, statistical postprocessing has remained a vital supplemental mechanism for enhancing the skill and spatial resolution of forecasts and for quantifying forecast uncertainties. Today, a plethora of conventional statistical postprocessing schemes exist that serve these purposes. These range from the analog method (e.g., Hamill and Whitaker 2006; Hamill et al. 2015), variants of Bayesian approach (Krzysztofowicz 2008; Wu et al. 2011; Robertson et al. 2013; Reggiani and Boyko 2019;

Darbandsari and Coulibaly 2022), and regression-based mechanisms (e.g., Hamill et al. 2004; Sloughter et al. 2007; Wilks 2009; Scheuerer and Hamill 2015; Taillardat et al. 2019; Ghazvinian et al. 2020).

The extant techniques, in particular the parametric schemes, have demonstrated wide success in augmenting the skill of PQPFs from diverse NWP systems for a range of lead times. Yet, there is growing recognition that additional room for improving the schemes might be limited, much due to the inflexible model structures and difficulties in selecting training samples for establishing predictor–predictand relationships (Ghazvinian et al. 2021). This recognition prompted various authors to explore newer, more flexible machine learning (ML) techniques as alternative postprocessing mechanisms (Herman and Schumacher 2018; Taillardat et al. 2019; Rasp and Lerch 2018; Bremnes 2020; Scheuerer et al. 2020; Baran and Baran 2021; Ghazvinian et al. 2021, Veldkamp et al. 2021; Chapman et al. 2022; Schulz and Lerch 2021; Li et al. 2022). In the field of precipitation forecasts postprocessing, Scheuerer et al. (2020) developed a multiclass artificial neural network (ANN) scheme for the subseasonal-to-seasonal range (week 2–4). Herman and Schumacher (2018) created a random forest–based postprocessing algorithm that has demonstrated prowess in producing skillful probabilistic guidance for days 1 and 2 during recent Flash Flood and Intense Rainfall (FFaIR) experiments (WPC 2019, 2020). More recently, Ghazvinian et al. (2021), drawing inspirations from Rasp and Lerch (2018), formulated a hybrid ANN–parametric framework

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-22-0021.s1>.

Ghazvinian's current affiliation: Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California.

Corresponding author: Mohammadvaghef Ghazvinian, mghazvinian@ucsd.edu

DOI: 10.1175/JHM-D-22-0021.1

© 2022 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

that fuses the ANN with the censored, shifted gamma distribution (CSGD; Scheuerer and Hamill 2015), namely, the ANN-CSGD. Relative to the traditional parametric methods, all these ML schemes offer flexibility in modeling predictor–predictand relationship and in integrating ancillary predictors and allow for adaptive selection of spatiotemporal training windows. Ghazvinian et al. (2021) demonstrated that ANN-CSGD broadly outperforms the original CSGD and mixed-type meta-Gaussian distribution (MMGD; Wu et al. 2011). This enhanced performance, as the authors explained, is attributable primarily to the adaptive and therefore more effective stratifications of training samples.

The challenges faced by postprocessing as a field go beyond the aforementioned limitations of parametric schemes alone. To date, a vast majority of contemporary schemes, including the more recent ML schemes, were formulated on the premise that extensive historical observations and retrospective forecasts (reforecasts) are available for training and calibration. In reality, however, reforecasts are often unavailable for many operational NWP systems in the United States and abroad. The U.S. National Weather Service's computing platform, for example, maintains archives of real-time forecast only for the past 60 days (see Hamill et al. 2017), and at present it is a practical necessity for any operational schemes to adapt to this short training window (Hamill 2018; Vannitsem et al. 2021). Note that the limited length of the training window aggravates the paucity in training sample that has already been an issue in the postprocessing of precipitation forecasts, necessitating the inclusion of compensatory measures. The quantile mapping and dressing (QMAP) algorithm (Hamill et al. 2017), the current operational algorithm of the U.S. National Blend of Models (NBM), addresses the data paucity by incorporating supplemental locations, i.e., locations sharing similar climatology and physiographic features such as elevation and topographic facets. Experiments performed by the authors have confirmed that this practice leads to sizable improvements in the skills of probability of precipitation (PoP) forecasts obtained through quantile mapping.

The aforementioned strengths of ANN models, in particular their ability to discern and establish complex predictor–predictand relationship from a large, heterogeneous sample, as we postulate, would render them particularly effective in alleviating the data paucity issue by intelligently expanding the domains where forecast–observation pairs are pooled. We further conjecture that ANN models' adaptive way of stratifying samples can lead to superior calibration beyond what is attainable by the QMAP that relies on prescribed supplemental locations. In this paper, we address these hypotheses by experimentally adapting and extending two ANN algorithms, namely, the ANN-Mclass (Scheuerer et al. 2020) and the ANN-CSGD (Ghazvinian et al. 2021), to short, 60-day training window. We perform a set of hindcast experiments over a 3-yr window for the entire CONUS, wherein we appraise the efficacy of the adapted schemes relative to the QMAP in processing ensemble QPF for day 1–8.

The present study expands the work of Ghazvinian et al. (2021) in three major directions. First, it entails comparisons of the skills of PQPFs produced by the hybrid ANN-CSGD and ANN-Mclass to determine the merit of the former. Second,

both ANN algorithms integrate ensemble attributes beyond the ensemble mean and incorporate geographic locations as well as physiographic features, thus allowing for exploitation of skill gains associated with the introduction of these predictors. Third, this study examines geographic variations in the relative performance of ANN and QMAP schemes to identify the dependence of skill differentials on precipitation regimes and accuracy in NWP forecast. Furthermore, this study assesses the skills of PQPFs for a range of thresholds much beyond the PoP, thereby providing critical information about the robustness of various schemes in forecasting intense precipitation events that is absent in extant studies in the context of NBM (Hamill et al. 2017; Hamill and Scheuerer 2018).

The remainder of this article is structured as follows. Section 2 describes the two ANN schemes as well as the benchmark QMAP model adapted for this study, layout of the hindcast experiment, and the datasets. Section 3 presents results of the hindcast experiment and discusses findings. Section 4 summarizes the work and offers concluding remarks.

2. Materials and methods

a. Postprocessing schemes

1) ANN WITH CATEGORICAL PROBABILITY PREDICTIONS (ANN-MCLASS)

Scheuerer et al. (2020) proposed a dense neural network-based postprocessing scheme that produces probabilities of 7-day precipitation totals falling into discrete categories at the subseasonal scale (i.e., weeks 2, 3, and 4). To elaborate, the scheme creates $m + 1$ possible classes of future observed precipitation probabilities. This is achieved by constructing precipitation climatology established from observation or analysis. Let $C_i = [q_{i-1}, q_i]$ denote the i th class where $i \in \{0, \dots, m\}$. Empirical quantile boundaries q_i are associated with the following probability levels:

$$p_{cl,i} = (1 - \text{pop}_{cl}) + \text{pop}_{cl}(i/m), \quad i = 0, \dots, m, \quad (1)$$

where pop_{cl} represent the probability of precipitation (>0.254 mm) from climatology. The first class = $[q_{-1}, q_0]$ represents precipitation values below 0.254 mm (dry conditions) and $q_m = \infty$. Scheuerer et al. (2020) chose to derive empirical quantiles from precipitation analysis for each grid point and day of the year using a 61-day window centered around the day of interest and all years of available data.

Scheuerer et al. (2020) proposed a modified categorical cross-entropy loss (MCCES)

$$L_{0,\dots,m}(\mathbf{p}, \mathbf{y}) = -\log\left(\sum_{i=0}^m y_i p_i\right), \quad (2)$$

where $\mathbf{p} = (p_0, \dots, p_m)$ is vector of estimated probabilities for each of $m + 1$ classes, and $\mathbf{y} = (y_0, \dots, y_m)$ is corresponding binary (one-hot encoded) truth vector that describes whether analyzed value falls into the respective category in a training case. This modification is necessary as the category assignment can be ambiguous (multiple cases with 1 due to duplicative

values in climatology). MCCES reduces to standard categorical cross entropy when the assignments are unambiguous (see appendix B of [Scheuerer et al. 2020](#)).

Continuous predictive distribution can be derived from the categorical probabilities by approximating the cumulative hazard function $H(x) = -\log[1 - F(x)]$ using piecewise linear interpolation/extrapolations for the points inside/outside the data ranges. In the hazard function, $F(x)$ represents cumulative probabilities estimated by summing up the probabilities specific for individual categories and for each forecast case. Exploratory analysis by [Scheuerer et al. \(2020\)](#) showed that the interpolation provides a reasonable reconstruction of predictive cumulative distribution function (CDF). A possible drawback of this model could be its reliance on the parameter m . As the number of classes directly impacts cross-entropy loss function value, other metrics such as continuous ranked probability score (CRPS; [Matheson and Winkler 1976](#), [Wilks 2019](#)) should be used for configuration of optimal number of classes for final predictions (see [Scheuerer et al. 2020](#)).

2) HYBRID ANN-CENSORED, SHIFTED GAMMA DISTRIBUTION (ANN-CSGD)

Censored, shifted gamma nonhomogeneous regression model (CSGD) first was introduced by [Scheuerer and Hamill \(2015\)](#). This technique and its extensions have been shown to be capable of generating reliable and skillful medium-range PPDFs over different regions of the world (see, [Baran and Nemoda 2016](#); [Zhang et al. 2017](#); [Baran and Lerch 2018](#); [Taillardat et al. 2019](#); [Ghazvinian et al. 2020](#); [Valdez et al. 2022](#)). Nonetheless, [Ghazvinian et al. \(2020\)](#) noted a caveat of the CSGD that stems from the direct use of climatological shift without tuning and showed that this led to a negative bias in predicted PoP, particularly for shorter lead times.

Heeding the success of the hybrid neural network-parametric postprocessing scheme of [Rasp and Lerch \(2018\)](#), [Ghazvinian et al. \(2021\)](#) formulated a similar, hybrid ANN-CSGD framework. The new framework retains the use of CSGD as the predictive distribution but employs a fully connected neural network structure that links the three CSGD parameters to ensemble statistics and ancillary predictors.

In this work, we follow the notations of [Ghazvinian et al. \(2021\)](#) and denote by $F_{k,\theta}$ the CDF of gamma distribution with parameters shape $k > 0$, scale $\theta > 0$. CDF of CSGD denoted by $F_{k,\theta,\delta}^0(y)$ with the shift parameter $\delta < 0$ is given by ([Scheuerer and Hamill 2015](#))

$$F_{k,\theta,\delta}^0(y) = \begin{cases} F_{k,\theta}(y - \delta), & y \geq 0 \\ 0, & y < 0 \end{cases}. \quad (3)$$

The three parameters of ANN-CSGD are optimized by minimizing the average value of CRPS closed form expression for CSGD over training dataset ([Scheuerer and Hamill 2015](#)).

[Ghazvinian et al. \(2021\)](#) demonstrated that ANN-CSGD alleviates the negative bias in PoP by directly learning shift parameter as an arbitrary function of predictors. In addition, the authors showed that ANN-CSGD's efficient way of modeling complex interactions between three CSGD parameters is a

major factor that contributed to its superior predictive skill at higher thresholds relative to the original CSGD.

3) QUANTILE MAPPING STENCIL

Chosen as the benchmark technique is the QMAP algorithm a coarse-grid approximation to the current baseline postprocessing scheme for a single ensemble prediction system component of the NBM. The algorithm was first described by [Hamill et al. \(2017\)](#) and was later modified in [Hamill and Scheuerer \(2018\)](#) to incorporate quantile dressing. In essence, it establishes matching quantiles from training data (forecasts and corresponding analyses for designated locations). The resulting quantile map function is then applied to real-time forecasts to produce probabilistic guidance grids. To augment sample size, the algorithm incorporates the so-called supplemental locations. The supplemental locations are locations that share similar precipitation climatology (as represented by climatological CDF), elevation, and terrain orientation (facet). These were determined using an approach similar to Parameter-Elevation Regressions on Independent Slopes Model (PRISM; see, [Daly et al. 2008](#)), which involves the use of a composite distance measure that is based on the aforementioned variables. Additional details of the method can be found in supplemental material of [Hamill et al. \(2017\)](#).

