

Journal of Hydrometeorology

Improving Probabilistic Quantitative Precipitation Forecasts Using Short Training Data through Deep Learning --Manuscript Draft--

Manuscript Number:	JHM-D-22-0021
Full Title:	Improving Probabilistic Quantitative Precipitation Forecasts Using Short Training Data through Deep Learning
Article Type:	Article
Corresponding Author:	Mohammadvaghef Ghazvinian Center for Western Weather and Water Extremes UNITED STATES
Corresponding Author's Institution:	Center for Western Weather and Water Extremes
First Author:	Mohammadvaghef Ghazvinian
Order of Authors:	Mohammadvaghef Ghazvinian Yu Zhang Thomas M. Hamill Dong-Jun Seo Nelun Fernando
Abstract:	<p>Conventional statistical postprocessing techniques offer limited ability to improve the skills of probabilistic guidance for heavy precipitation. This paper introduces two deep learning (DL) based, geographically aware, and computationally efficient postprocessing schemes namely the Artificial Neural Network – 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 forecasts (PQPFs) over the contiguous United States (CONUS) using a short (60-day), rolling training window. The performances of these schemes are assessed through a set of hindcast experiments, wherein postprocessed 24-h PQPFs from the two DL 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 DL 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 DL 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., > 50mm/day), though the outperformance comes at a slight expense of sharpness and spatial specificity. Collectively, these results confirm the ability of the DL 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.</p>
Suggested Reviewers:	Michael Scheuerer scheuerer@nr.no Sandor Baran baran.sandor@inf.unideb.hu Rochelle Worsnop rochelle.worsnop@noaa.gov

Mohammadvaghef Ghazvinian, Ph.D.

Center for Western Weather and Water Extremes

Scripps Institution of Oceanography

University of California San Diego

La Jolla, CA

Email: mghazvinian@ucsd.edu

(817) 437 6440

February 14, 2022

Dear Dr. Crow,
Editor in Chief
Journal of Hydrometeorology

I am writing to submit our manuscript entitled “Improving Probabilistic Quantitative Precipitation Forecasts Using Short Training Data through Deep Learning” for consideration by *Journal of Hydrometeorology* as original research article.


In this study we introduce two unified deep learning-based schemes capable of creating medium-range, reliable, skillful, and spatially detailed probabilistic quantitative precipitation forecasts simultaneously over the contiguous United States. In contrast to a vast majority of conventional postprocessing schemes for precipitation forecast postprocessing, Machine learning schemes included, our proposed methods rely on only short training datasets for training thus adequately address the operational forecasting requirements where extensive historical observation-reforecasts data are not available for training.

When aggregated over the entire CONUS, the results show that, forecasts from proposed schemes broadly improve upon those from quantile mapping stencil algorithm (the current operational algorithm of the US National Blend of Models) and raw forecast in terms of reliability and predictive skill. The superior performances are more pronounced at predicting higher precipitation thresholds which are relevant to flood forecasting and real time reservoir management.

Please note that the manuscript currently exceeds the AMS word limit. This is mainly because of extended forecast verification. We appreciate your consideration of our manuscript, and we are happy to adjust the length during review process. Please do not hesitate to contact me should you have any questions or concerns.

Sincerely,

Mohammadvaghef Ghazvinian
Corresponding author



Click here to access/download

Cost Estimation and Agreement Worksheet
Journals Estimation Worksheet New Submission
Format.pdf

1 **Improving Probabilistic Quantitative Precipitation Forecasts Using**
2 **Short Training Data through Deep Learning**

3
4 Mohammadvaghef Ghazvinian,^{a,b} Yu Zhang,^a Thomas M. Hamill,^c

5 Dong-Jun Seo,^a Nelun Fernando^d

6
7 ^a *Dept. of Civil Engineering, The University of Texas at Arlington, Arlington, TX*

8 ^b *Current affiliation: Center for Western Weather and Water Extremes, Scripps Institution of Oceanography,*
9 *University of California San Diego, La Jolla, CA*

10 ^c *NOAA Physical Science Laboratory, Boulder, CO*

11 ^d *Texas Water Development Board, Austin, TX*

12
13
14
15 *Corresponding author: Mohammadvaghef Ghazvinian, mghazvinian@ucsd.edu*
16

17 ABSTRACT

18 Conventional statistical postprocessing techniques offer limited ability to improve the skills of
19 probabilistic guidance for heavy precipitation. This paper introduces two deep learning (DL)
20 based, geographically aware, and computationally efficient postprocessing schemes namely the
21 Artificial Neural Network – Multiclass (ANN-Mclass) and the ANN-Censored, Shifted
22 Gamma Distribution (ANN-CSGD). Both schemes are implemented to postprocess Global
23 Ensemble Forecast System (GEFS) forecasts to produce forecasts (PQPFs) over the contiguous
24 United States (CONUS) using a short (60-day), rolling training window. The performances of
25 these schemes are assessed through a set of hindcast experiments, wherein postprocessed 24-h
26 PQPFs from the two DL schemes were compared against those produced using the benchmark
27 quantile mapping algorithm for lead times ranging from 1 to 8 days. Outcomes of the hindcast
28 experiments show that DL schemes overall outperform the benchmark as well as the raw
29 forecast over the CONUS in predicting probability of precipitation over a range of thresholds.
30 The relative performance varies among geographic regions, with the two DL schemes broadly
31 improving upon quantile mapping over the central, south, and southeast, and slightly
32 underperforming along the Pacific coast where skills of raw forecasts are the highest. Between
33 the two schemes, the hybrid ANN-CSGD outperforms at higher rainfall thresholds (i.e., >
34 50mm/day), though the outperformance comes at a slight expense of sharpness and spatial
35 specificity. Collectively, these results confirm the ability of the DL algorithms to produce
36 skillful PQPFs with a limited training window, and point to the prowess of the hybrid scheme
37 for calibrating PQPFs for rare-to-extreme rainfall events.

38
39
40
41
42
43
44
45
46

47 **1. Introduction**

48 Accurate, spatially detailed quantitative precipitation forecasts (QPFs) are of paramount
49 importance for applications ranging from flash flood forecasting to reservoir management
50 (Cloke and Pappenberger 2009; Pappenberger and Buizza 2009; Brown et al. 2014; Scheuerer
51 et al. 2017). Despite continuing improvements in the accuracy of QPFs from numerical weather
52 prediction (NWP) models, statistical postprocessing has remained a vital supplemental
53 mechanism for enhancing the skill and spatial resolution of forecasts, and for quantifying
54 forecast uncertainties. Today, a plethora of conventional statistical postprocessing schemes
55 exist that serve these purposes. These range from the analog method (e.g., Hamill and Whitaker
56 2006; Hamill et al. 2015); variants of Bayesian approach (Krzysztofowicz, 2008; Wu et al.
57 2011; Robertson et al. 2013; Reggiani and Boyko 2019; Darbandsari and Coulibaly 2022); and
58 regression-based mechanisms (Hamill et al. 2004; Sloughter et al. 2007; Wilks 2009; Scheuerer
59 and Hamill 2015; Taillardat et al. 2019; Ghazvinian et al. 2020; to name a few).

60 The extant techniques, in particular the parametric schemes, have demonstrated wide
61 success in augmenting the skill of PQQPFs from diverse NWP systems for a range of lead times.
62 Yet, there is growing recognition that additional room for improving the schemes might be
63 limited, much due to the inflexible model structures and difficulties in selecting training
64 samples for establishing predictor-predictand relationships (Ghazvinian et al. 2021). This
65 recognition prompted various authors to explore newer, more flexible machine learning (ML)
66 techniques as alternative postprocessing mechanisms (Herman and Schumacher 2018;
67 Taillardat et al. 2019; Rasp and Lerch 2018; Bremnes 2020; Scheuerer et al. 2020; Baran and
68 Baran 2021; Ghazvinian et al 2021, Veldkamp et al. 2021; Chapman et al. 2021; Schulz and
69 Lerch 2021; Li et al. 2022). In the field of precipitation forecasts postprocessing, Scheuerer et
70 al. (2020) developed a multi-class Artificial Neural Network (ANN) scheme for subseasonal-
71 to-seasonal range (week 2-4). Herman and Schumacher (2018) created a random forest-based
72 postprocessing algorithm that has demonstrated prowess in producing skillful probabilistic
73 guidance for day 1 and 2 during recent Flash Flood and Intense Rainfall (FFaIR) experiments
74 (WPC; 2019, 2020). More recently, Ghazvinian et al. (2021), drawing inspirations from Rasp
75 and Lerch (2018), formulated a hybrid Deep learning (DL)-parametric framework that fuses
76 the ANN with the Censored, Shifted Gamma Distribution (CSGD; Scheuerer and Hamill
77 2015), namely the ANN-CSGD. Relative to the traditional parametric methods, all these ML
78 schemes offer flexibility in modeling predictor-predictand relationship and in integrating

79 ancillary predictors and allow for adaptive selection of spatio-temporal training windows.
80 Ghazvinian et al. (2021) demonstrated that ANN-CSGD broadly outperforms the original
81 CSGD and Mixed-type Meta-Gaussian Distribution (MMGD; Wu et al. 2011). This enhanced
82 performance, as the authors explained, is attributable primarily to the adaptive and therefore
83 more effective stratifications of training samples.

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

101 The aforementioned strengths of ANN models, in particular their ability to discern and
102 establish complex predictor-predictand relationship from a large, heterogeneous sample, as we
103 postulate, would render them particularly effective in alleviating the data paucity issue by
104 intelligently expanding the domains where forecast-observation pairs are pooled. We further
105 conjecture that ANN models' adaptive way of stratifying samples can lead to superior
106 calibration beyond what is attainable by the QMAP that relies on prescribed supplemental
107 locations. In this paper, we address these hypotheses by experimentally adapting and extending
108 two DL algorithms, namely the ANN-Mclass (Scheuerer et al. 2020) and the ANN-CSGD
109 (Ghazvinian et al. 2021), to short, 60-day training window. We perform a set of hindcast

110 experiments over a 3-year window for the entire CONUS, wherein we appraise the efficacy of
111 the adapted schemes relative to the QMAP in processing ensemble QPF for day 1-8.

112 The present study expands the work of Ghazvinian et al. (2021) in three major directions.
113 First, it entails comparisons of the skills of PQPFs produced by the hybrid ANN-CSGD and
114 ANN-Mclass to determine the merit of the former. Second, both DL algorithms integrate
115 ensemble attributes beyond the ensemble mean and incorporate geographic locations as well
116 as physiographic features, thus allowing for exploitation of skill gains associated with the
117 introduction of these predictors. Third, this study examines geographic variations in the relative
118 performance of DL and QMAP schemes to identify the dependence of skill differentials on
119 precipitation regimes and accuracy in NWP forecast. Furthermore, this study assesses the skills
120 of PQPFs for a range of thresholds much beyond the PoP, thereby providing critical
121 information about the robustness of various schemes in forecasting intense precipitation events
122 that is absent in extant studies in the context of NBM (Hamill et al. 2017; Hamill and Scheuerer,
123 2018).

124 The remainder of this article is structured as follows. Section 2 describes the two DL
125 schemes as well as the benchmark QMAP model adapted for this study, layout of the hindcast
126 experiment, and the training/validation data sets. Section 3 presents results of the hindcast
127 experiment and discusses findings. Section 4 summarizes the work and offers concluding
128 remarks.

129 **2. Materials and methods**

130 *a. Postprocessing schemes*

131 1) DEEP LEARNING WITH CATEGORICAL PROBABILITY PREDICTIONS (ANN-MCLASS)

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

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

139 Where pop_{cl} represent the probability of precipitation (> 0.254 mm) from climatology. The
 140 first class = $[q_{-1}, q_0]$ represents precipitation values below 0.254mm (dry conditions) and
 141 $q_m = \infty$. Scheuerer et al. (2020) chose to derive empirical quantiles from precipitation analysis
 142 for each grid point and day of the year using a 61-day window centered around the day of
 143 interest and all years of available data.

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

$$L_{1,\dots,m+1}(\mathbf{p}, \mathbf{y}) = -\log \left(\sum_{i=1}^{m+1} y_i p_i \right) \quad (2)$$

145 Where $\mathbf{p} = (p_1, \dots, p_{m+1})$ is vector of estimated probabilities for each of $m + 1$ classes,
 146 and $\mathbf{y} = (y_1, \dots, y_{m+1})$ is corresponding binary (one-hot encoded) truth vector that describes
 147 whether analyzed value falls into the respective category in a training case. This modification
 148 is necessary as the category assignment can be ambiguous (multiple cases with 1 due to
 149 duplicative values in climatology). MCCES reduces to standard categorical cross entropy when
 150 the assignments are unambiguous (see Appendix B of Scheuerer et al. 2020).

151 Continuous predictive distribution can be derived from the categorical probabilities by
 152 approximating the cumulative hazard function $H(x) = -\log [1 - F(x)]$ using piecewise
 153 linear interpolation/ extrapolations for the points inside/outside the data ranges. In the hazard
 154 function, $F(x)$ represents cumulative probabilities estimated by summing up the probabilities
 155 specific for individual categories and for each forecast case. Exploratory analysis by Scheuerer
 156 et al. (2020) showed that the interpolation provides a reasonable reconstruction of predictive
 157 CDF. A possible drawback of this model could be its reliance on the parameter m . As the
 158 number of classes directly impacts cross-entropy loss function value, other metrics such as
 159 continuous ranked probability score (CRPS; Matheson and Winkler 1976, Wilks 2011) should
 160 be used for configuration of optimal number of classes for final predictions (see Scheuerer et
 161 al. 2020).

162 2) HYBRID DEEP LEARNING-CSGD (ANN-CSGD)

163 Censored, shifted gamma nonhomogeneous regression model (CSGD) first was introduced
 164 by Scheuerer and Hamill (2015). This technique and its extensions have been shown to be
 165 capable of generating reliable and skillful medium-range PQPFs over different regions of the

166 world (see, Baran and Nemoda 2016; Zhang et al. 2017; Baran and Lerch 2018; Taillardat et
 167 al. 2019; Ghazvinian et al. 2020; Valdez et al. 2021). Nonetheless, Ghazvinian et al. (2020)
 168 noted a caveat of the CSGD that stems from the direct use of climatological shift without tuning
 169 and showed that this led to a negative bias in predicted PoP, particularly for shorter lead times.

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

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

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

178 The three parameters of ANN-CSGD are optimized by minimizing the average value of
 179 crps over training sample. As shown in Scheuerer and Hamill (2015), a closed-form expression
 180 for crps exists for CSGD, and it takes the following form:

$$\begin{aligned} \text{crps}(F_{k_i,\theta_i,\delta_i}, y_i) = & (y_i - \delta_i)[2F_{k_i,\theta_i}(y_i - \delta_i) - 1] - \frac{\theta_i k_i}{\pi} B\left(\frac{1}{2}, k_i + \frac{1}{2}\right) [1 - \\ & F_{2k_i,\theta_i}(-2\delta_i)] + \theta_i k_i [1 + 2F_{k_i,\theta_i}(-\delta)F_{k_i+1,\theta_i}(-\delta_i) - \\ & F_{k_i,\theta_i}(-\delta_i)^2 - 2F_{k_i+1,\theta_i}(y_i - \delta_i)] + \delta F_{k,\theta}(-\delta)^2 \end{aligned} \quad (4)$$

181 where $B(0,0)$ is the beta function, and $(k_i, \theta_i, \delta_i)$ are three parameters of i th predictive
 182 CSGD with y_i being the corresponding verifying observation. Ghazvinian et al. (2021)
 183 demonstrated that ANN-CSGD alleviates the negative bias in PoP by directly learning shift
 184 parameter as an arbitrary function of predictors. In addition, the authors showed that ANN-
 185 CSGD's efficient way of modeling complex interactions between three CSGD parameters is a
 186 major factor that contributed to its superior predictive skill at higher thresholds relative to the
 187 original CSGD.

188 3) QUANTILE MAPPING STENCIL

189 Chosen as the benchmark technique is the QMAP algorithm a coarse-grid approximation
190 to the current baseline postprocessing scheme for a single ensemble prediction system
191 component of the NBM. The algorithm was first described by Hamill et al. (2017), and was
192 later modified in Hamill and Scheuerer (2018) to incorporate quantile dressing. In essence, it
193 establishes matching quantiles from training data (forecasts and corresponding analyses for
194 designated locations). The resulting quantile map function is then applied to real-time forecasts
195 to produce probabilistic guidance grids. To augment sample size, the algorithm incorporates
196 the so-called supplemental locations. The supplemental locations are locations that share
197 similar precipitation climatology (as represented by climatological CDF), elevation, and terrain
198 orientation (facet).

199 In this study, we implement the original version of QMAP as described by Hamill et al.
200 (2017), but for simplicity chose to forgo the dressing mechanism – we do so with the tacit
201 assumption that the incremental improvements in skills as a result of dressing would be
202 insufficiently large to alter the relative performance of the algorithms. In our implementation
203 of the QMAP scheme, up to 100 supplemental locations are identified for each target grid point
204 (0.25×0.25 degree in size), and for each month of the year using the data from the respective
205 month and two surrounding months. For each lead time and the grid point of interest, empirical
206 quantiles $q(p), p \in \{1/100, 2/100, \dots, 99/100\}$ are constructed from the augmented forecast
207 and analysis data sets, which are accumulated within a rolling window and supplemental
208 locations. Note that in this step, ensemble members from Global Ensemble Forecast System
209 (GEFS), are pooled to populate forecast CDFs based on the assumption that these members are
210 identically distributed. To account for wider sampling variability in larger forecast quantiles
211 (larger than 0.95 quantile of forecasts), a linear approximation of quantile mapping function is
212 applied. A detailed explanation of this procedure can be found in Hamill et al. 2017.

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

220 size on the skills of postprocessed PQPFs, the QMAP method was implemented on 1×1
221 (center point) and 5×5 stencils.

222 *b. Architecture of two DL schemes*

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

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

- 239 • Input layer where predictors are introduced to the network.
- 240 • Batch normalization (Ioffe and Szegedy 2015). This practice normalizes input to
241 maintain the mean of each feature close to 0 and the standard deviation close to 1.
- 242 • Hidden layers (*Dense*) with nonlinear activation functions.
- 243 • Output layer

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

- 245 • ANN-CSGD uses output layer with linear activation function and three nodes to
246 represent functions of three CSGD parameters. Additional functions are required to
247 limit CSGD parameters in allowable ranges. We follow Ghazvinian et al. (2021) to set
248 CSGD shift parameter: $\delta = -\text{sqrt}(O_1^2)$ and use inverse logarithmic link functions for

249 location and scale parameters (\exp) , $\mu = \exp(O_2^2)$, and $\sigma = \exp(O_3^2)$. Where oi
250 represents i th output node.

251 • ANN-Mclass uses an output layer with 50 nodes and softmax activation function to
252 ensure that output probabilities are in the range [0,1] and sum to 1. For each forecast
253 day, observation quantiles are derived using CCPA data of previous 60 days and all
254 CCPA grids. This yields a sample of $60 \text{ days} \times 13528 \text{ locations}$ worth of data from
255 which the empirical quantiles are calculated daily. One-hot encoded CCPA data in each
256 location and day are then assigned to the designated classes. The latter approach helps
257 maintain balanced assignments between classes $i \in \{1, \dots, m\}$. As another structural
258 modification to the work by Scheuerer et al. 2020, we do not include climatological
259 information in the last layer. This practice was necessary for postprocessing
260 subseasonal forecasts to ensure that estimated QPPFs revert to climatology in the cases
261 where signal to noise ratio was limited. In our application this is considered redundant
262 as it complicates the model and makes overfitting more likely.

263 For each lead time and each day, previous 60 days' worth of data over the entire CONUS are
264 available for training. To reduce generalization errors we use early stopping, one of the most
265 efficient and widely used regularization techniques (See Goodfellow et al. 2016). In our
266 application we keep the last 6 days of training data for validation (not used in training) and
267 monitor its average loss value. The training is terminated when no further decrease in the loss
268 is seen after three epochs (*patience*). To simplify matters, we decided to avoid extensive grid
269 search for hyperparameter tuning. Loss functions were minimized using adaptive moment
270 estimation (*Adam*) algorithm (Kingma and Ba, 2014) with the learning rate kept fixed at $lr =$
271 0.01. Following model architectures for hidden layer were tested:

- 272 • Number of nodes in hidden layer (s) (ANN-CSGD): {[10], [20],[10,10]}
- 273 • Number of nodes in hidden layer (s) (ANN-mclass): {[20],[50], [20,20]}

274 Our initial assessment showed that expanding the number of hidden layers does not
275 systematically improve validation loss (possibly due to overfitting). Thus, we did not test
276 number of hidden layers > 2 in final model configurations. The batch size was set to 10000 for
277 training both networks. Networks were trained with a common random number generator
278 (seed) and the configuration with the lowest validation loss was saved for each out-of-sample
279 day prediction.

280 *c. Hindcast experiment setup*

281 In this study we focus on postprocessing forecasts of 24-hourly accumulated precipitation
282 for 1-8 day lead time over the CONUS. We leveraged Global Ensemble Forecast System
283 (GEFS) -version 11 (v11) reforecast data sets (Hamill et al. 2013).. The GEFS-v11 data we use
284 were produced on a quadratic Gaussian mesh with $\sim 0.5^\circ$ resolution for the first 8 days and \sim
285 0.67° for 9–16 days lead times. The reforecasts are composed of 11 ensemble members (10
286 perturbed and one control) issued every 24-h at 00 UTC. The reforecast data were retrieved,
287 extracted for CONUS on native Gaussian grid, bilinearly interpolated to a 0.25° grid mesh and
288 accumulated to 24-h sums. As the analyses we use Climatology-Calibrated Precipitation
289 Analysis (CCPA; Hou et al. 2014), which is available on 6-h increments spanning from 1
290 January 2002 to 31 Dec 2019 on the CONUS National Digital Forecast Database (NDFD;
291 Glahn and Ruth 2003) grid resolution (see [https://vlab.noaa.gov/web/mdl/ndfd-spatial-
292 reference-system](https://vlab.noaa.gov/web/mdl/ndfd-spatial-reference-system)). The latter data were upscaled to 0.25° resolution and accumulated to 24-h
293 sums. To mimic the US NBM operations as described in Hamill et al. (2017), training for each
294 scheme is performed each day on forecasts for the lead times of 1-8 days and corresponding
295 analysis over previous 60-day rolling window. The trained schemes are then applied to
296 forecasts of prediction day to create PQPFs that are then verified against coincidental analysis.
297 This training-verification cycle repeats progressively in time over a 3-year window extending
298 from 1 January 2017 to 31 December 2019.