In this study, we implement the original version of QMAP as described by [Hamill et al. \(2017\)](#), but for simplicity chose to forgo the dressing mechanism—we do so with the tacit assumption that the incremental improvements in skills as a result of dressing would be insufficiently large to alter the relative performance of the algorithms. In our implementation of the QMAP scheme, up to 100 supplemental locations are identified for each target grid point ($0.25^\circ \times 0.25^\circ$ in size), and for each month of the year using the data from the respective month and two surrounding months. For each lead time and the grid point of interest, empirical quantiles $q(p)$, $p \in (1/100, 2/100, \dots, 99/100)$ are constructed from the augmented forecast and analysis datasets, which are accumulated within a rolling window and supplemental locations. Note that in this step, ensemble members from Global Ensemble Forecast System (GEFS), are pooled to populate forecast CDFs based on the assumption that these members are identically distributed. To account for wider sampling variability in larger forecast quantiles (larger than 0.95 quantile of forecasts), a linear approximation of quantile mapping function is applied. A detailed explanation of this procedure can be found in [Hamill et al. \(2017\)](#).

The quantile map thus established is then used to transform members of ensemble forecasts from following day and in a 5×5 stencil of surrounding grid points using the forecast CDF of each point and analysis CDF of the center grid point. Using expanded spatial domain enlarges the ensemble size and reduces the sampling error in ensemble and helps mitigate potential mismatches in forecast and analyses quantiles due to displacement errors in forecast ([Hamill et al. 2017](#)). The exceedance probability for each precipitation threshold is computed using the fraction of the transformed members exceeding that threshold. To assess the impact of stencil size on the skills of postprocessed

PQPFs, the QMAP method was implemented on 1×1 (center point) and 5×5 stencils.

b. Architecture of two ANN schemes

To rigorously evaluate the performance of the two ANN schemes, we configure both to use identical predictors and on identical grid mesh ($0.25^\circ \times 0.25^\circ$). These predictors are ensemble statistics including ensemble mean, PoP and ensemble spread. Following the practice of [Scheuerer and Hamill \(2015\)](#), for each point we compute these statistics from a super ensemble constructed using members in an expanded spatial domain. In this study, each expanded domain is the 5×5 stencil surrounding the target point, thus maintaining consistency with the training scheme of QMAP. To simultaneously postprocess forecast over the entire grid mesh, we use geographical coordinates (latitude and longitude) of analyses grid points as predictors to the networks. As additional spatial features, we introduce grid terrain height and local terrain orientation information (facet). Note that we chose to exclude additional ensemble-based predictors (e.g., additional statistics from ensemble, control member; see, e.g., [Taillardat et al. 2019](#)); as our initial evaluation showed that the inclusion of these predictors failed to yield systematic improvement in predictive skill, possibly due to an increased risk of overfitting.

Both ANN models share a similar model architecture with differences in the shape of output layer where model predictions are derived. The architecture consists of the following elements:

- Input layer where predictors are introduced to the network.
- Batch normalization ([Ioffe and Szegedy 2015](#)). This practice normalizes input to maintain the mean of each feature close to 0 and the standard deviation close to 1.
- Hidden layers (*dense*) with nonlinear activation functions.
- Output layer

For the output layer depending on the model, specific configuration is required:

- ANN-CSGD uses output layer with linear activation function and three nodes to represent functions of three CSGD parameters. Additional functions are required to limit CSGD parameters in allowable ranges. We follow [Ghazvinian et al. \(2021\)](#) to set CSGD shift parameter, $\delta = -\sqrt{O_1^2}$, and use inverse logarithmic link functions for location and scale parameters, $\mu = \exp(O_2)$ and $\sigma = \exp(O_3)$, where O_i represents the i th output node.
- ANN-Mclass uses an output layer with 50 nodes and softmax activation function to ensure that output probabilities are in the range $[0, 1]$ and sum to 1. For each forecast day, observation quantiles are derived using CCPA data of previous 60 days and all CCPA grids. This yields a sample of 60 days $\times$ 13 528 locations worth of data from which the empirical quantiles are calculated daily. One-hot encoded CCPA data in each location and day are then assigned to the designated classes. The latter approach helps maintain balanced assignments between classes $i \in \{1, \dots, m\}$. As another structural modification to the work by [Scheuerer et al. \(2020\)](#), we do not include climatological

information in the last layer. This practice was necessary for postprocessing subseasonal forecasts to ensure that estimated PQPFs revert to climatology in the cases where signal to noise ratio was limited. In our application this is considered redundant as it complicates the model and makes overfitting more likely.

For each lead time and each day, previous 60 days' worth of data over the entire CONUS are available for training. To reduce generalization errors we use early stopping, one of the most efficient and widely used regularization techniques (see [Goodfellow et al. 2016](#)). In our application we keep the last 6 days of training data for validation (not used in training) and monitor its average loss value. The training is terminated when no further decrease in the loss is seen after three epochs (*patience*). To simplify matters, we decided to avoid extensive grid search for hyperparameter tuning. Loss functions were minimized using adaptive moment estimation (Adam) algorithm ([Kingma and Ba 2014](#)) with the learning rate kept fixed at $\text{lr} = 0.01$. The following model architectures for hidden layer were tested:

- Number of nodes in hidden layer(s) (ANN-CSGD): $\{[10], [20], [10, 10]\}$.
- Number of nodes in hidden layer(s) (ANN-Mclass): $\{[20], [50], [20, 20]\}$.

Our initial assessment showed that expanding the number of hidden layers does not systematically improve validation loss (possibly due to overfitting). Thus, we did not test number of hidden layers > 2 in final model configurations. The batch size was set to 10 000 for training both networks. Networks were trained with a common random number generator (seed) and the configuration with the lowest validation loss was saved for each out-of-sample day prediction. See the [appendix](#) for further implementation details.

c. Hindcast experiment setup

In this study we focus on postprocessing forecasts of 24-hourly accumulated precipitation for 1–8-day lead time over the CONUS. We leveraged Global Ensemble Forecast System (GEFS) version 11 (v11) reforecast datasets ([Hamill et al. 2013](#)). The GEFSv11 data we use were produced on a quadratic Gaussian mesh with $\sim 0.5^\circ$ resolution for the first 8 days and $\sim 0.67^\circ$ for 9–16 days lead times. The reforecasts are composed of 11 ensemble members (10 perturbed and one control) issued every 24 h at 0000 UTC. The reforecast data were retrieved, extracted for CONUS on native Gaussian grid, bilinearly interpolated to a 0.25° grid mesh and accumulated to 24-h sums. As the analyses we use Climatology-Calibrated Precipitation Analysis (CCPA; [Hou et al. 2014](#)), which is available on 6-h increments spanning from 1 January 2002 to 31 December 2019 on the CONUS National Digital Forecast Database (NDFD; [Glahn and Ruth 2003](#)) grid resolution (see <https://vlab.noaa.gov/web/mdl/ndfd-spatial-reference-system>). The latter data were upscaled to 0.25° resolution and accumulated to 24-h sums. To mimic the U.S. NBM operations as described in [Hamill et al. \(2017\)](#), training for each scheme is performed each day on forecasts for the lead times of 1–8 days and corresponding analysis

TABLE 1. Different experiment names and configurations used.

Model	Experiment name	5×5 stencil	Supplemental locations
ANN	ANN-CSGD	Yes (super ensemble)	No
	ANN-Mclass	Yes (super ensemble)	No
Quantile mapping	QMAP (5×5)	Yes	Yes
	QMAP (center point)	No	Yes
Raw ensemble	RAW (5×5)	Yes	No
	RAW (center point)	No	No

over previous 60-day rolling window. The trained schemes are then applied to forecasts of prediction day to create PQPFs that are then verified against coincidental analysis. This training-verification cycle repeats progressively in time over a 3-yr window extending from 1 January 2017 to 31 December 2019.

Note that the aforementioned studies used archive of real-time GEFS forecasts (20-member ensemble). As this dataset is only available for a short time window, the present study relies instead on the reforecast available for a longer time span but with fewer ensemble members. In addition, we apply a longer verification window than that used in the previous studies, and this allows for more robust assessments of the time and region-dependent performance differences.

3. Results

a. CRPSS

To assess the relative overall predictive performance of three postprocessing schemes, we first examine continuous ranked probability skill score (CRPSS). We use climatological CRPS as the reference. Climatological CRPS was calculated for each grid point, separately for each month using CCPA data pooled across a 3-month window surrounding that month from years between 2002 and 2015. For each forecast suite, we first compute CONUS-wide, lead-time specific CRPSS by aggregating either the raw ensemble PQPFs or postprocessed PQPFs among all grid points in CONUS and across all verification days within the 3-yr window. To highlight the impacts of using an expanded spatial domain, we computed the results using raw forecasts over each target grid point (center point) and corresponding 5×5 stencil. Note that Table 1 summarizes different model configurations and corresponding abbreviations used in the figures of this study.

Figure 1 shows the CRPSS results. It is evident that ANN based schemes show the best overall predictive performance across lead times. Quantile mapping from each source (5×5 stencil and center point) highly improves the overall performance of corresponding raw forecasts. Inclusion of forecasts from neighboring grid points (5×5 stencil) improves the skill of raw and quantile mapped forecast across lead times. This can be explained by the fact that expanding the spatial window helps alleviate displacement errors in the raw forecast and increase the spread, thus improving the calibration.

b. Brier skill scores

To assess the relative efficacy of three postprocessing schemes in predicting specific events, we examine the Brier

skill scores (BSS; Wilks 2019; Hamill et al. 2015) computed for four daily accumulation thresholds, namely, 0.254, 10, 25, and 50 mm. Figure 2 shows the resulting BSS from the six forecast sources. At the two lowest thresholds here, it is evident that raw GEFS forecast (11-member ensemble) from the center point is the most poorly performing forecast of all. Similar to CRPSS, quantile mapping appreciably improves PoP forecasts without and with the expanded domain (Fig. 2a). However, the gap between raw ensemble and quantile mapped PQPFs diminishes at higher thresholds. It appears that quantile mapping mainly improves BSS of PQPF from raw forecast at the lowest thresholds, which have disproportionate impacts on its overall prediction performance as shown in CRPSS results (Fig. 1). In fact, quantile mapped PQPFs broadly underperform raw ensemble forecasts at the highest threshold (50 mm; Fig. 2d). In addition, without domain expansion, PQPF from quantile mapping center point underperforms climatology across lead times (Fig. 2d). Post-processed PQPFs via the two ANN schemes manage to improve the forecast skill from raw and quantile mapped sources across all thresholds and throughout the lead times. It is also worth noting that ANN-CSGD and ANN-Mclass both

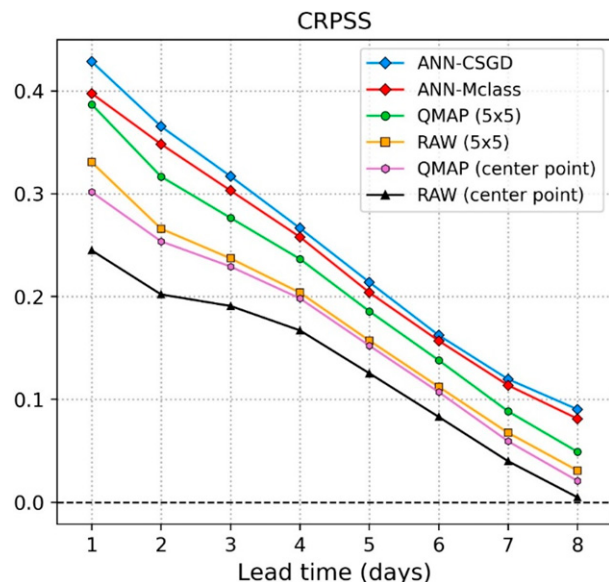


FIG. 1. Continuous ranked probability skill score (CRPSS) results computed over the CONUS and shown as a function of lead time. Climatological forecasts are used as the reference.

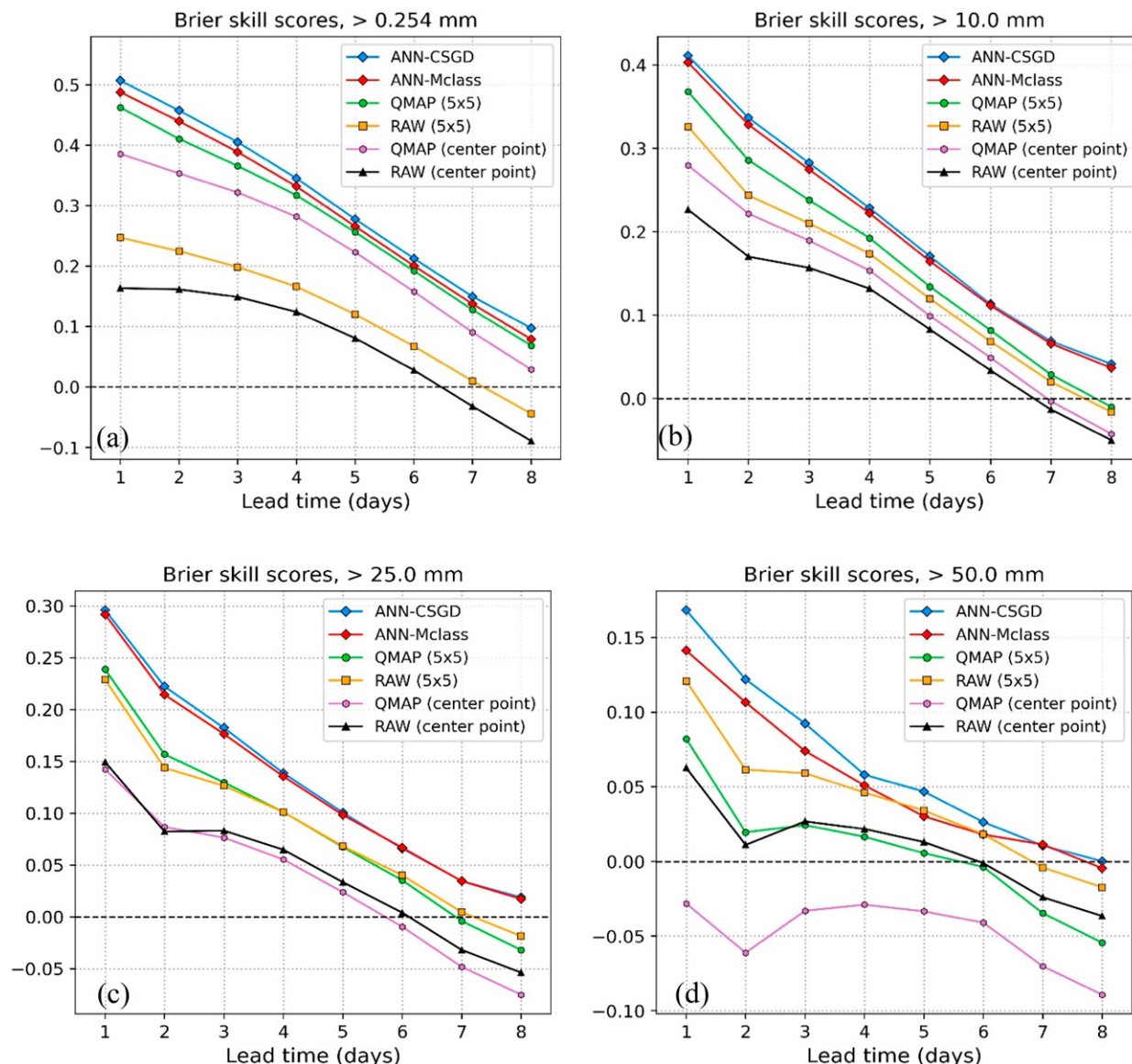


FIG. 2. Brier skill scores (BSSs) for exceeding events (a) >0.254 (PoP), (b) >10 , (c) >25 , and (d) >50 mm, computed over the CONUS and shown as a function of lead time. Climatological forecasts are used as the reference.

improve the skill of raw forecast at longer lead times where raw and quantile mapped forecasts are unskillful relative to climatology. Between the two ANN schemes, ANN-CSGD slightly outperforms ANN-Mclass, and this performance differential is the most evident at the highest threshold of 50 mm (Fig. 2d).

Figures 3–5 characterize the geographically dependent skills of raw ensemble and three postprocessed PQPFs, obtained by applying QMAP, ANN-Mclass, and ANN-CSGD, within the CONUS, where the skills are again characterized by BSS with climatology as the reference. We retain only the forecasts generated using 5×5 stencil as these tend to outperform those without domain expansion and focus on lead times from +48 to +72 h as the results for this lead time range are broadly

representative of the performance differentials of postprocessing schemes.

Figure 3 shows the maps of BSS computed with 0.254-mm threshold (PoP). Some of the prominent features mirror those noted in past studies. In particular, of all regions in the CONUS, the skill of raw GEFS ensemble appears highest along the Pacific coast to the west of the Cascade and Sierra Nevada mountain ranges (Fig. 3a). This reflects the high predictability of orographically induced, synoptically forced precipitation systems that are predominant rainfall producers in these regions (Brown et al. 2014; Hamill et al. 2015; Scheuerer and Hamill 2015). By contrast, forecast skills of GEFS are the lowest over parts of Texas and southern Florida, where BSS is overwhelming negative (Fig. 3a). The region of low BSS values extends northward to cover much

Brier skill score for lead time +48h to +72h, PoP (> 0.254 mm/24h)

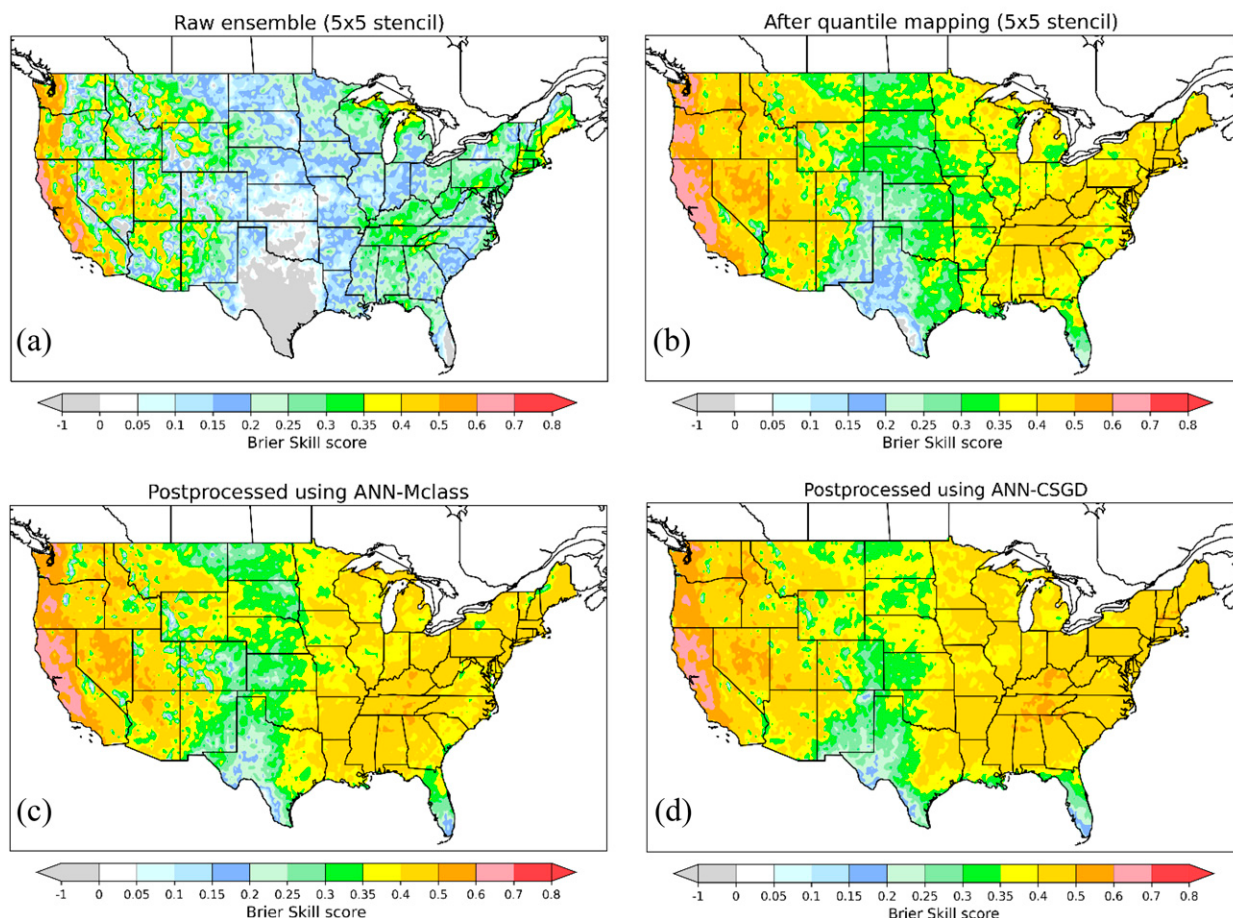


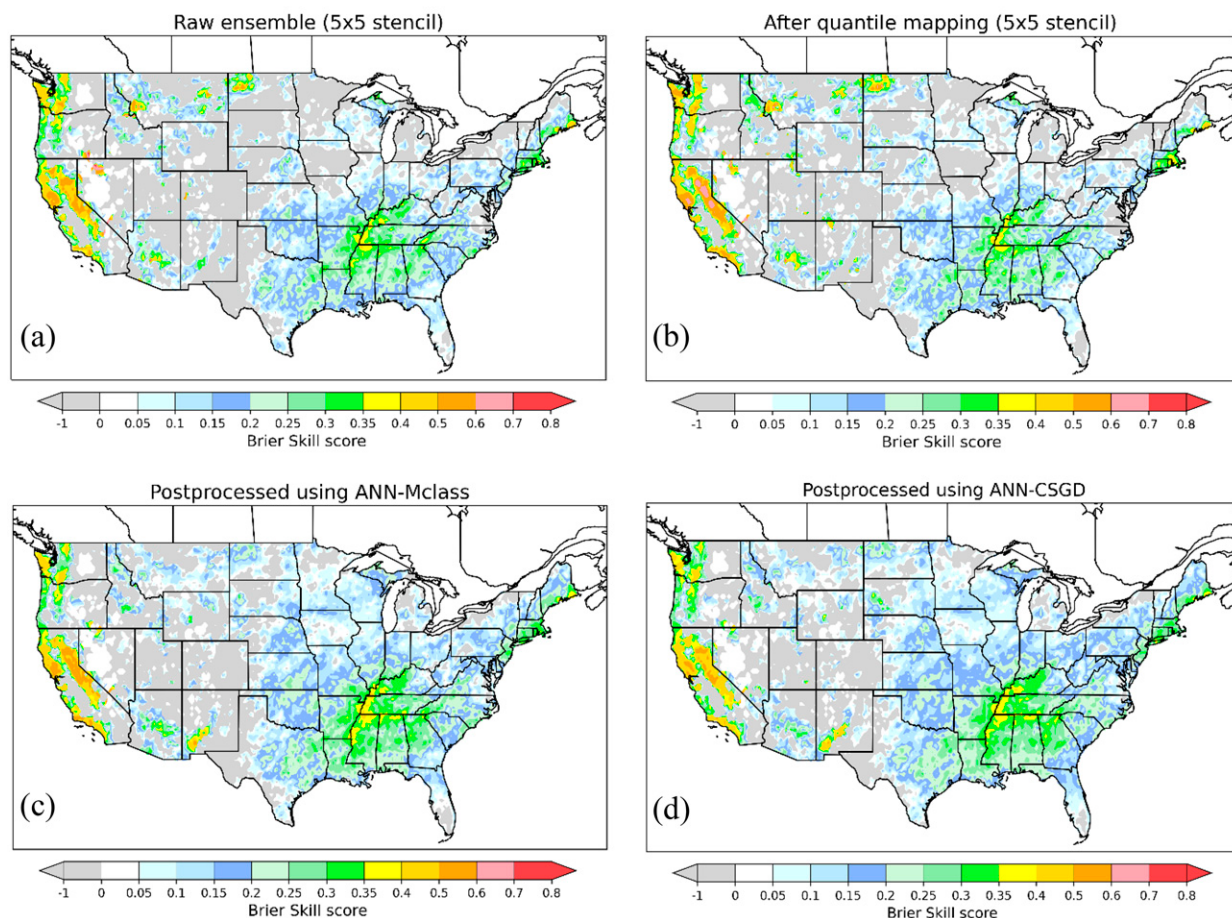
FIG. 3. Maps of Brier skill scores values of PoP forecasts aggregated over all days and for lead time from +48 to +72 h over the CONUS. Climatology is used as the reference. Forecasts are generated from (a) raw GEFS forecast using 5×5 stencil of grid points. (b) Quantile mapped forecast using 5×5 stencil of grid points. (c) ANN-Mclass and (d) ANN-CSGD.