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

305 **3. Results**

306 *a. CRPSS*

307 To assess the relative overall predictive performance of three postprocessing schemes, we
308 first examine continuous ranked probability skill score (CRPSS). We use climatological CRPS
309 as the reference. Climatological CRPS was calculated for each grid point, separately for each
310 month using CCPA data pooled across a 3-month window surrounding that month from years

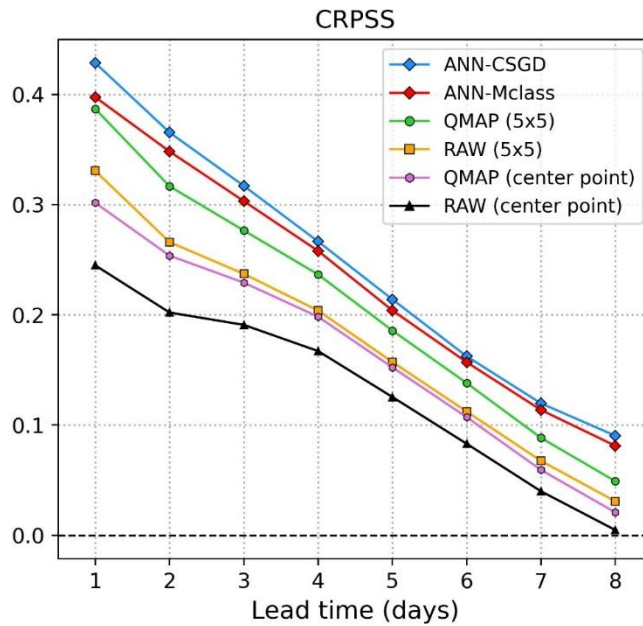
311 between 2002 and 2015. For each forecast suite, we first compute CONUS-wide, lead-time
 312 specific CRPSS by aggregating either the raw ensemble QPFs or postprocessed PQPFs among
 313 all grid points in CONUS and across all verification days within the 3-year window. To
 314 highlight the impacts of using an expanded spatial domain, we computed the results using raw
 315 forecasts over each target grid point (center point) and corresponding 5×5 stencil. Note that,
 316 Table 1 summarizes different model configurations and corresponding abbreviations used in
 317 the figures of this study.

318

Model	Experiment name	5×5 stencil	Supplemental locations
DL	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	-
	RAW (center point)	No	-

319 Table 1. Different experiment names and configurations used

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



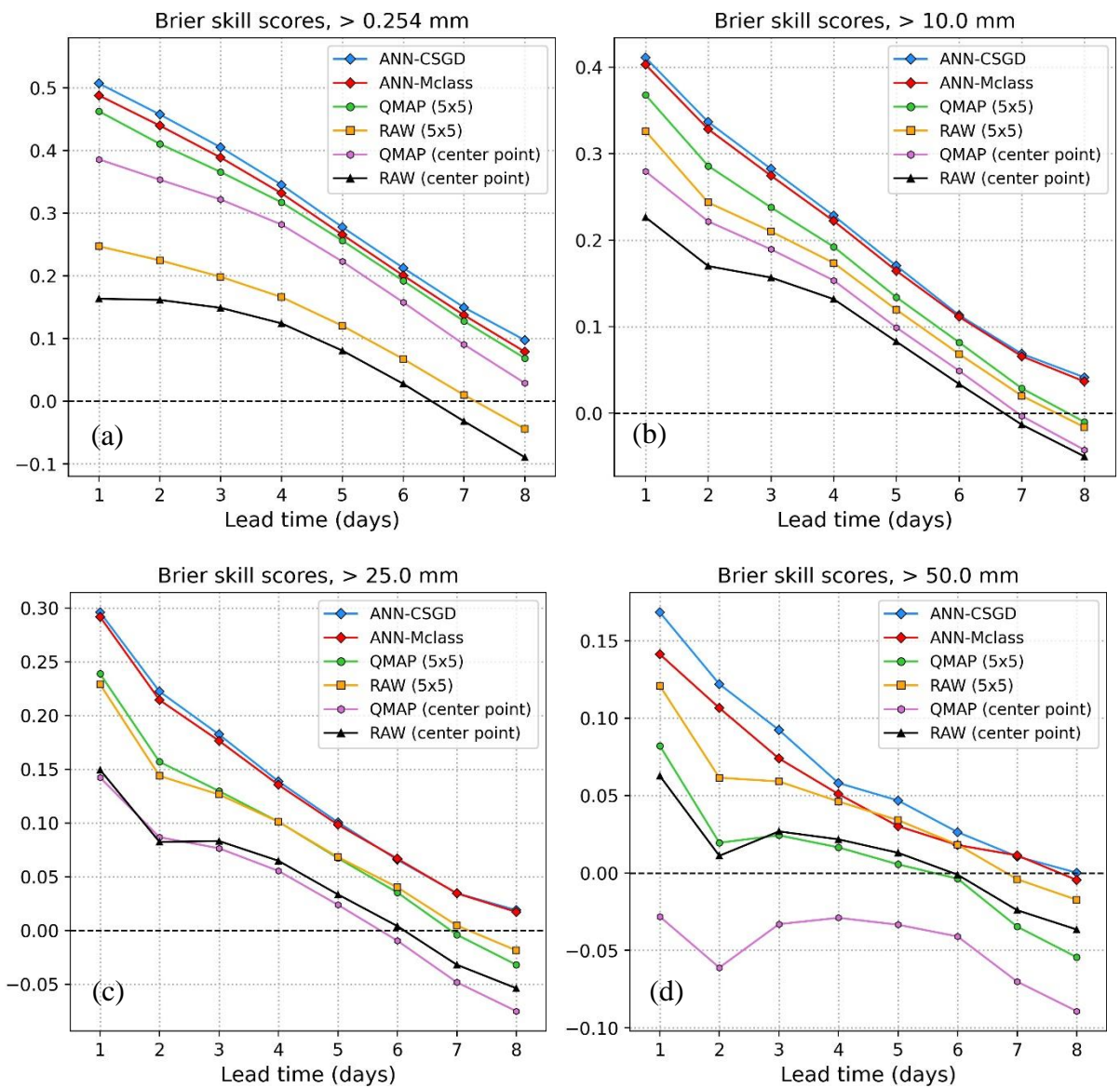
327

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

330 *b. Brier skill scores*

331 To assess the relative efficacy of three postprocessing schemes in predicting specific
 332 events, we examine the Brier skill scores (BSS; Wilks 2011; Hamill et al. 2015) computed for
 333 four daily accumulation thresholds, namely 0.254-, 10-, 25-, and 50-mm. Fig. 2 shows the
 334 resulting BSS from the six forecast sources. At the two lowest thresholds here, it is evident that
 335 raw GEFS forecast (11-member ensemble) from the center point are the most poorly
 336 performing forecast of all. Similar to CRPSS, quantile mapping appreciably improves PoP
 337 forecasts without and with the expanded domain (Fig. 2a). However, the gap between raw
 338 ensemble and quantile mapped PQQFs diminishes at higher thresholds. It appears that quantile
 339 mapping mainly improves BSS of PQQF from raw forecast at the lowest thresholds, which have
 340 disproportionate impacts on its overall prediction performance as shown in CRPSS results (Fig.
 341 1). In fact, quantile mapped PQQFs broadly underperform raw ensemble forecasts at the highest
 342 threshold (50 mm; Fig. 2d). In addition, without domain expansion, PQQF from quantile
 343 mapping center point underperforms climatology across lead times (Fig. 2d). Postprocessed
 344 PQQFs via the two DL schemes manage to improve the forecast skill from raw and quantile
 345 mapped sources across all thresholds and throughout the lead times. It is also worth noting that
 346 ANN-CSGD and ANN-Mclass both improve the skill of raw forecast at longer lead times
 347 where raw and quantile mapped forecasts are unskillful relative to climatology. Between the

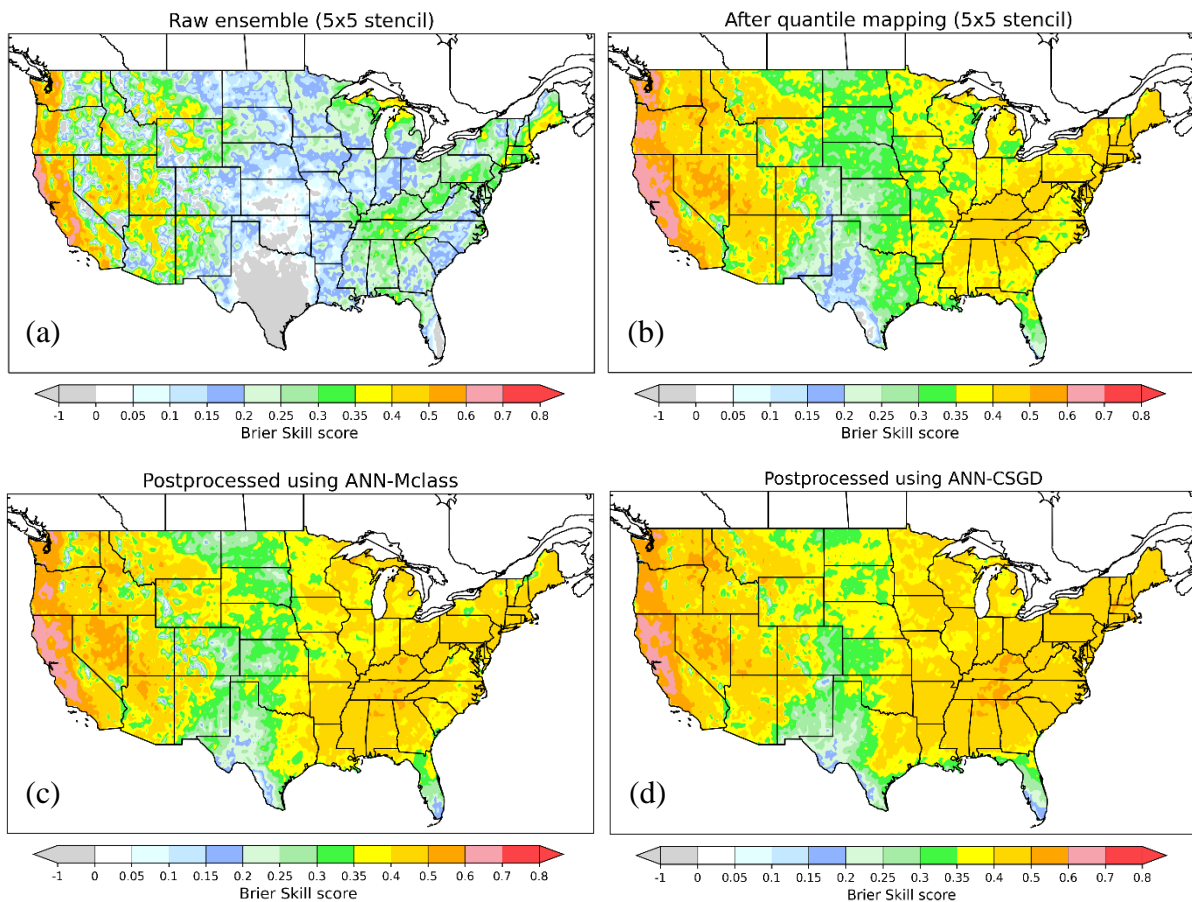
348 two DL schemes, ANN-CSGD slightly outperforms ANN-Mclass, and this performance
 349 differential is the most evident at the highest threshold of 50 mm (Fig. 2d).