of the Great Plains, whereas clusters of areas with high BSS values are found along the windward side of the Appalachians, and between the Cascades and the Rockies (Fig. 3a). QMAP drastically improves the skills with respect to the forecast of PoP for nearly all regions in CONUS (Fig. 3b), though negative remains in small areas over the southern tip of Texas. Both ANN schemes broadly outperform QMAP, and the outperformance is the most conspicuous to east of the Rockies (Figs. 3c,d). Between the two schemes, ANN-CSGD extends the skill of PQPF over the upper Midwest, the South, and the Southeast. Nonetheless, it is worth noting that the two ANN schemes slightly underperform QMAP along the Pacific coast where skills of raw GEFS are high.

Similar comparisons of BSS but for two higher thresholds, namely, 25 and 50 mm, are shown in Figs. 4 and 5, respectively. Notable features are summarized as follows. First, raw GEFS ensemble remains skillful relative to climatology along the Pacific coast, eastern portion of the southern Great Plains, much of the Midwest, South, Southeast, and along the mid-northeast Atlantic coast, but it underperforms climatology

over the upper Midwest, the Rockies, and the western portion of the Great Plains (Figs. 4a and 5a). Second, the performance of QMAP is mixed across the nation, in direct contrast to the wide skill improvements evident at the 0.254-mm threshold (Figs. 4b and 5b). The improvement is still appreciable over a few regions including the Sierra Nevada, but over other parts of the country, e.g., the South and Southeast, QMAP appears to degrade the skills of raw ensemble. Third, both ANN schemes are able to bring modest skill improvements for regions east of the Rockies. Between the two schemes, ANN-CSGD results in skill improvements over wider geographic regions, consistent with the earlier observation that it offers the best overall performance for CONUS (Figs. 2c,d).

Another subtle, yet important observation in Figs. 4 and 5 is that, despite the broad outperformance of ANN schemes, in a few regions the schemes conspicuously underperform the QMAP. Examples include the Sierra Nevada, southeast Texas, and the Carolinas. In each of these regions, raw GEFS ensemble offers good skills ($BSS > 0.35$); the skills are

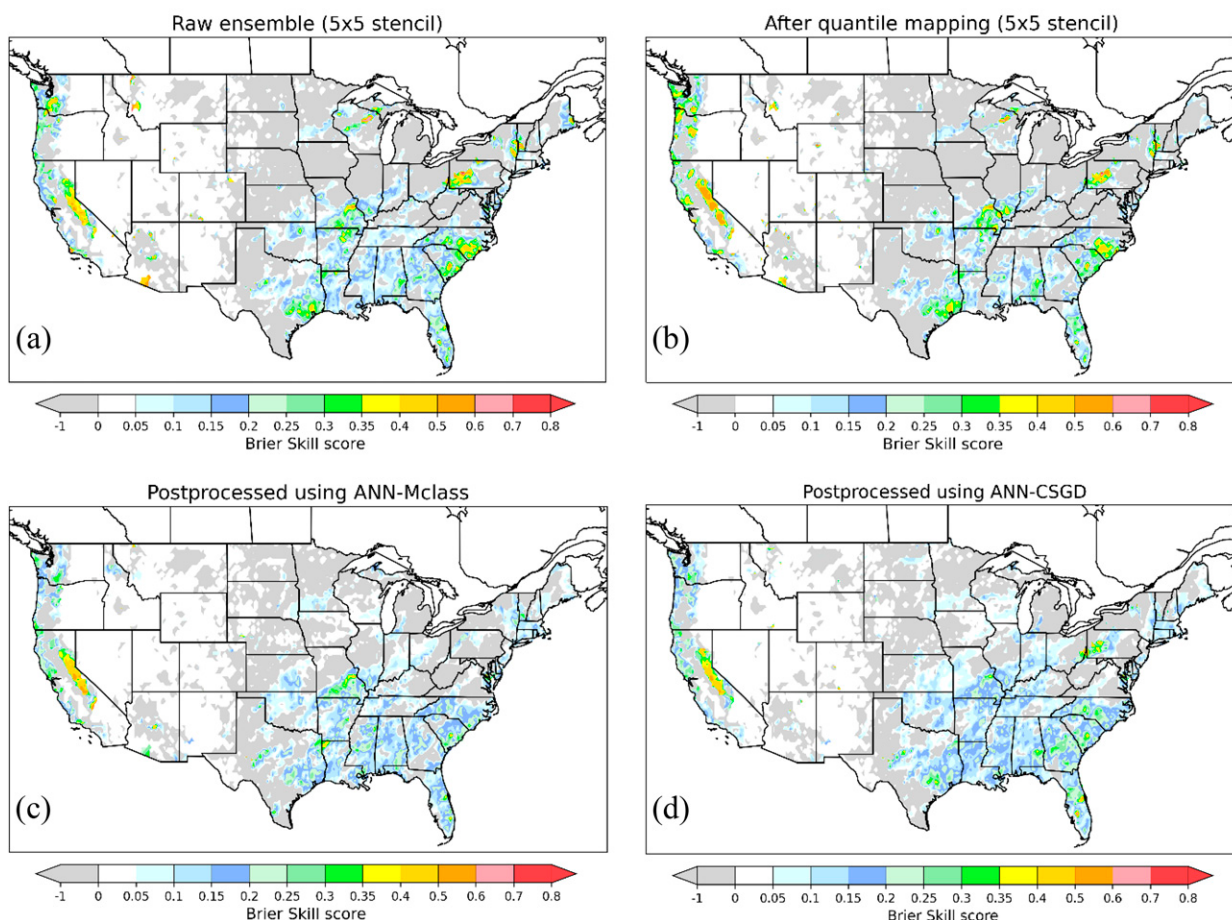
Brier skill score for lead time +48h to +72h, (> 25 mm/24h)FIG. 4. As in Fig. 3, but for the events > 25 mm.

retained in QMAPed results but are clearly degraded in the postprocessed PQPFs produced by applying the two ANN schemes. Therefore, it appears that the CONUS-wide skill gains associated with the application of the ANN schemes are mostly a result of improvements over domains where raw ensemble forecasts are marginally or modestly skillful. These improvements, however, are achieved at the expense of reduced skills for a few regions where raw ensemble forecasts are particularly accurate. As the former regions are far larger in size, the improvements seen therein tend to overshadow degradations observed over latter locations. It is heretofore unclear what contributes to QMAP's outperformance over the latter regions, but QMAP's reliance on a restricted set of supplemental locations most likely plays a critical role. This practice, as we surmise, may have limited overdispersion caused by the use of an unduly large amount of grid points with heterogeneous predictor–predictand relationships.

c. Reliability diagrams

Figures 6–8 show reliability diagrams computed for the raw GEFS ensemble and three sets of postprocessed PQPFs for

three thresholds, i.e., 0.254, 25, and 50 mm, again for 48–72-h lead time, with corresponding histograms of relative frequency of usage (sharpness histograms) superimposed. The reliability diagrams allow us to assess forecast attributes including reliability and resolution (see Brocker and Smith 2007; Wilks 2019). In constructing the reliability diagrams, for raw and quantile mapped forecasts based on center point only (11-member ensemble) we use 12 equally spaced probability categories within the ranges of $[0, 1]$ (see online supplemental material for center point results and results for additional thresholds). The reliability diagrams for ANN-based PQPFs, as well as raw and quantile mapped using 5×5 stencil, are computed by stratifying probabilities into 21 bins. We further perform decomposition of Brier score (Brier 1950) into reliability, resolution, and uncertainty as proposed by Murphy (1973) and assess the contribution of each component to PQPF skill in predicting specific events. Among these, resolution characterizes the forecast's ability to discriminate between different events and is identical to sharpness for perfectly reliable forecasts (see Jolliffe and Stephenson 2012). The resulting BSS, reliability (REL), and resolution (RES)

Brier skill score for lead time +48h to +72h, PoP (> 50 mm/24h)FIG. 5. As in Fig. 3, but for the events > 50 mm.

are superimposed on each reliability diagram. We choose to leave out uncertainty as it is independent of forecast source.

For the PoP forecasts (Fig. 6), it is clear that the raw ensemble is unreliable and tends to overforecast across the entire probability range (Fig. 6a). Quantile mapping improves both reliability and resolution of raw forecast, which in aggregate helps improve the forecast skill as measured in BSS (Fig. 6b). Nevertheless, it is evident that the QMAP scheme yields under/overforecast at low/high probability categories, and this feature is consistent with the findings of Hamill et al. (2017).

ANN-Mclass produces POPFs with further improved reliability and resolution but tends to consistently overforecast (Fig. 6c). ANN-CSGD POPFs outperform the rest in terms of reliability and resolution, and there is no conspicuous tendency to over/underforecast (Fig. 6d). That said, it is worth noting that ANN-CSGD POPFs exhibit the lowest sharpness as judged by the sharpness histogram. At the 25-mm threshold (Fig. 7), raw and quantile mapped forecasts perform comparably, though the latter is slightly more skillful due to its improvement in resolution (Figs. 7a,b). POPFs generated by ANN schemes on the other hand are much more reliable and

skillful than the former two forecast sources, but they are not as sharp: these POPFs feature lower frequencies for high probability categories (Figs. 7c,d).