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

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

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

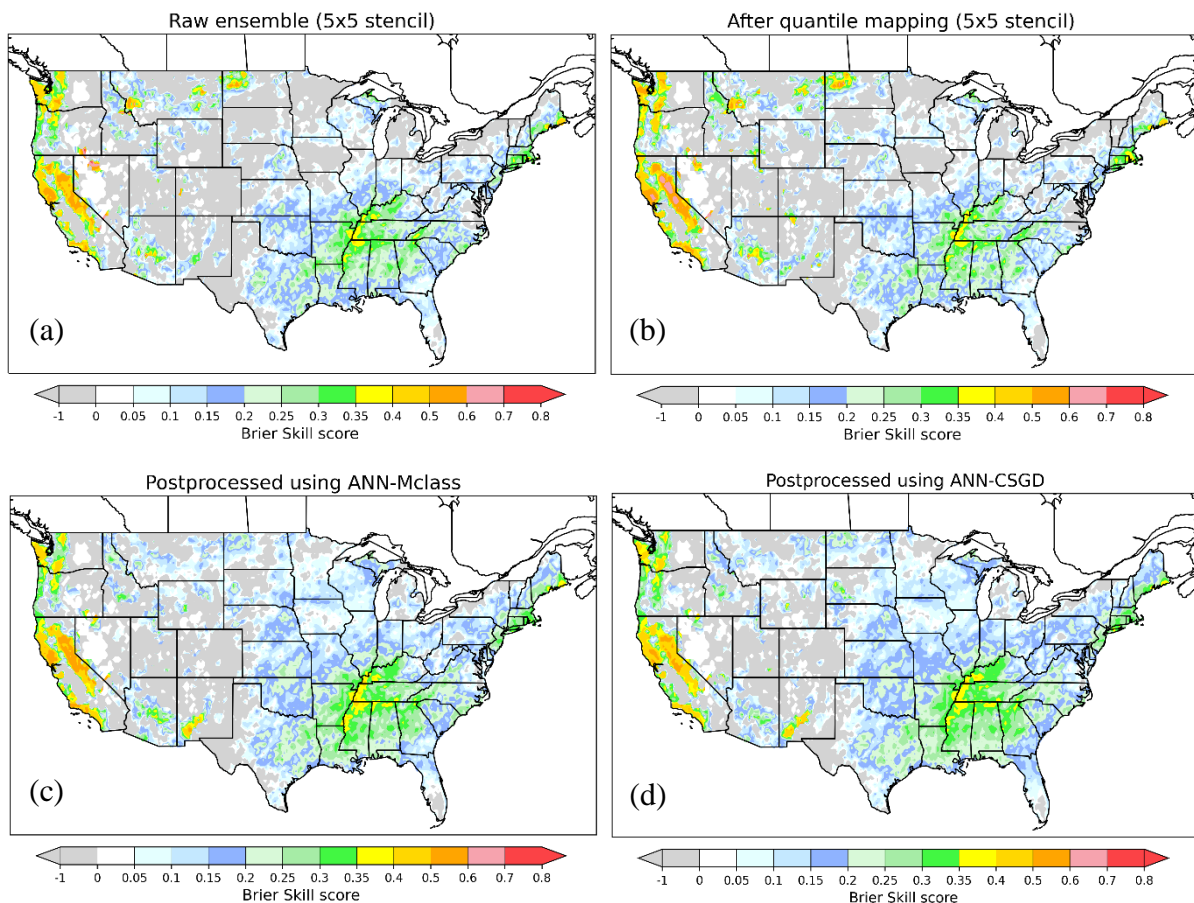


359 Fig. 3. Maps of Brier skill scores values of PoP forecasts aggregated over all days and for lead time +48h to
 360 +72h over the CONUS. Climatology is used as the reference. Forecasts are generated from (a) raw GEFS
 361 forecast using 5×5 stencil of grid points. (b) quantile mapped forecast using 5×5 stencil of grid points.
 362 (c) ANN-Mclass and (d) ANN-CSGD.

363 Fig. 3 shows the maps of BSS computed with 0.254mm threshold (PoP). Some of the
 364 prominent features mirror those noted in past studies. In particular, of all regions in the
 365 CONUS, the skill of raw GEFS ensemble appears highest along the Pacific coast to the west
 366 of the Cascade and the Sierra-Nevada Mountain ranges (Fig. 3a). This reflects the high
 367 predictability of orographically induced, synoptically forced precipitation systems that are
 368 predominant rainfall producers in these regions (Brown et al. 2014; Hamill et al. 2015;
 369 Scheuerer and Hamill 2015). By contrast, forecast skills of GEFS are the lowest over parts of
 370 Texas and southern Florida, where BSS is overwhelming negative (Fig. 3a). The region of low
 371 BSS values extends northward to cover much of the Great Plains, whereas clusters of areas
 372 with high BSS values are found along the windward side of the Appalachians, and between the
 373 Cascades and the Rockies (Fig. 3a). QMAP drastically improves the skills with respect to the
 374 forecast of PoP for nearly all regions in CONUS (Fig. 3b), though negative remains in small

375 areas over the southern tip of Texas. Both DL schemes broadly outperform QMAP, and the
 376 outperformance is the most conspicuous to east of the Rockies (Figs 3c and d). Between the
 377 two schemes, ANN-CSGD extends the skill of PQPF over the upper Midwest, the South, and
 378 the Southeast. Nonetheless, it is worth noting that the two DL schemes slightly underperform
 379 QMAP along the Pacific coast where skills of raw GEFS are high.

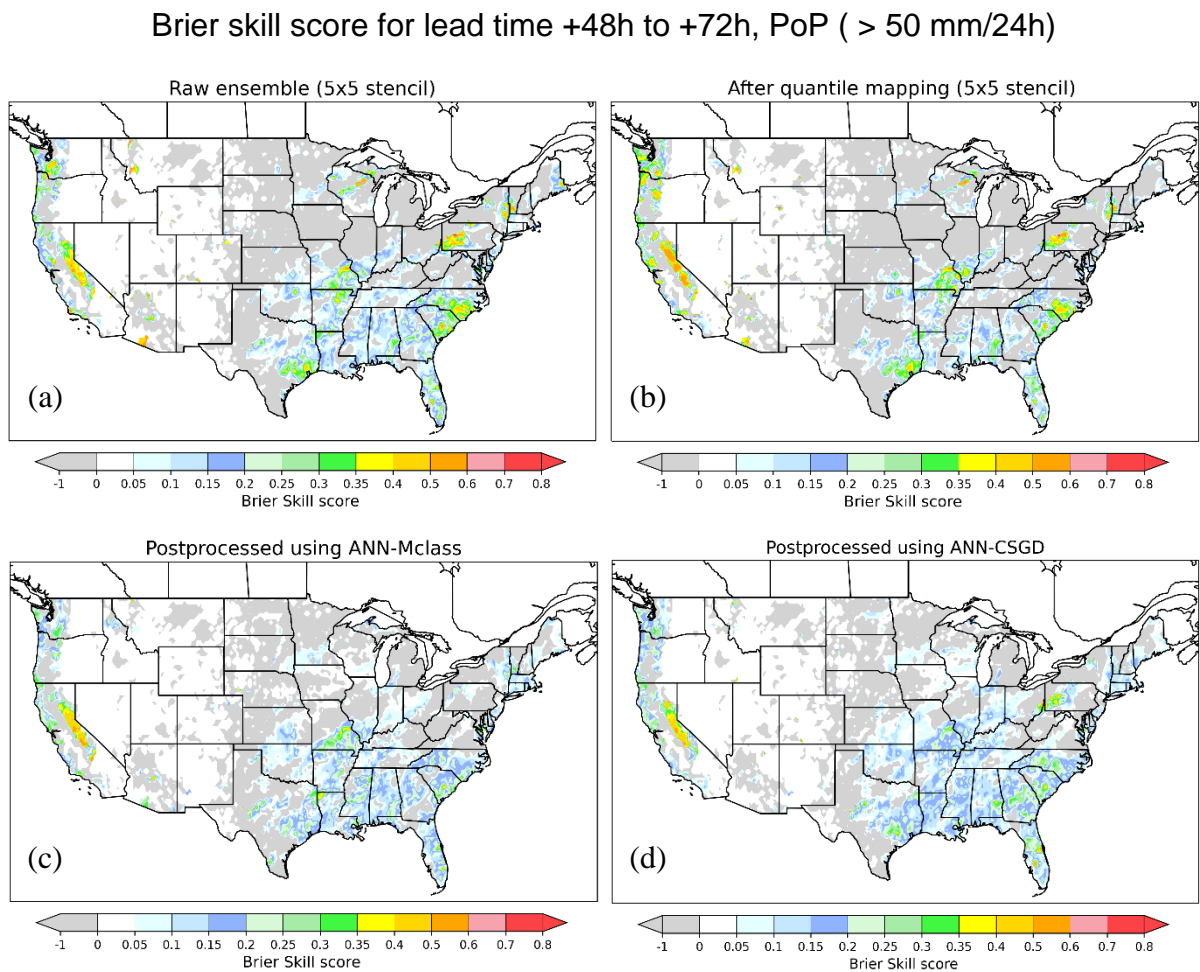
Brier skill score for lead time +48h to +72h, (> 25 mm/24h)



380 Fig. 4. As in Fig. 3, except for the events > 25 mm.

381 Similar comparisons of BSS but for two higher thresholds, namely 25 and 50 mm, are
 382 shown in Figs. 4 and 5 respectively. Notable features are summarized as follows. First, raw
 383 GEFS ensemble remains skillful relative to climatology along the Pacific Coast, eastern portion
 384 of the Southern Great Plains, much of the Midwest, South, Southeast, and along the Mid-
 385 Northeast Atlantic coast, but it underperforms climatology over the upper Midwest, the
 386 Rockies, and the western portion of the Great Plains (Figs 4a and 5a). Second, the performance
 387 of QMAP is mixed across the nation, in direct contrast to the wide skill improvements evident
 388 at the 0.254mm threshold (Figs. 4b and 5b). The improvement is still appreciable over a few

389 regions including the Sierra Nevada, but over other parts of the country, e.g., South and
 390 Southeast, QMAP appears to degrade the skills of raw ensemble. Third, both DL schemes are
 391 able to bring modest skill improvements for regions east of the Rockies. Between the two
 392 schemes, ANN-CSGD results in skill improvements over wider geographic regions, consistent
 393 with the earlier observation that it offers the best overall performance for CONUS (Fig. 2c and
 394 d).