As previously shown in the comparisons of BSS, the performance gap between raw ensemble and QMAP POPFs tends to narrow at higher thresholds. From Fig. 7b, it appears that QMAP slightly degrades the reliability but improves the resolution and, to a limited extent, the sharpness. This divergent outcome is rooted in the mismatch between forecast and analysis. As noted by Hamill and Whitaker (2006), the raw GEFS ensemble has a strong tendency to overpredict precipitation amounts for events with light-to-moderate intensity over for much of CONUS, and meanwhile it contains a small but substantial number of instances where it underpredicts precipitation amounts for events associated with larger accumulations. For larger forecast amounts, quantile adjustment tends to increase the forecast amounts. But due to the forecast–analysis mismatch, this increase serves to inflate the amounts for a vast majority of events where analyzed accumulations are actually lower than the forecasted, thus amplifying the wet bias that is already existent in the raw forecast (Fig. 7b). At the

Reliability diagrams (PoP) for lead time +48h to +72h

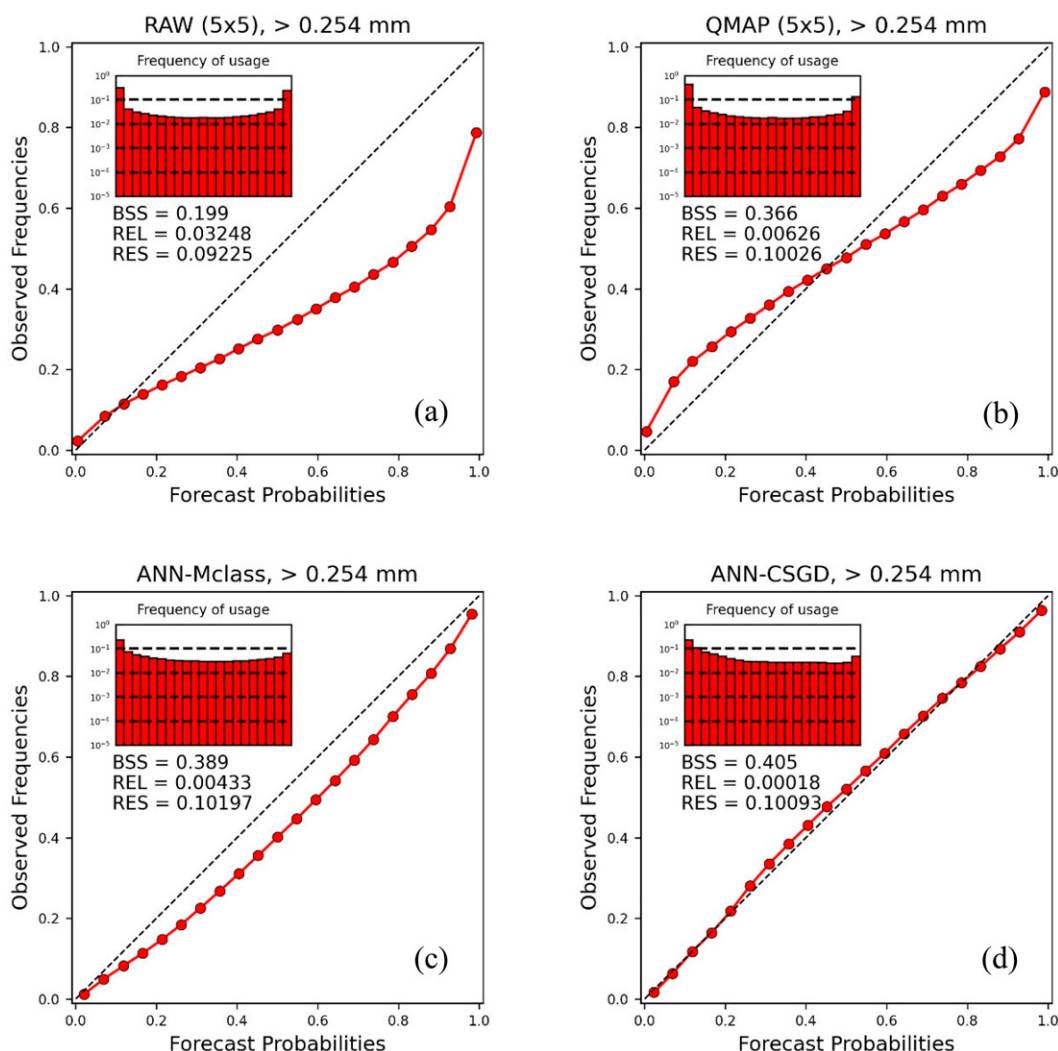


FIG. 6. Reliability diagrams (PoP) for lead time from +48 to +72 h over the CONUS. Histograms show relative frequency of forecast issuance for each of 21 forecast probability bins in log10 scale. BSSs and Brier score decompositions are shown in each panel. (a) Raw forecast using 5×5 stencil of grid points. (b) Quantile mapped forecast using 5×5 stencil of grid points. (c) ANN-Mclass and (d) ANN-CSGD.

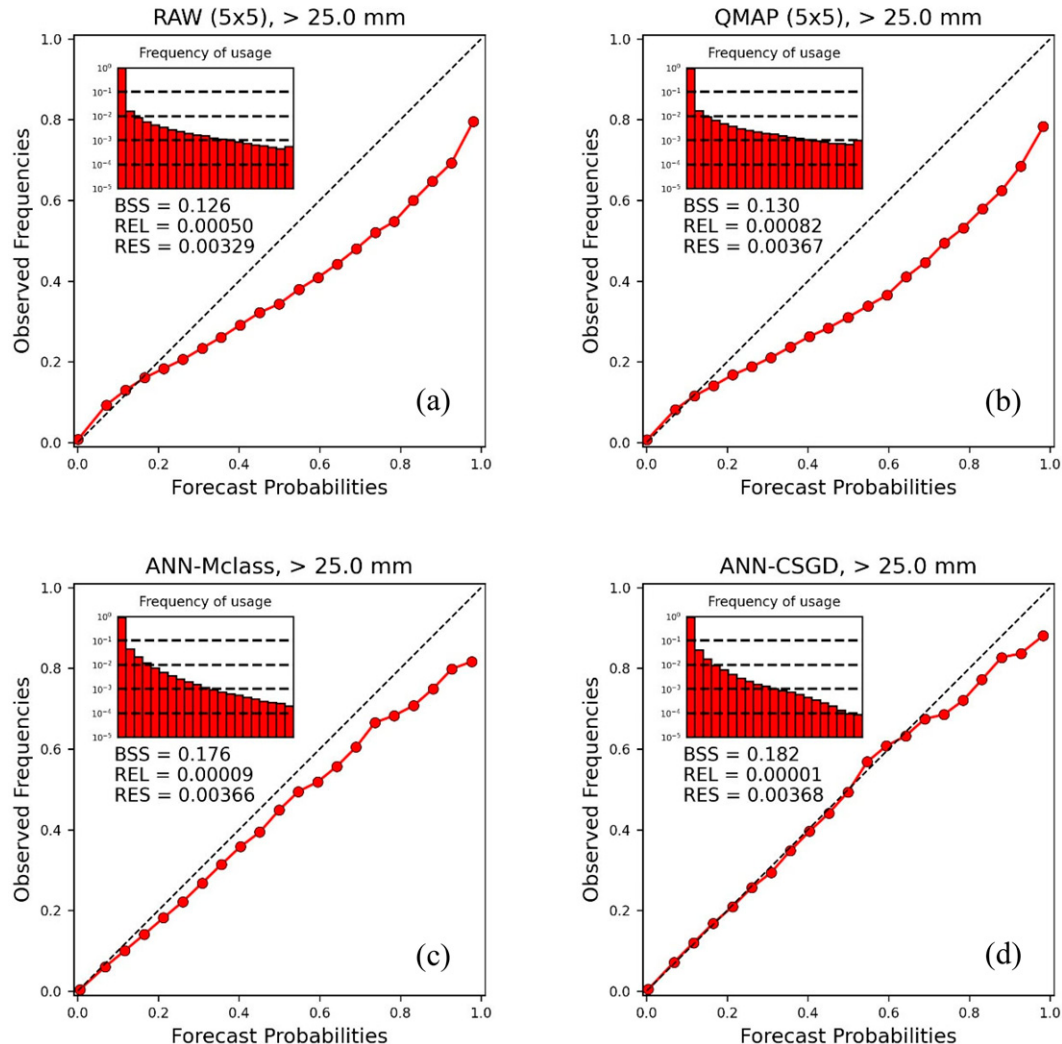
highest threshold, i.e., 50 mm (Fig. 8), the relative performance of three schemes broadly echoes those at the 25-mm threshold but a few distinctions are apparent.

First, QMAP scheme tends to more severely degrade the reliability that it does at the 25-mm threshold, resulting in a conspicuous overforecast across probability categories, though there is a sign that it improves the sharpness (Fig. 8b). Second, while both ANN schemes (Figs. 8c,d) again yield PQPFs with improved reliability relative to the raw ensemble, but the margin of improvements narrows somewhat, and at the two highest probability categories both schemes underperform the QMAP by featuring more severe positive bias (overforecast). Between the two ANN schemes, ANN-CSGD

clearly outperforms in terms of calibration but at the cost of reduced sharpness.

d. Case study

To further illustrate the skills of PQPFs and shed light on their geographic and precipitation-regime dependence, we construct forecast guidance from raw ensemble and postprocessed PQPFs in a way that mimics the NBM operation and compared these against areas where corresponding thresholds are exceeded in the analysis. Such a practice is widely adopted in NWS forecast verifications (see, e.g., WPC 2019). This verification exercise focuses on a one-day window ending at 0000 UTC 4 January 2017. As shown in Fig. 9, this window was so

Reliability diagrams (> 25 mm) for lead time +48h to +72hFIG. 7. As in Fig. 6, but for events > 25 mm.

chosen that there were several large precipitation clusters simultaneously present over the west coast (Northern California), between the Midwest and Mid-Atlantic coast, and over the southeast (Alabama, Georgia, South Carolina, and part of northern Florida). Maximum 1-day accumulations for these clusters all exceeded 50 mm.

We computed exceedance probabilities from day 2 (from +24 to +48 h) GEFS ensemble forecasts for the valid date ending at 0000 UTC 4 January 2017. Figure 10 displays maps of PoP (> 0.254 mm) computed based on raw ensemble (with 5×5 stencil) and derived from three suites of postprocessed POPFs, namely, those produced by QMAP (with 5×5 stencil), ANN-Mclass, and ANN-CSGD.

Average values of Brier score (averaged over all grid points for this day) are overlaid on each map to gauge the CONUS-

wide performance of each product. The following features are evident. First, PoP from raw ensemble is broadly higher (close to 1) and there is a conspicuous lack of spatial details (Fig. 10a). Second, in many parts of the CONUS, the PoP is close to one despite a lack of precipitation in CCPA, consistent with the severe overforecast seen in earlier reliability diagrams for the raw ensemble (Fig. 6a). Quantile mapping drastically improves locational precision by reducing the areas where raw ensemble produces high PoP (Fig. 10b). This reduction is particularly noticeable over the Intermountain West to the east of the Sierra Nevada, where gradients in PoP values emerge after quantile mapping. The BS is much reduced (improved), corresponding to this reduction in areas with overforecasted PoP. The two ANN schemes produce further improvements for the Intermountain West, by suppressing the PoP values outside the areas where

Reliability diagrams (> 50 mm) for lead time +48h to +72h

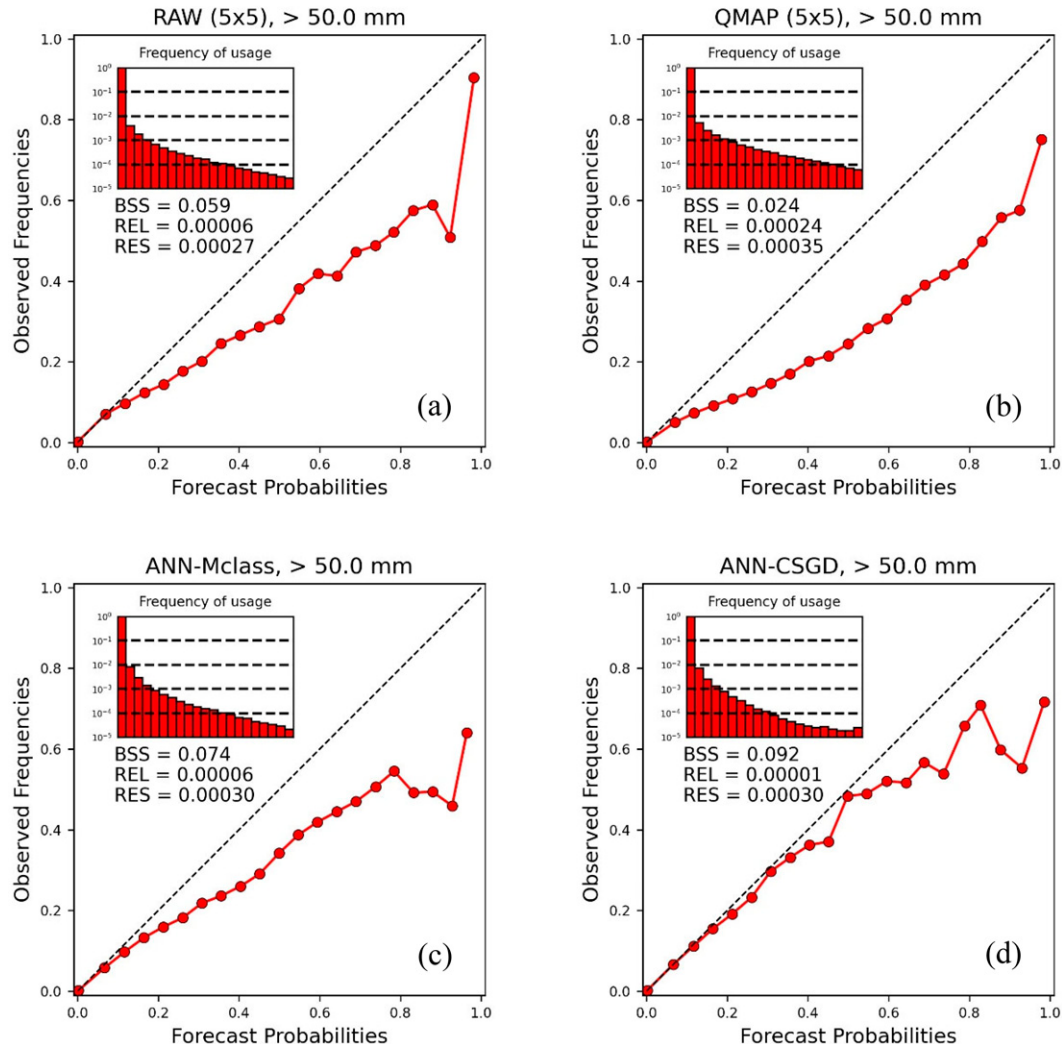


FIG. 8. As in Fig. 6, but for events > 50 mm.

CCPA indicates wet conditions (Figs. 10c,d). The postprocessed PoPs from the two schemes feature much lower overforecast errors over this region, thus allowing the actual precipitation clusters to be more precisely defined. Nonetheless, application of these schemes leads to sharp expansions of areas with low, but positive PoP (<0.2) to regions where no precipitation was observed (e.g., Wyoming). Between the two ANN schemes, ANN-CSGD produces the most skillful PoP that exhibits the least spatial mismatch with the analysis, and its PoP features the lowest BS among the four sets of products. On the other hand, however, it tends to produce wider expansion of PoP, and suppress the high probability values in regions where precipitation was observed (Fig. 10d). One possible explanation for this apparent trade-off lies in ANN schemes' inclusion of training samples over wider areas where the precipitation amounts are dissimilar to those near the target. While this practice improves the overall

skills for CONUS, it introduces diffusion in areal coverage and reduces the sharpness. This phenomenon is analogous to the dilution effect noted in regression literature (Fuller 1987; Hughes 1993; Frost and Thompson 2000; Jozaghi et al. 2021).

At the threshold of 25 mm (P25, Fig. 11), raw ensemble fails to detect the precipitation clusters in northern California while all three schemes help recover these. Over this region, P25 produced by quantile mapping exhibits the highest locational precision and sharpness, a feature consistent with the earlier observation shown in Fig. 4. By contrast, over the eastern part of the country, quantile mapping broadly degrades the skills—it produces excessively high P25 over broader areas where accumulations per CCPA are much below 25 mm. This introduces additional wet biases and contributes to a reduction in overall reliability relative to raw as shown in reliability diagrams (Fig. 7b). Both ANN schemes improve the overall accuracy of

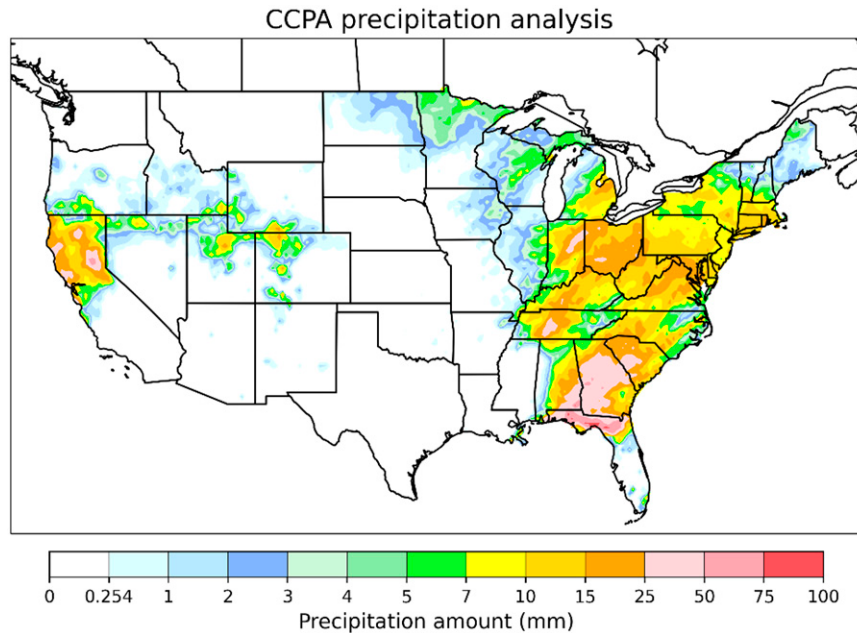


FIG. 9. CCPA precipitation analysis for 24-h accumulated data ending at 0000 UTC 4 Jan 2017.

guidance for CONUS and over major precipitation clusters (Figs. 11c,d). Over northern California, both yield higher P25 that overlap with the observed clusters, but with lower spatial precision. ANN-CSGD, for example, populates the entire Northern California with positive P25, grossly exaggerating the areal coverage of rainfall risks. ANN-Mclass fares somewhat better with more subdued areal coverage bias. Along the eastern United States, the most notable feature is that both ANN schemes perform well in capturing the rainfall risks over the cluster that encompasses parts of Georgia, Alabama, Tennessee, Florida, and South Carolina. The relative performance of the two techniques is mixed. ANN-Mclass excels by creating higher P25 within the cluster, yet it in the meantime inflates the P25 outside the cluster. By comparison, the P25 produced by ANN-CSGD is nearly uniformly lower over the region, either inside or outside the cluster. The BS of the ANN-based QPFs is broadly comparable at this threshold and is slightly better than that for QMAP.

Figure 12 shows the relative performance of the three schemes at the highest threshold, i.e., 50 mm (P50). Heavy rainfall exceeding the threshold is concentrated over smaller clusters along the Sierra Nevada, in central Georgia, and over the northern Florida Panhandle. Raw ensemble is apparently unable to capture any of these clusters. All three schemes create areas with positive P50. Among these, QMAP and ANN-CSGD perform comparably by creating small P50 values that marginally overlap with the clusters in central Georgia and Florida Panhandle. By comparison, ANN-Mclass creates positive P50 over a wider area over the southeast with more substantial overlap with all clusters in the region. This improved detection, however, is offset by the false coverage elsewhere. The BS values for all four products are nearly identical, most likely a result of the limited sample size for computing the metric. The ability of the two ANN schemes in mitigating the

overforecast as seen in Fig. 8 is not confirmed by the spatial verification for possibly the same reason.

4. Summary and conclusions

This study marks one of the first attempts to explore the use of unified ANN mechanisms for postprocessing medium-range (1–8 days) ensemble QPFs over a large domain using *short training datasets*. Chosen for the experimentation are two recent ANN postprocessing schemes. The first approach (ANN-Mclass) creates forecast probabilities for discrete precipitation categories, and then interpolates/extrapolates these probabilities to construct full CDF (Scheuerer et al. 2020). The second one is the ANN-CSGD, a newly developed, hybrid ANN-parametric postprocessing scheme that uses ANN to relate set of predictors to parameters of predictive censored, shifted gamma distribution (Ghazvinian et al. 2021). Both networks have rather simple structure (dense) and shared similar predictors, yet they differ in the specification of predictive distribution and loss function. In fact, these two schemes were so chosen to identify potential merits of retaining the parametric form of the predictive distribution in the prediction of rare, heavy rainfall events.

To assess the performance of the ANN schemes, we designed hindcast experiments to postprocess 24-h accumulated GEFS reforecast data using a rolling training scheme and with previous 60 days' forecast and analyzed data. As the benchmark statistical postprocessing technique we implemented QMAP stencil method (Hamill et al. 2017) that is used to produce probabilistic guidance that populates the National Digital Forecast Database (NDFD; Glahn and Ruth 2003). Note that a key difference between the QMAP approach and the ANN schemes concerns the mechanism of augmenting the spatial sampling domain to compensate for the limited time window. QMAP does

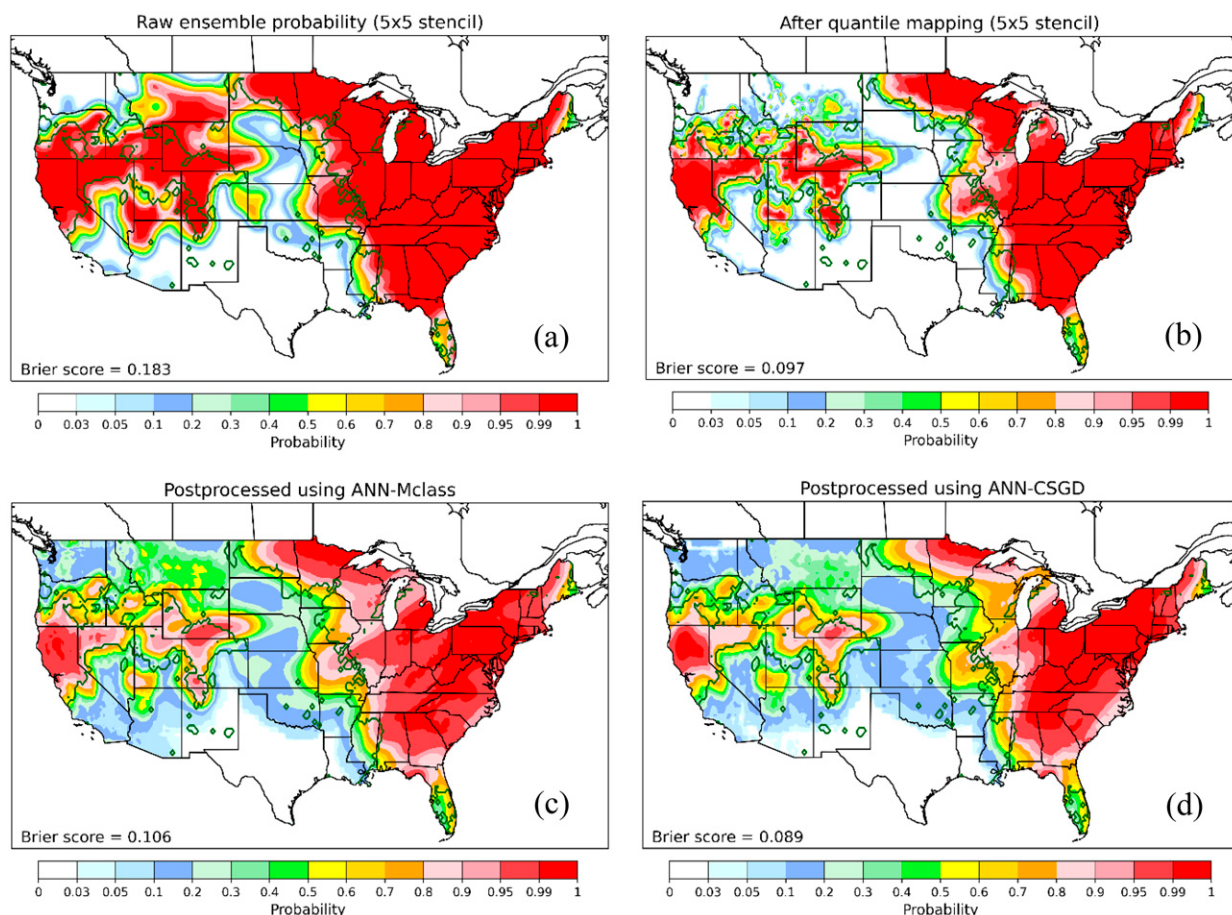


FIG. 10. POP forecast guidance for lead time from +24 to +48 h over CONUS for valid date ending at 0000 UTC 4 Jan 2017. (a) Raw ensemble forecast using 5×5 stencil, (b) quantile mapped forecasts using 5×5 stencil, (c) forecasts generated using ANN-Mclass, and (d) forecasts generated using ANN-CSGD. Areas inside dark green contours show that event > 0.254 mm has been observed by CCPA.

so by incorporating so-called supplemental locations, i.e., locations that share similar elevation and topographic facets, and, presumably, may share similar precipitation climatology. By contrast, the ANN schemes leverage data at all grid points within the domain and infusing geographical information including latitude and longitude as ancillary predictors.