395 Fig. 5. As in Fig. 3, except for the events > 50 mm.

396 Another subtle, yet important observation in Figs 4 and 5 is that, despite the broad
 397 outperformance of DL schemes, in a few regions the schemes conspicuously underperform the
 398 QMAP. Examples include the Sierra Nevada, southeast Texas, and the Carolinas. In each of
 399 these regions, raw GEFS ensemble offers good skills (BSS > 0.35); the skills are retained in
 400 QMAP-ed results, but are clearly degraded in the postprocessed PQPFs produced by applying
 401 the two DL schemes. Therefore, it appears that the CONUS-wide skill gains associated with
 402 the application of the DL schemes are mostly a result of improvements over domains where

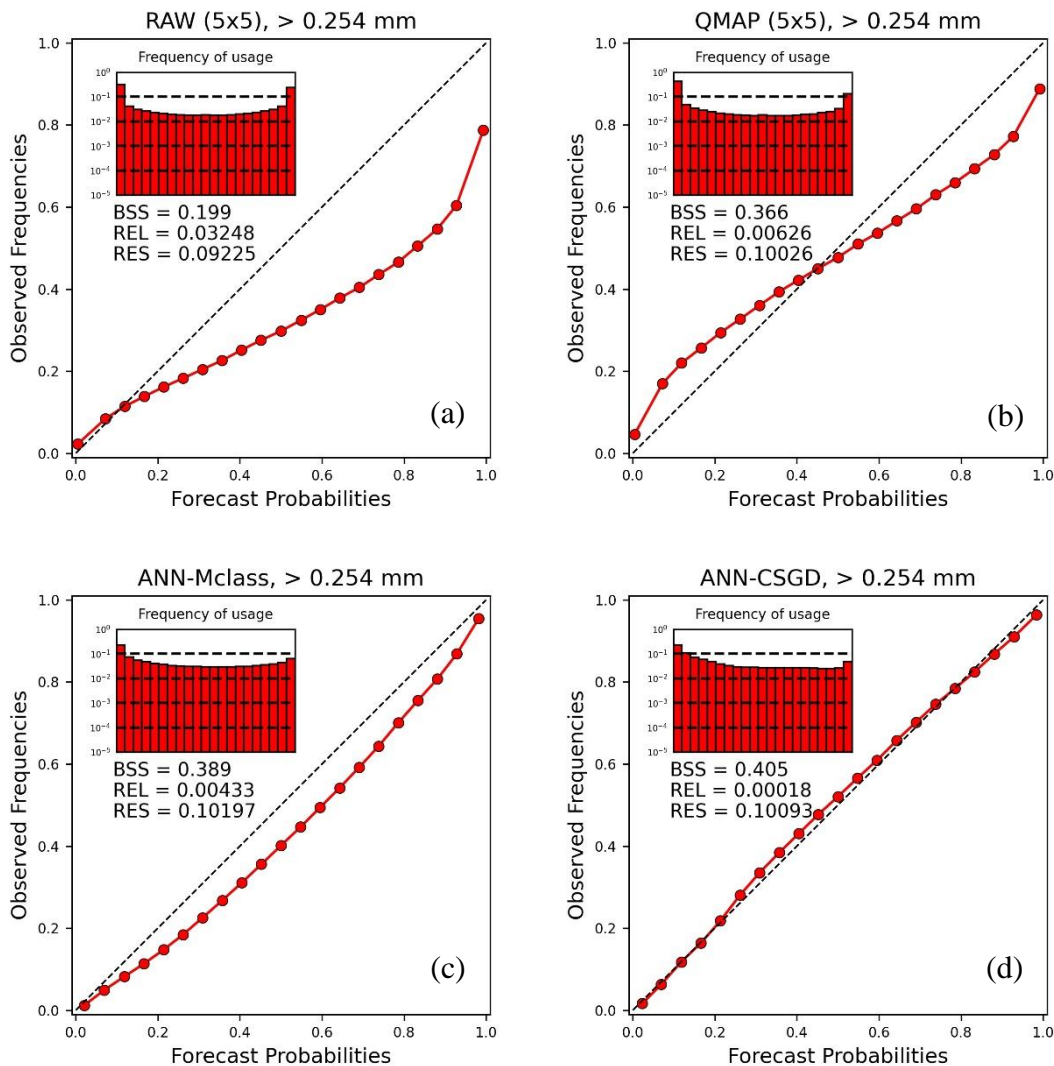
403 raw ensemble forecasts are marginally or modestly skillful. These improvements, however, are
404 achieved at the expense of reduced skills for a few regions where raw ensemble forecasts are
405 particularly accurate. As the former regions are far larger in size, the improvements seen therein
406 tend to overshadow degradations observed over latter locations. It is heretofore unclear what
407 contributes to QMAP's outperformance over the latter regions, but QMAP's reliance on a
408 restricted set of supplemental locations most likely plays a critical role. This practice, as we
409 surmise, may have limited overdispersion caused by the use of an unduly large amount of grid
410 points with heterogenous predictor-predictand relationships.

411 *c. Reliability diagrams*

412 Figs 6-8 show reliability diagrams computed for the raw GEFS ensemble and three sets of
413 postprocessed PQPFs for three thresholds, i.e., 0.254, 25 and 50mm, again for 48-72 h lead
414 time, with corresponding histograms of relative frequency of usage (sharpness histograms)
415 superimposed. The reliability diagrams allow us to assess forecast attributes including
416 reliability and resolution (see, Brocker and Smith 2007; Wilks 2011). In constructing the
417 reliability diagrams, for raw and quantile mapped forecasts based on center point only (11-
418 member ensemble) we use 12 equally spaced probability categories within the ranges of [0,1]
419 (see supplemental materials for center point results and results for additional threshold). The
420 reliability diagrams for DL-based PQPFs, as well as raw and quantile mapped using 5×5
421 stencil, are computed by stratifying probabilities into 21 bins. We further perform
422 decomposition of Brier score (Brier 1950) into reliability, resolution and uncertainty as
423 proposed by Murphy (1973) and assess the contribution of each component to PQPF skill in
424 predicting specific events. Among these, resolution characterizes the forecast's ability to
425 discriminate between different events and is identical to sharpness for perfectly reliable
426 forecasts (see Jolliffe and Stephenson 2012). The resulting BSS, reliability (REL) and
427 resolution (RES) are superimposed on each reliability diagram. We choose to leave out
428 uncertainty as it is independent of forecast source.

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

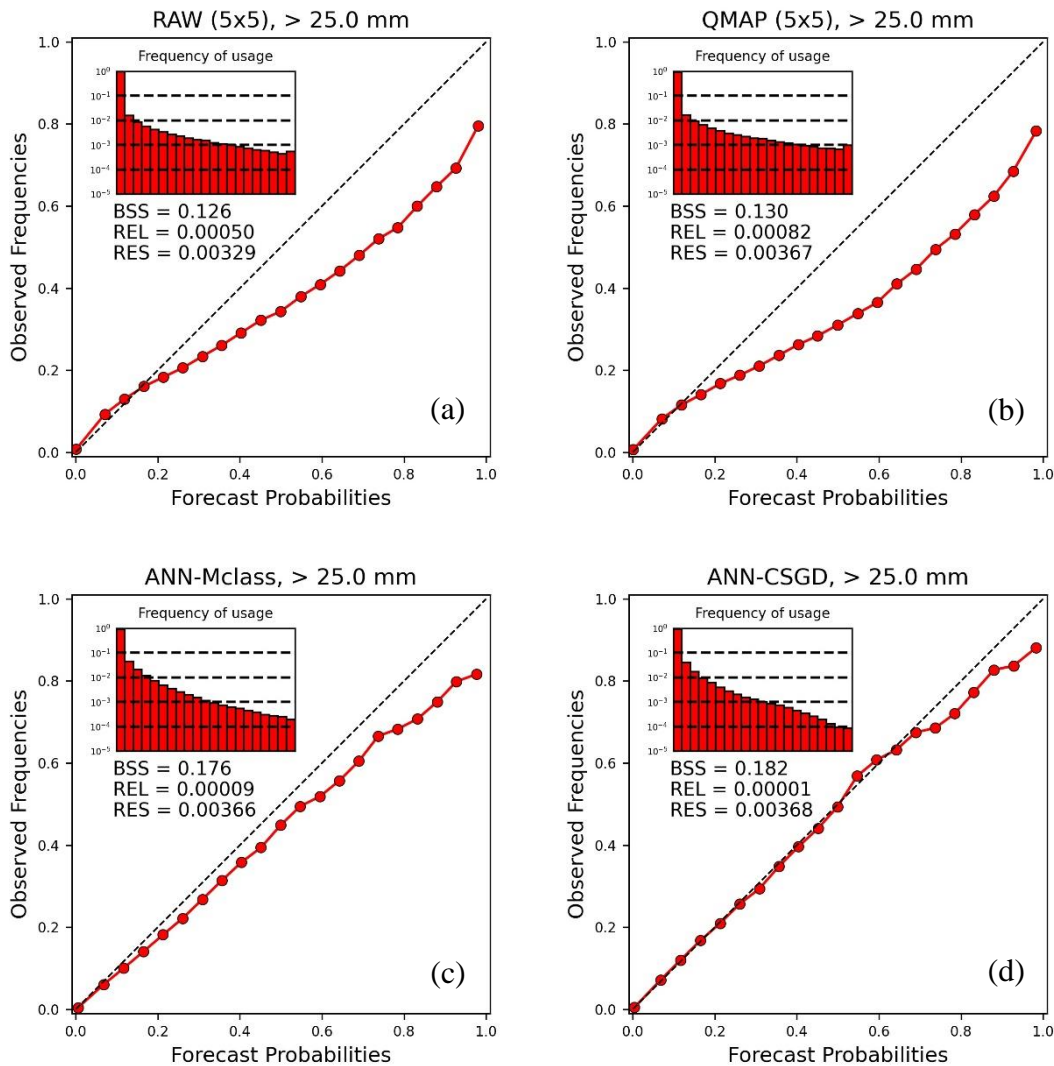
Reliability diagrams (PoP) for lead time +48h to +72h



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

439 ANN-Mclass produces PQPFs with further improved reliability and resolution, but tends
 440 to consistently overforecast (Fig. 6c). ANN-CSGD PQPFs outperform the rest in terms of
 441 reliability and resolution, and there is no conspicuous tendency to over/underforecast (Fig. 6d).
 442 That said, it is worth noting that ANN-CSGD PQPFs exhibit the lowest sharpness as judged
 443 by the sharpness histogram. At the 25 mm threshold (Fig. 7), raw and quantile mapped forecasts
 444 perform comparably, though the latter is slightly more skillful due to its improvement in
 445 resolution (Figs. 7a-b). PQPFs generated by DL schemes on the other hand are much more
 446 reliable and skillful than the former two forecast sources, but they are not as sharp: these PQPFs
 447 feature lower frequencies for high probability categories (Figs. 7c and d).