When aggregated over the entire CONUS, as the results demonstrate, POPFs from ANN schemes broadly outperform the raw and quantile mapped forecasts in terms of reliability and overall skill. This outperformance is seen for a range of thresholds and across all lead times, but it tends to be more pronounced at higher precipitation thresholds (e.g., 25 and 50 mm)—thresholds that are closely relevant to flood forecasting and real time reservoir management. It was also found that while quantile mapping broadly improves upon raw forecasts in predicting PoP, its performance declines at higher thresholds. At the highest threshold (50 mm day^{-1}), the POPFs from QMAP underperforms the raw ensemble.

The two ANN schemes perform comparably at low-to-middle thresholds in terms of calibration, but the performance differentials widen at higher thresholds with ANN-CSGD conspicuously outperforming. A major weakness of ANN-Mclass is the lack of reliability of its POPFs at the highest

threshold (50 mm day^{-1}): when compared to ANN-CSGD, it tends to produce more severe overforecasts and is broadly incapable of enhancing the skills of raw ensemble at this threshold. This underperformance of ANN-Mclass is potentially related to the inadequate number of output categories implemented in the study. Note that the selection of optimal number of output categories is not a straightforward task—as high observed precipitation values are much less frequent than lighter precipitation values, empirical quantiles cannot be easily extended to large amounts without incurring substantial uncertainties. In this regard, ANN-CSGD's explicit use of a parametric distribution proves advantageous as it offers a more consistent way of estimating higher forecast quantiles.

Our validation experiments also reveal a distinctive geographic dependence of the relative performance of different schemes. QMAP, while broadly underperforming the ANN schemes for the entire CONUS, outperforms the latter competitors along the West Coast and over the Sierra Nevada. Its overall underperformance is mostly a result of its inability to produce skill gains for much of the central and eastern United States. Closer examinations suggest that variations in the skill of the raw ensemble for different precipitation regimes, along

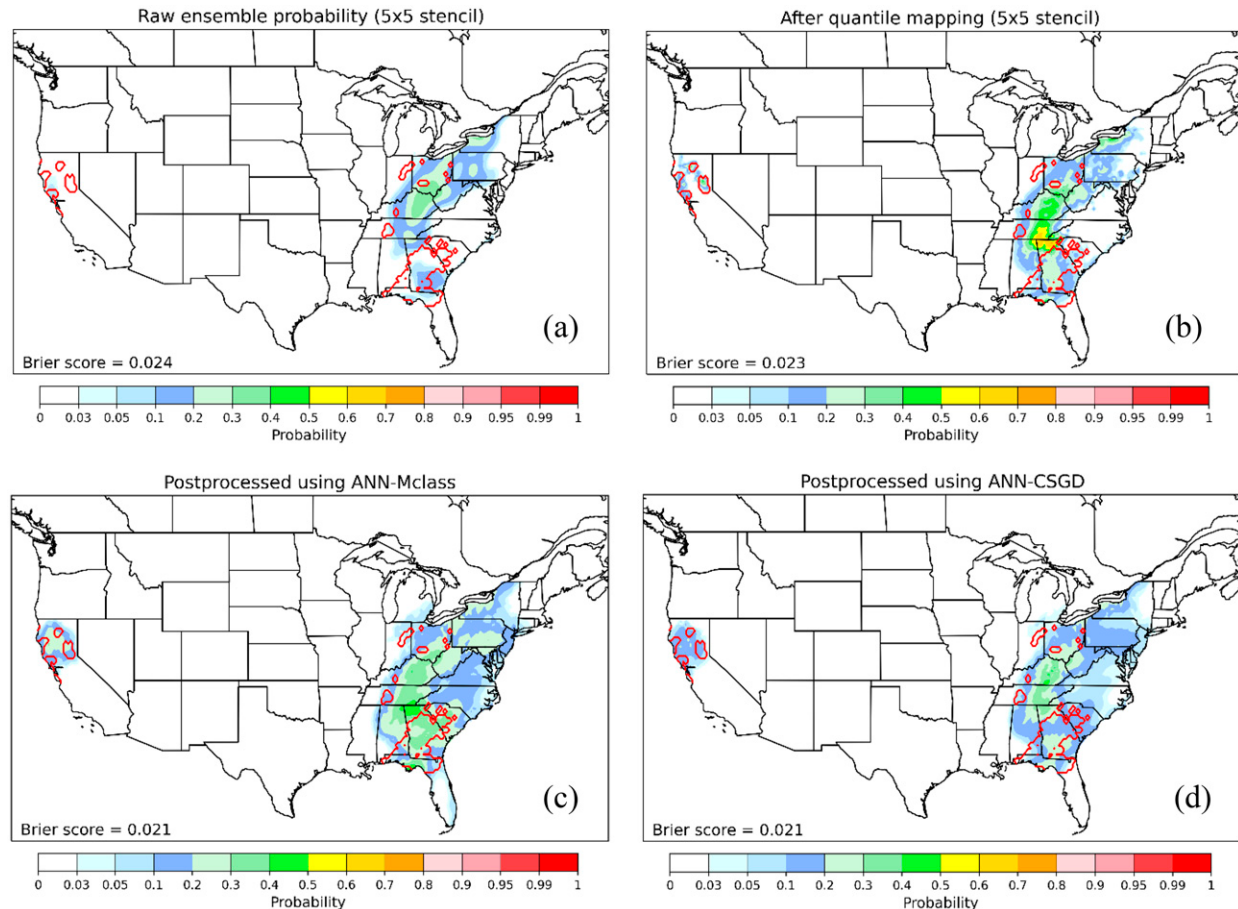


FIG. 11. As in Fig. 10, but for >25 mm. Areas inside red contours show that event > 25 mm has been observed by CCPA.

with the spatial variability of rainfall brought by these regimes, may have played pivotal roles in shaping the geographic disparity. To elaborate, over the Pacific coast–Sierra Nevada, landfalling atmospheric river events dominate large precipitation amounts over the region. The GEFS ensemble exhibits good skills in predicting the occurrence of these events as well as the associated geographic distribution of precipitation. QMAP’s use of limited, prescribed supplemental locations prove effective in correcting forecast biases, whereas the ANN schemes’ simultaneous use of samples across locations may have overexpanded the training sample and thereby impaired the robustness of predictor–predictand relationship derived therefrom. By contrast, heavy precipitation over the central and eastern United States can arise from a mix of organized convection, frontal systems, and tropical cyclones, and predictability of these systems varies both by location and season. The overall skills of GEFS ensemble are low over this region, and spatial displacement errors are a major contributor to the lack of skills. For these locations, the ability of ANN schemes in adaptively incorporating forecast–observation pairs over broader areas for training more effectively address the displacement errors. Another potential factor underlying the contrasting performance of QMAP, as the authors postulate, is the degree of similarity

among supplemental locations. It is possible that forecast–observation relationships are broadly dissimilar among supplemental locations over central and eastern United States as elevations and facets play lesser roles in modulating precipitation climatology. Other data augmentation techniques, such as semilocal models described in Lerch and Baran (2017), may improve quantile mapping performance. It is also worth pointing out that the gains in calibration for the higher thresholds as achieved by the ANN schemes often come at the expense of subdued forecast sharpness—the exceedance probability in regions where precipitation was observed is often lower in the postprocessed guidance produced by these schemes, a feature reminiscent of the findings of Herman and Schumacher (2018). These issues warrant further, more thorough investigations to confirm and illuminate.

As ANN techniques are evolving rapidly, there are many emerging opportunities for further enhancing the ANN postprocessing schemes illustrated in the study. Future research will be directed toward identifying and integrating mechanisms that will allow for (i) more effective use of geographic information in the networks. Location embedding can be used to project discrete pairs of each latitude and longitude values onto a continuous, larger vector of latent inputs using IDs specific for each CCPA grid. Embedding permit the model to optimize grid IDs representations by the training

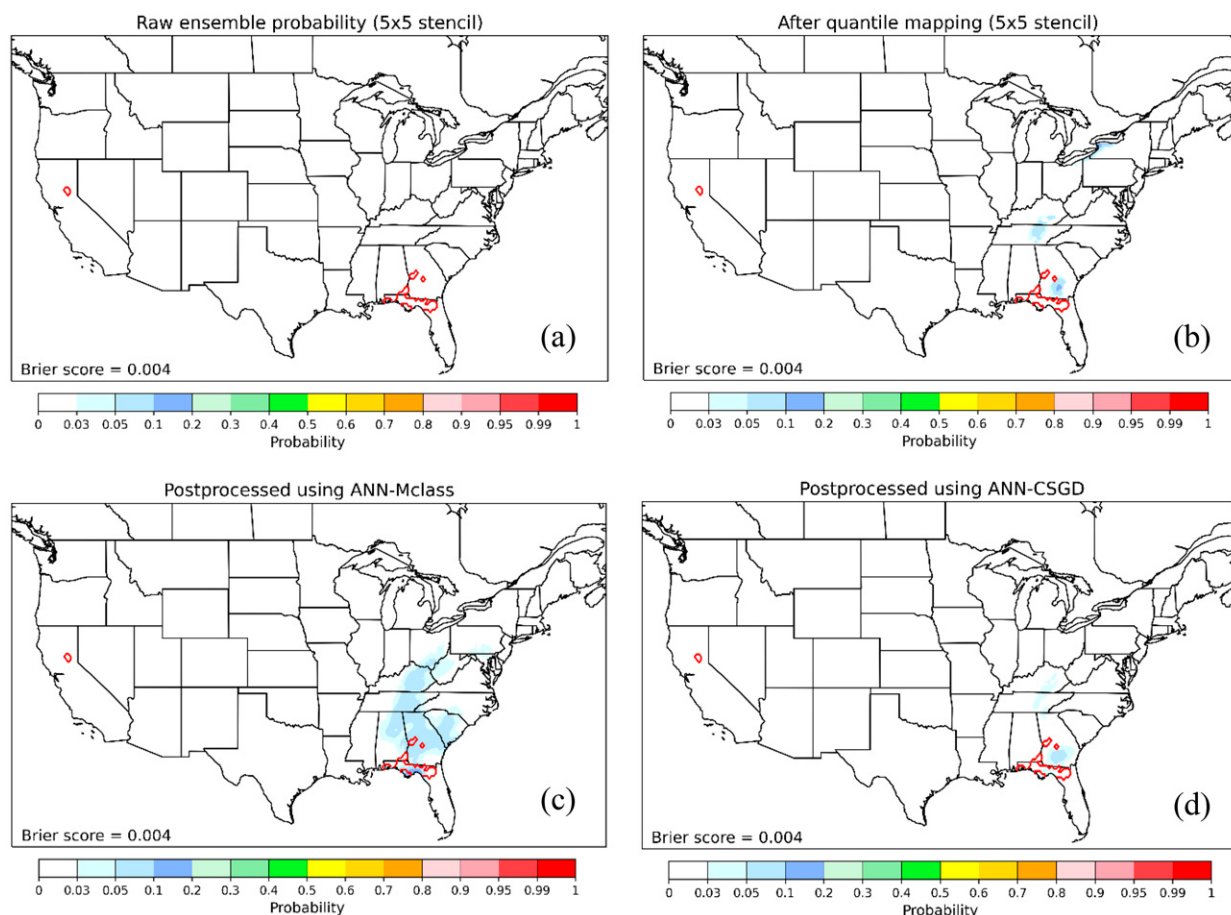


FIG. 12. As in Fig. 10, but for >50 mm. Areas inside red contours show that event > 50 mm has been observed by CCPA.

process and potentially help better capture local characteristics as shown in past postprocessing studies (see Schulz and Lerch 2021; Chapman et al. 2022). Incorporating auxiliary predictors such as dot product of moisture advection with terrain gradient, total column precipitable water and convective available potential energy (CAPE) might be another way of getting more terrain-related detail; (ii) more efficient modeling of complex-arbitrary nonlinear predictor–predictand relationships such as use one dimensional convolution or attention layers (Collobert et al. 2011; Vaswani et al. 2017; Devlin et al. 2018) on top of an embedding layer to better capture predictor interactions with each other and with spatial features, and (iii) more robust training of networks to avoiding overfitting. This work used a rather basic but popular and effective regularization technique to stop training, which is based on validation dataset loss (Goodfellow et al. 2016). It is possible that additional gains can be realized by simply increasing the validation window with lead time and introducing additional regularization parameters.

Acknowledgments. The authors would like to acknowledge financial support for the first and second authors over the years provided by the faculty startup package for Dr. Yu Zhang from UT Arlington, NOAA Grant NA18OAR4590370-01, NSF

Grant 1909367, and Texas Water Development Board Contract 1800012276. The work benefits from input from a large number of individuals within the National Weather Service, including Jeff Craven, David Rudack, and Eric Engle of the National Blended Model team, Kris Lander at the West Gulf River Forecast Center, Bruce Veenhuis at the Weather Prediction Center, and John J. Brost at the Operational Proving Ground. We would also like to thank Kevin He from the California Department of Water Resources for stimulating discussions that helped shape the work, and to members of the Unified Forecast System Steering Committee for critiques and suggestions.

Data availability statement. The analyses–forecasts and output dataset on which the results of this work are based are too large to be publicly archived with available resources. The codes to reproduce results can be made available based on individual requests and only for research purposes.

APPENDIX

Implementation Details

We used Python (Python Software Foundation 2018), R (R Core Team 2017), and FORTRAN in this project.

Specifically, R was used for initial data processing. We implemented our ANN codes in Python using Google's platform, Tensorflow (Abadi et al. 2016) and Keras API (Chollet et al. 2015). For quantile mapping a research version was implemented using Python. FORTRAN routines to generate supplemental locations were provided by Dr. Tom Hamill from NOAA PSL and were tailored to our setting. Other computations (verification, graphics, etc.) were performed with Python.

REFERENCES

- Abadi, M., and Coauthors, 2016: Tensorflow: A system for large-scale machine learning. *Proc. USENIX 12th Symp. on Operating Systems Design and Implementation*, Savannah, GA, Advanced Computing Systems Association, 265–283, <https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf>.
- Baran, S., and D. Nemoda, 2016: Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, **27**, 280–292, <https://doi.org/10.1002/env.2391>.
- , and Á. Baran, 2021: Calibration of wind speed ensemble forecasts for power generation. *Idojaras*, **125**, 609–624, <https://doi.org/10.28974/idojaras.2021.4.4>.
- , and S. Lerch, 2018: Combining predictive distributions for statistical post-processing of ensemble forecasts. *Int. J. Forecast.*, **34**, 477–496, <https://doi.org/10.1016/j.ijforecast.2018.01.005>.
- Bremnes, J. B., 2020: Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Mon. Wea. Rev.*, **148**, 403–414, <https://doi.org/10.1175/MWR-D-19-0227.1>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, <https://doi.org/10.1175/WAF993.1>.
- Brown, J. D., L. Wu, M. He, S. Regonda, H. Lee, and D. J. Seo, 2014: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. *Hydrol.*, **519**, 2869–2889, <https://doi.org/10.1016/j.jhydrol.2014.05.028>.
- Chapman, W. E., L. Delle Monache, S. Alessandrini, A. C. Subramanian, F. M. Ralph, S. Xie, S. Lerch, and N. Hayatbini, 2022: Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Mon. Wea. Rev.*, **150**, 215–234, <https://doi.org/10.1175/MWR-D-21-0106.1>.
- Chollet, F., and Coauthors, 2015: Keras: The Python Deep Learning library. Accessed 2020, <https://keras.io>.
- Cloke, H. I., and F. Pappenberger, 2009: Ensemble flood forecast: A review. *J. Hydrol.*, **375**, 613–626, <https://doi.org/10.1016/j.jhydrol.2009.06.005>.
- Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, 2011: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, **12**, 2493–2537, <https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf>.
- Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, <https://doi.org/10.1002/joc.1688>.
- Darbandsari, P., and P. Coulibaly, 2022: Assessing entropy-based Bayesian model averaging method for probabilistic precipitation forecasting. *J. Hydrometeorol.*, **23**, 421–440, <https://doi.org/10.1175/JHM-D-21-0086.1>.
- Devlin, J., M. W. Chang, K. Lee, and K. Toutanova, 2018: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv, 1810.04805, <https://arxiv.org/abs/1810.04805>.
- Frost, C., and S. G. Thompson, 2000: Correcting for regression dilution bias: Comparison of methods for a single predictor variable. *J. Roy. Stat. Soc.*, **163**, 173–189, <https://doi.org/10.1111/1467-985X.00164>.
- Fuller, W. A., 1987: *Measurement Error Models*. Wiley, 440 pp.
- Ghazvinian, M., Y. Zhang, and D. J. Seo, 2020: A nonhomogeneous regression-based statistical postprocessing scheme for generating probabilistic quantitative precipitation forecast. *J. Hydrometeorol.*, **21**, 2275–2291, <https://doi.org/10.1175/JHM-D-20-0019.1>.
- , —, D.-J. Seo, M. He, and N. Fernando, 2021: A novel hybrid artificial neural network - Parametric scheme for postprocessing medium-range precipitation forecasts. *Adv. Water Resour.*, **151**, 103907, <https://doi.org/10.1016/j.advwatres.2021.103907>.
- Glahn, H. R., and D. P. Ruth, 2003: The new digital forecast database of the national weather service. *Bull. Amer. Meteor. Soc.*, **84**, 195–202, <https://doi.org/10.1175/BAMS-84-2-195>.
- Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. MIT Press, 775 pp.
- Hamill, T. M., 2018: Practical aspects of statistical postprocessing. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. W. Messner, Eds., Elsevier, 187–217.
- , 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>.
- , J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: 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](https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2).
- , G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, 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 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>.
- Herman, G. R., and R. S. Schumacher, 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>.
- Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeorol.*, **15**, 2542–2557, <https://doi.org/10.1175/JHM-D-11-0140.1>.

- Hughes, M. D., 1993: Regression dilution in the proportional hazards model. *Biometrics*, **49**, 1056–1066, <https://doi.org/10.2307/2532247>.
- Ioffe, S., and C. Szegedy, 2015: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proc. 32nd Int. Conf. Machine Learning*, Vol. 37, Lille, France, JMLR, 448–456. <http://proceedings.mlr.press/v37/ioffe15.pdf>.
- Jolliffe, I. T., and D. B. Stephenson, Eds., 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. John Wiley & Sons, 292 pp. <https://doi.org/10.1002/9781119960003>.
- Jozaghi, A., H. Shen, M. Ghazvinian, D.-J. Seo, Y. Zhang, E. Welles, and S. Reed, 2021: Multi-model streamflow prediction using conditional bias-penalized multiple linear regression. *Stochastic Environ. Res. Risk Assess.*, **35**, 2355–2373, <https://doi.org/10.1007/s00477-021-02048-3>.
- Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. arXiv, 1412.6980, <https://arxiv.org/abs/1412.6980>.
- Krzysztofowicz, R., 2008: Bayesian processor of ensemble: Concept and development. *Proc. 19th Conf. Probability and Statistics*, New Orleans, LA, Amer. Meteor. Soc., 4.5, https://ams.confex.com/ams/88Annual/techprogram/paper_131722.htm.
- Lerch, S. and S. Baran, 2017: Similarity-based semilocal estimation of post-processing models. *J. Roy. Stat. Soc.*, **66**, 29–51, <https://doi.org/10.1111/rssc.12153>.
- Li, W., B. Pan, J. Xia, and Q. Duane, 2022: Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *J. Hydrol.*, **605**, 127301, <https://doi.org/10.1016/j.jhydrol.2021.127301>.
- Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1096, <https://doi.org/10.1287/mnsc.22.10.1087>.
- Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, [https://doi.org/10.1175/1520-0450\(1973\)012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2).
- Pappenberger, F., and R. Buizza, 2009: The skill of ECMWF precipitation and temperature predictions in the Danube basin as forcings of hydrological models. *Wea. Forecasting*, **24**, 749–766, <https://doi.org/10.1175/2008WAF2222120.1>.
- Python Software Foundation, 2018: Python Language Reference, version 3.7. <http://www.python.org>.
- R Core Team, 2017: R: A language and environment for statistical computing. R Foundation for Statistical Computing, <https://www.R-project.org/>.
- Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, <https://doi.org/10.1175/MWR-D-18-0187.1>.
- Reggiani, P., and O. Boyko, 2019: A Bayesian processor of uncertainty for precipitation forecasting using multiple predictors and censoring. *Mon. Wea. Rev.*, **147**, 4367–4387, <https://doi.org/10.1175/MWR-D-19-0066.1>.
- Robertson, D. E., D. L. Shrestha, and Q. J. Wang, 2013: Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrol. Earth Syst. Sci.*, **17**, 3587–3603, <https://doi.org/10.5194/hess-17-3587-2013>.
- Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.
- , —, B. Whitin, M. He, and A. Henkel, 2017: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.*, **53**, 3029–3046, <https://doi.org/10.1002/2016WR020133>.
- , M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic sub-seasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, <https://doi.org/10.1175/MWR-D-20-0096.1>.
- Schulz, B., and S. Lerch, 2021: Machine learning methods for post-processing ensemble forecasts of wind gusts: A systematic comparison. arXiv, 2106.09512, <https://arxiv.org/abs/2106.09512>.
- 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>.
- Taillardat, M., A. Fougères, P. Naveau, and O. Mestre, 2019: Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Wea. Forecasting*, **34**, 617–634, <https://doi.org/10.1175/WAF-D-18-0149.1>.
- Valdez, E. S., F. Anctil, and M.-H. Ramos, 2022: Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems. *Hydrol. Earth Syst. Sci.*, **26**, 197–220, <https://doi.org/10.5194/hess-26-197-2022>.
- Vannitsem, S., 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>.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, 2017: Attention is all you need. *31st Conf. on Advances in Neural Information Processing Systems*, Long Beach, CA, Neural Information Processing Systems, 5998–6008. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- Veldkamp, S., K. Whan, S. Dirksen, and M. Schmeits, 2021: Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Mon. Wea. Rev.*, **149**, 1141–1152, <https://doi.org/10.1175/MWR-D-20-0219.1>.
- Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, <https://doi.org/10.1002/met.134>.
- , 2019: *Statistical Methods in the Atmospheric Sciences*. 4th ed. Elsevier, 840 pp., <https://doi.org/10.1016/C2017-0-03921-6>.
- WPC, 2019: 2019 Flash Flood and Intense Rainfall Experiment: Findings and results. NCEP/Weather Prediction Center, 123 pp., https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2019_FFaIR.pdf.
- , 2020: 2020 Flash Flood and Intense Rainfall Experiment: Findings and Results. NCEP/Weather Prediction Center, 99 pp., https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2020_FFaIR_Experiment_Nov13.pdf.
- Wu, L., D. J. Seo, J. Demargne, J. Brown, S. Cong, and J. Schaake, 2011: Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *J. Hydrol.*, **399**, 281–298, <https://doi.org/10.1016/j.jhydrol.2011.01.013>.
- Zhang, Y., L. Wu, M. Scheuerer, J. Schaake, and C. Kongoli, 2017: Comparison of probabilistic quantitative precipitation forecasts from two postprocessing mechanisms. *J. Hydrometeorol.*, **18**, 2873–2891, <https://doi.org/10.1175/JHM-D-16-0293.1>.