Reliability diagrams (> 25 mm) for lead time +48h to +72h

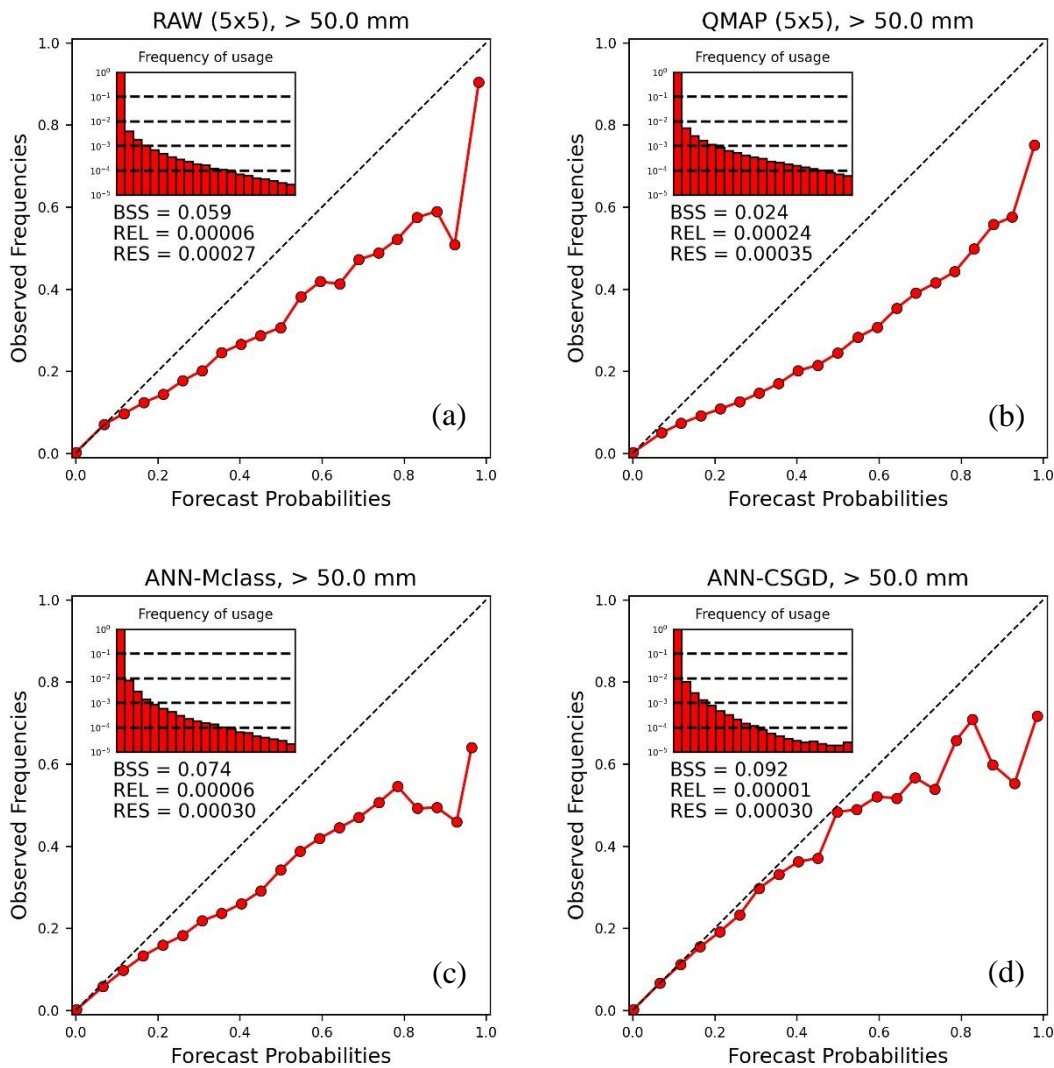


448 Fig. 7. As in Fig. 6 except for > 25 mm events.

449 As previously shown in the comparisons of BSS, the performance gap between raw ensemble
 450 and QMAP PQPFs tends to narrow at higher thresholds. From Fig. 7b, it appears that QMAP
 451 slightly degrades the reliability but improves the resolution and, to a limited extent, the
 452 sharpness. This divergent outcome is rooted in the mismatch between forecast and analysis.
 453 As noted by Hamill and Whitaker (2006), the raw GEFS ensemble has a strong tendency to
 454 overpredict precipitation amounts for events with light-to-moderate intensity over for much of
 455 CONUS, and meanwhile it contains a small, but substantial number of instances where it
 456 underpredicts precipitation amounts for events associated with larger accumulations. For larger
 457 forecast amounts, quantile adjustment tends to increase the forecast amounts. But due to the
 458 forecast-analysis mismatch, this increase serves to inflate the amounts for a vast majority of

459 events where analyzed accumulations are actually lower than the forecasted, thus amplifying
 460 the wet bias that is already existent in the raw forecast (Fig. 7b). At the highest threshold, i.e.,
 461 50 mm (Fig. 8), the relative performance of three schemes broadly echoes those at the 25mm
 462 threshold but a few distinctions are apparent.

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



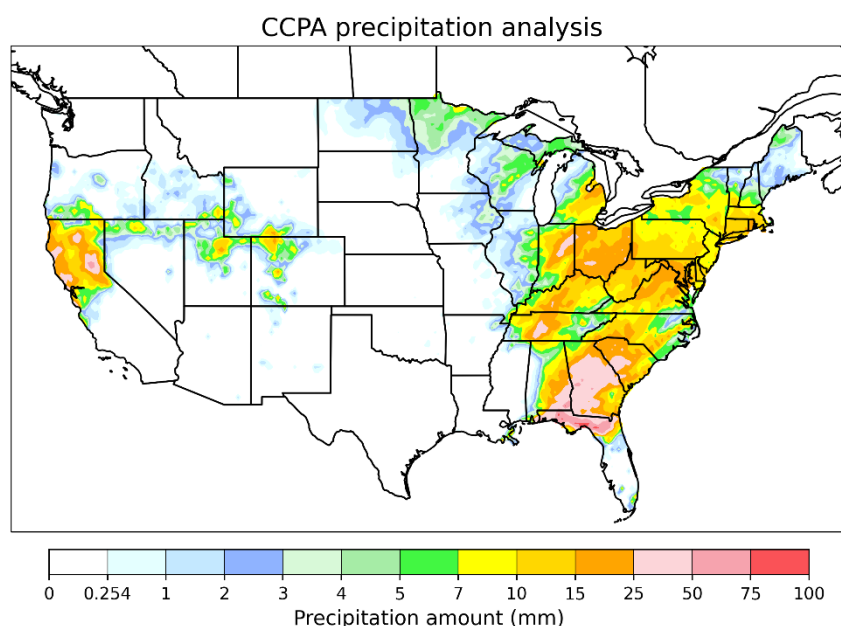
463 Fig. 8. As in Fig. 6 except for > 50 mm events.

464 First, QMAP scheme tends to more severely degrade the reliability that it does at the 25mm
 465 threshold, resulting in a conspicuous overforecast across probability categories, though there
 466 is a sign that it improves the sharpness (Fig. 8b). Second, while both DL schemes (Figs. 8c
 467 and d) again yield PQPFs with improved reliability relative to the raw ensemble, but the margin
 468 of improvements narrows somewhat, and at the two highest probability categories both

469 schemes underperform the QMAP by featuring more severe positive bias (overforecast).
470 Between the two DL schemes, ANN-CSGD clearly outperforms in terms of calibration but at
471 the cost of reduced sharpness.

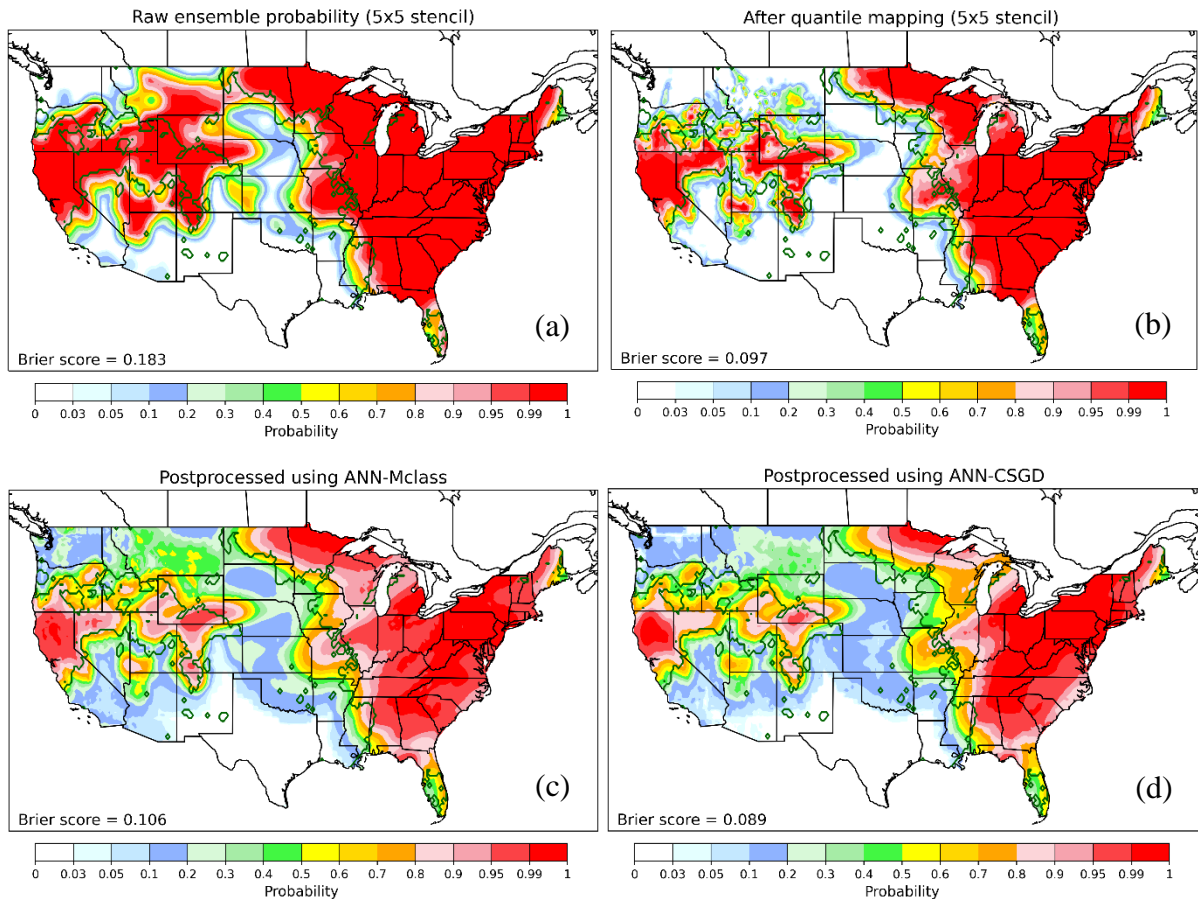
472 *d. Case study*

473 To further illustrate the skills of PQPFs and shed light on their geographic and
474 precipitation-regime dependence, we construct forecast guidance from raw ensemble and
475 postprocessed PQPFs in a way that mimics the NBM operation and compared these against
476 areas where corresponding thresholds are exceeded in the analysis. Such a practice is widely
477 adopted in NWS forecast verifications (see e.g., WPC 2019). This verification exercise focuses
478 on a one-day window ending at 00 UTC on 4 January 2017. As shown in Fig. 9, this window
479 was so chosen that there were several large precipitation clusters simultaneously present over
480 the west coast (Northern California), between the Midwest and Mid-Atlantic coast, and over
481 the southeast (Alabama, Georgia, South Carolina, and part of northern Florida). Maximum 1-
482 day accumulations for these clusters all exceeded 50mm.



483
484 Fig. 9. CCPA precipitation analysis for 24h accumulated data ending at 00 UTC 4 January 2017.

485 We computed exceedance probabilities from Day 2 (+24h to +48h) GEFS ensemble
486 forecasts for the valid date ending at 00 UTC 4 January 2017. Fig. 10 displays maps of PoP
487 ($> 0.254\text{mm}$) computed based on raw ensemble (with 5×5 stencil) and derived from three
488 suites of postprocessed PQPFs, namely those produced by QMAP (with 5×5 stencil), ANN-
489 Mclass, and ANN-CSGD.

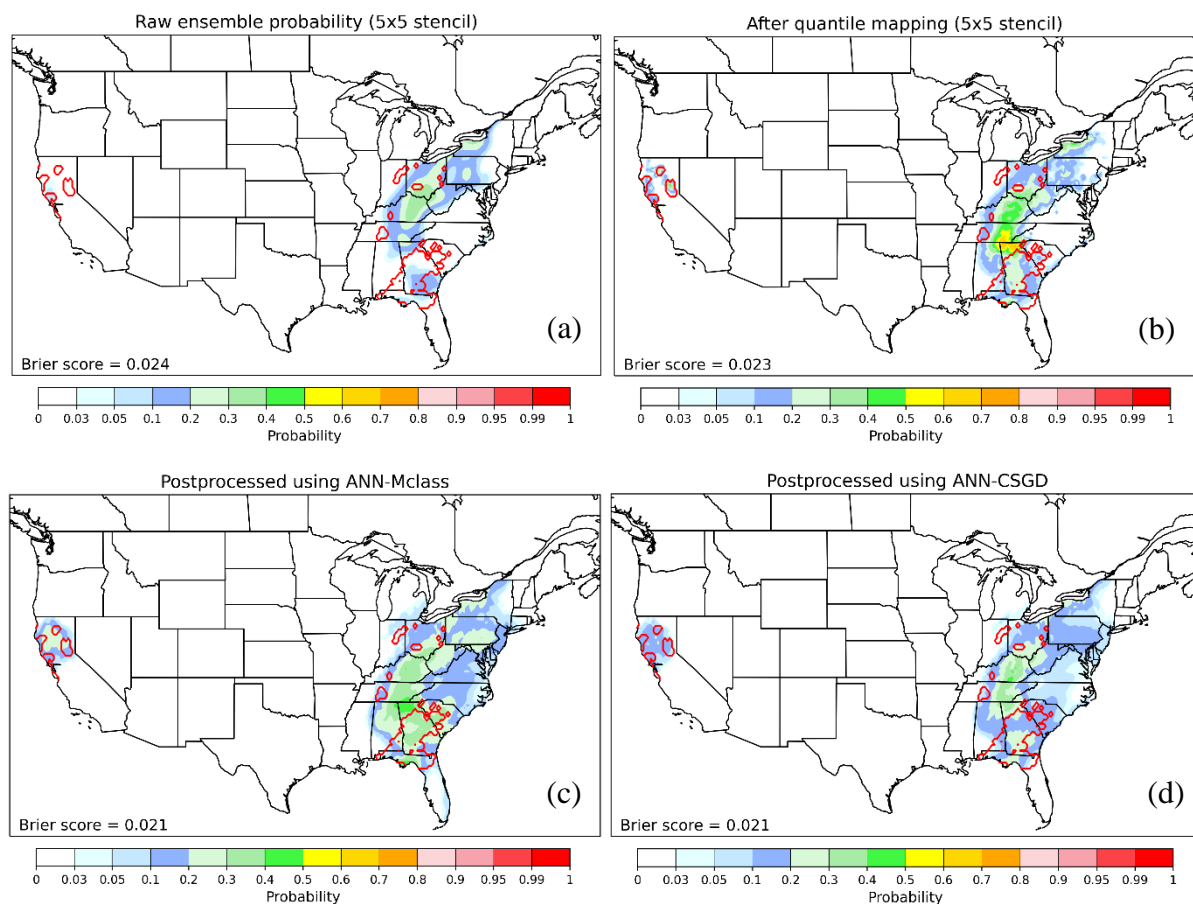


490 Fig. 10. POP forecast guidance for +24h to +48h lead time over CONUS for valid date ending at 00 UTC 4
 491 January 2017. (a) Raw ensemble forecast using 5×5 stencil, (b) quantile mapped forecasts using 5×5
 492 stencil, (c) forecasts generated using ANN-Mclass and (d) forecasts generated using ANN-CSGD. Areas
 493 inside dark green contours show that event > 0.254 mm has been observed by CCPA.

494 Average values of Brier score (averaged over all grid points for this day) are overlaid on
 495 each map to gauge the CONUS-wide performance of each product. The following features are
 496 evident. First, PoP from raw ensemble is broadly higher (close to 1) and there is a conspicuous
 497 lack of spatial details (Fig. 10a). Second, in many parts of the CONUS, the PoP is close to one
 498 despite a lack of precipitation in CCPA, consistent with the severe overforecast seen in earlier
 499 reliability diagrams for the raw ensemble (Fig. 6a). Quantile mapping drastically improves
 500 locational precision by reducing the areas where raw ensemble produces high PoP (Fig. 10b).
 501 This reduction is particularly noticeable over the intermountain west to the east of the Sierra
 502 Nevada, where gradients in PoP values emerge after quantile mapping. The BS is much
 503 reduced, corresponding to this reduction in areas with overforecasted PoP. The two DL
 504 schemes produce further improvements for the Intermountain West, by suppressing the PoP
 505 values outside the areas where CCPA indicates wet conditions (Figs. 10c and d). The
 506 postprocessed PoPs from the two schemes feature much lower overforecast errors over this
 507 region, thus allowing the actual precipitation clusters to be more precisely defined.

508 Nonetheless, application of these schemes leads to sharp expansions of areas with low, but
509 positive PoP (< 0.2) to regions where no precipitation was observed (e.g., Wyoming). Between
510 the two DL schemes, ANN-CSGD produces the most skillful PoP that exhibits the least spatial
511 mismatch with the analysis, and its PoP features the lowest BS among the four sets of products.
512 On the other hand, however, it tends to produce wider expansion of PoP, and suppress the high
513 probability values in regions where precipitation was observed (Fig. 10d). One possible
514 explanation for this apparent tradeoff lies in DL schemes' inclusion of training samples over
515 wider areas where the precipitation amounts are dissimilar to those near the target. While this
516 practice improves the overall skills for CONUS, it introduces diffusion in areal coverage and
517 reduces the sharpness. This phenomenon is analogous to the dilution effect noted in regression
518 literature (Fuller, 1987; Hughes, 1993; Frost and Thompson, 2000; and Jozaghi et al., 2021).

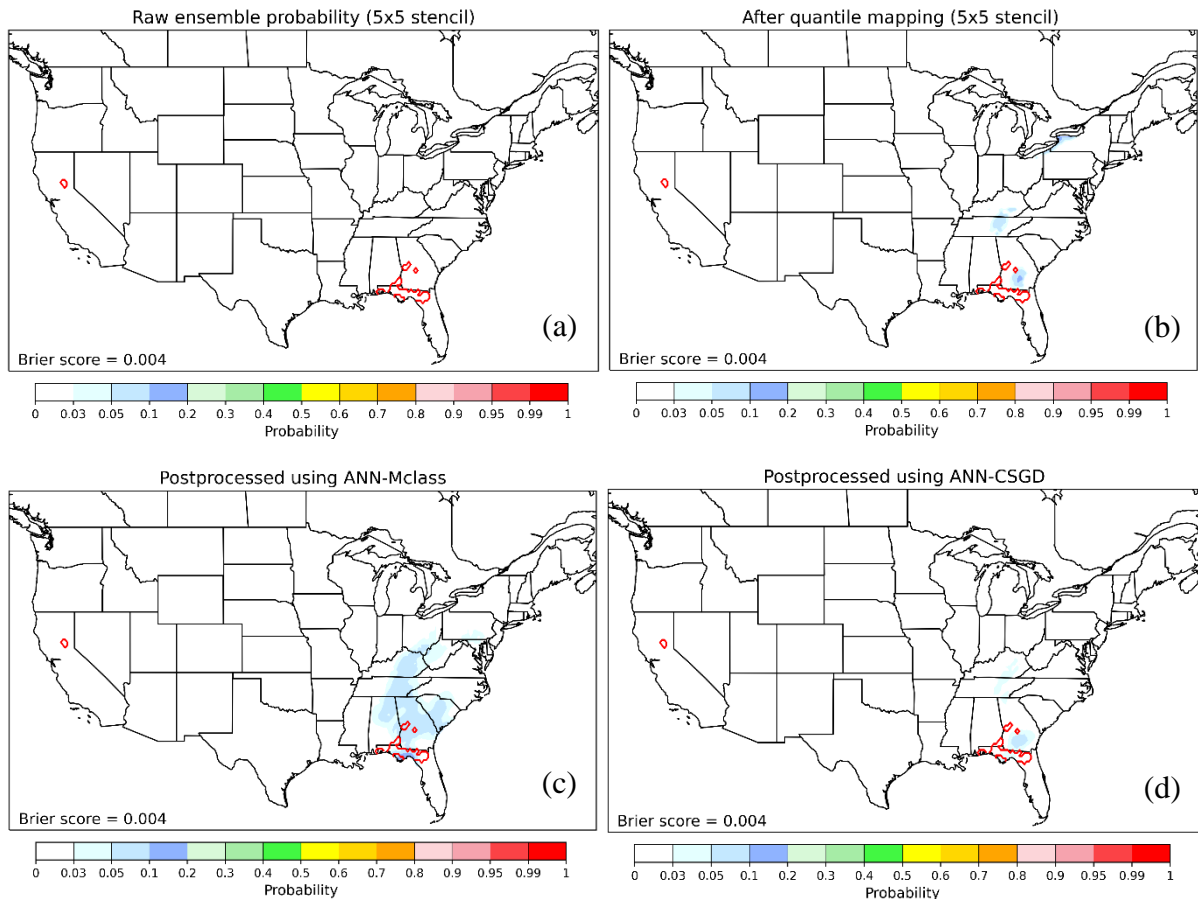
519 At the threshold of 25 mm (P25, Figure 11), raw ensemble fails to detect the precipitation
520 clusters in northern California while all three schemes help recover these. Over this region, P25
521 produced by quantile mapping exhibits the highest locational precision and sharpness, a feature
522 consistent with the earlier observation shown in Fig. 4. By contrast, over the eastern part of the
523 country, quantile mapping broadly degrades the skills – it produces excessively high P25 over
524 broader areas where accumulations per CCPA are much below 25mm. This introduces
525 additional wet biases and contributes to a reduction in overall reliability relative to raw as
526 shown in reliability diagrams (Fig. 7b). Both DL schemes improve the overall accuracy of
527 guidance for CONUS and over major precipitation clusters (Figs. 11b and c). Over northern
528 California, both yield higher P25 that overlap with the observed clusters, but with lower spatial
529 precision. ANN-CSGD, for example, populates the entire northern California with positive
530 P25, grossly exaggerating the areal coverage of rainfall risks. ANN-Mclass fares somewhat
531 better with more subdued areal coverage bias. Along the eastern US, the most notable feature
532 is that both DL schemes perform well in capturing the rainfall risks over the cluster that
533 encompasses parts of Georgia, Alabama, Tennessee, Florida, and South Carolina. The
534 performance of the two techniques is mixed. ANN-Mclass excels by creating higher P25 within
535 the cluster, yet it in the meantime inflates the P25 outside the cluster. By comparison, the P25
536 produced by ANN-CSGD is nearly uniformly lower over the region, either inside or outside
537 the cluster. The BS of the DL-based PQQFs is broadly comparable at this threshold and is
538 slightly better than that for QMAP.



539 Fig. 11. As in Figure 9 except for > 25 mm. Areas inside red contours show that event > 25 mm has been
 540 observed by CCPA.

541 Fig. 12 shows the relative performance of the three schemes at the highest threshold, i.e.,
 542 50mm (P50). Heavy rainfall exceeding the threshold is concentrated over smaller clusters along
 543 the Sierra Nevada, in central Georgia, and over the northern Florida Panhandle. Raw ensemble
 544 is apparently unable to capture any of these clusters. All three schemes create areas with
 545 positive P50. Among these, QMAP and ANN-CSGD perform comparably by creating small
 546 P50 values that marginally overlap with the clusters in central Georgia and Florida Panhandle.
 547 By comparison, ANN-Mclass creates positive P50 over a wider area over the southeast with
 548 more substantial overlap with all clusters in the region. This improved detection, however, is
 549 offset by the false coverage elsewhere. The BS values for all four products are nearly identical,
 550 most likely a result of the limited sample size for computing the metric. The ability of the two
 551 DL schemes in mitigating the overforecast as seen in Fig. 8 is not confirmed by the spatial
 552 verification for possibly the same reason.

553



554 Fig. 12. As in Fig.10 except for > 50 mm. Areas inside red contours show that event > 50 mm has been
 555 observed by CCPA.

556 4. Summary and Conclusions

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

569 To assess the performance of the DL schemes, we designed hindcast experiments to
570 postprocess 24-h accumulated GEFS reforecast data using a rolling training scheme and with
571 previous 60 day's forecast and analyzed data. As the benchmark statistical postprocessing
572 technique we implemented QMAP stencil (Hamill et al. 2017) method that is used to produce
573 probabilistic guidance that populates the National Digital Forecast Database (NDFD; Glahn
574 and Ruth 2003). Note that a key difference between the QMAP approach and the DL schemes
575 concerns the mechanism of augmenting the spatial sampling domain to compensate for the
576 limited time window. QMAP does so by incorporating so-called supplemental locations, i.e.,
577 locations that share similar elevation and topographic facets, and, presumably, may share
578 similar precipitation climatology (Daly et al., 1994). By contrast, the DL schemes leverage data
579 at all grid points within the domain and infusing geographical information including latitude
580 and longitude as ancillary predictors.

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

589 The two DL schemes perform comparably at low-middle thresholds in terms of calibration,
590 but the performance differentials widen at higher thresholds with ANN-CSGD conspicuously
591 outperforming. A major weakness of ANN-Mclass is the lack of reliability of its PQPFs at the
592 highest threshold (50mm/day): when compared to ANN-CSGD, it tends to produce more
593 severe overforecasts and is broadly incapable of enhancing the skills of raw ensemble at this
594 threshold. This underperformance of ANN-Mclass is potentially related to the inadequate
595 number of output categories implemented in the study. Note that the selection of optimal
596 number of output categories is not a straightforward task - as high observed precipitation values
597 are much less frequent than lighter precipitation values, empirical quantiles cannot be easily
598 extended to large amounts without incurring substantial uncertainties. In this regard, ANN-
599 CSGD's explicit use of a parametric distribution proves advantageous as it offers a more
600 consistent way of estimating higher forecast quantiles.

601 Our validation experiments also reveal a distinctive geographic dependence of the relative
602 performance of different schemes. QMAP, while broadly underperforming the DL schemes for
603 the entire CONUS, outperforms the latter competitors along the West Coast and over the Sierra-
604 Nevada. Its overall underperformance is mostly a result of its inability to produce skill gains
605 for much of the central and eastern US. Closer examinations suggest that variations in the skill
606 of the raw ensemble for different precipitation regimes, along with the spatial variability of
607 rainfall brought by these regimes, may have played pivotal roles in shaping the geographic
608 disparity. To elaborate, over the Pacific coast-Sierra Nevada, landfalling Atmospheric River
609 events dominate large precipitation amounts over the region. The GEFS ensemble exhibits
610 good skills in predicting the occurrence of these events as well as the associated geographic
611 distribution of precipitation. QMAP's use of limited, prescribed supplemental locations prove
612 effective in correcting forecast biases, whereas the DL schemes' simultaneous use of samples
613 across locations may have over-expanded the training sample and thereby impaired the
614 robustness of predictor-predictand relationship derived therefrom. By contrast, heavy
615 precipitation over the central and eastern US can arise from a mix of organized convection,
616 frontal systems, as well as tropical cyclones, and predictability of these systems varies both by
617 location and season. The overall skills of GEFS ensemble are low over this region, and spatial
618 displacement errors are a major contributor to the lack of skills. For these locations, the ability
619 of DL schemes in adaptively incorporating forecast-observation pairs over broader areas for
620 training more effectively address the displacement errors. Another potential factor underlying
621 the contrasting performance of QMAP, as the authors postulate, is the degree of similarity
622 among supplemental locations. It is possible that forecast-observation relationships are broadly
623 dissimilar among supplemental locations over central and eastern US as elevations and facets
624 play lesser roles in modulating precipitation climatology. It is also worth pointing out that the
625 gains in calibration for the higher thresholds as achieved by the DL schemes often come at the
626 expense of subdued forecast sharpness – the exceedance probability in regions where
627 precipitation was observed is often lower in the postprocessed guidance produced by these
628 schemes, a feature reminiscent of the findings of Herman and Schumacher (2018). These issues
629 warrant further, more thorough investigations to confirm and illuminate.

630 As DL techniques are evolving rapidly, there are many emerging opportunities for further
631 enhancing the DL postprocessing schemes illustrated in the study. Future research will be
632 directed towards identifying and integrating mechanisms that will allow for i) more effective
633 use of geographic information in the networks. Location embedding can be used to project

634 discrete pairs of each latitude and longitude values onto a continuous, larger vector of latent
635 inputs using IDs specific for each CCPA grid. Embedding permit the model to optimize grid
636 IDs representations by the training process and potentially help better capture local
637 characteristics as shown in past postprocessing studies (see, Schulz and Lerch 2021; Chapman
638 et al. 2021). Incorporating auxiliary predictors such as dot product of moisture advection with
639 terrain gradient, total column precipitable water and CAPE might be another way of getting
640 more terrain-related detail; ii) more efficient modeling of complex-arbitrary nonlinear
641 predictor-predictand relationships such as use one dimensional convolution or attention layers
642 (Collobert et al. 2011; Vaswani et al. 2017; Delvin et al. 2018) on top of an embedding layer
643 to better capture predictor interactions with each other and with spatial features, and iii) more
644 robust training of networks to avoiding overfitting. This work used a rather basic but popular
645 and effective regularization technique to stop training, which is based on validation dataset loss
646 (Goodfellow et al. 2016). It is possible that additional gains can be realized by simply
647 increasing the validation window with lead time and introducing additional regularization
648 parameters.

649

650 *Acknowledgments.*

651 The authors would like to acknowledge financial support for the first and second authors
652 over the years provided by the faculty startup package for Dr. Yu Zhang from UT Arlington,
653 NOAA Grant NA18OAR4590370-01, NSF Grant 1909367, and Texas Water Development
654 Board Contract 1800012276. The work benefits from input from a large number of individuals
655 within the National Weather Service, including Jeff Craven, David Rudack, and Eric Engle of
656 the National Blended Model team, Kris Lander at the West Gulf River Forecast Center, Bruce
657 Veenhuis at the Weather Prediction Center, and John J. Brost at the Operational Proving
658 Ground. We would also like to thank Kevin He from the California Department of Water
659 Resources for stimulating discussions that helped shape the work, and to members of the
660 Unified Forecast System Steering Committee for critiques and suggestions.

661

662 *Data Availability Statement.*

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

666

667

APPENDIX

668

Implementation details

669 We used Python (Python Software Foundation 2018), R (R core team 2017) and Fortran in
670 this project. Specifically, R was used for initial data processing. We implemented our deep
671 learning codes in python using Google's platform, Tensorflow (Abadi et al. 2016) and Keras
672 API (Chollet et al. 2015). For quantile mapping a research version was implemented using
673 python. Fortran routines to generate supplemental locations were provided by Dr. Tom Hamill
674 from NOAA PSL and were tailored to our setting. Other computations (verification, graphics,
675 etc.) were performed with python.

676

677

REFERENCES

678 Abadi, M., and Coauthors, 2016: Tensorflow: A system for largescale machine learning.

679 Proc. USENIX 12th Symp. On Operating Systems Design and Implementation,

680 Savannah, GA, Advanced Computing Systems Association, 265–283,

681 <https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf>.

682 Baran, S., and D. Nemoda, 2016: Censored and shifted gamma distribution based EMOS

683 model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27, 280–

684 292, <https://doi.org/10.1002/env.2391>.

685 Baran, S., and S. Lerch, 2018: Combining predictive distributions for statistical post-

686 processing of ensemble forecasts. *Int. J. Forecast.*, 34, 477–496,

687 <https://doi.org/10.1016/j.ijforecast.2018.01.005>.

688 Baran, S., & Baran, Á. 2021: Calibration of wind speed ensemble forecasts for power

689 generation. ArXiv Preprint ArXiv:2104.14910, <https://arxiv.org/abs/2104.14910>.

690 Bremnes, J. B., 2020: Ensemble postprocessing using quantile function regression based on

691 neural networks and Bernstein polynomials. *Mon. Wea. Rev.*, 148, 403–414,

692 <https://doi.org/10.1175/MWR-D-19-0227.1>.

693 Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea.*
694 *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).

695 Bröcker, J. and L.A. Smith, 2007: Increasing the Reliability of Reliability Diagrams. *Wea.*
696 *Forecasting*, 22, 651–661, <https://doi.org/10.1175/WAF993.1>.

697 Brown, J. D., L. Wu, M. He, S. Regonda, H. Lee, and D.J. Seo, 2014a: Verification of
698 temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic
699 Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification.
700 *Hydrol*, 519, 2869–2889, <https://doi.org/10.1016/j.jhydrol.2014.05.028>.

701 Chapman, W. E., Delle Monache, L., Alessandrini, S., Subramanian, A. C., Ralph, F. M.,
702 Xie, S., Lerch, S., & Hayatbini, N. 2021: Probabilistic Predictions from Deterministic
703 Atmospheric River Forecasts with Deep Learning, *Mon. Wea. Rev.*, (published online
704 ahead of print 2021), <https://doi.org/10.1175/MWR-D-21-0106.1>.

705 Chollet, F., and Coauthors, 2015: Keras: The Python Deep Learning library. Accessed 2020,
706 <https://keras.io>.

707 Cloke, H. I., and F. Pappenberger, 2009: Ensemble flood forecast: A review. *J.*
708 *Hydrol.*, 375, 613–626. <https://doi.org/10.1016/j.jhydrol.2009.06.005>.

709 Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, 2011: Natural
710 Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*,
711 12:2493-2537. Available online at: <https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf>.

713 Darbandsari, P., and P. Coulibaly, 2022: Assessing Entropy-based Bayesian Model
714 Averaging Method for Probabilistic Precipitation Forecasting, *Journal of*
715 *Hydrometeorology* (published online ahead of print 2022). [https://doi.org/10.1175/JHM-](https://doi.org/10.1175/JHM-D-21-0086.1)
716 [D-21-0086.1](https://doi.org/10.1175/JHM-D-21-0086.1)

717 Devlin, J., M. W. Chang, K. Lee, and K. Toutanova, 2018: Bert: Pre-training of deep
718 bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
719 <https://arxiv.org/abs/1810.04805>.

720 Frost, C., and S. G. Thompson, 2000: Correcting for regression dilution bias: Comparison of
721 methods for a single predictor variable. *J. Roy. Stat. Soc.*, 163A, 173–190,
722 <https://doi.org/10.1111/1467-985X.00164>

723 Fuller, W. A., 1987: Measurement Error Models. Wiley, 440 pp.

724 Ghazvinian, M., Y. Zhang, and D. J. Seo, 2020: A Nonhomogeneous Regression-Based
725 Statistical Postprocessing Scheme for Generating Probabilistic Quantitative Precipitation
726 Forecast. *J. Hydrometeor.*, 21, 2275–2291, <https://doi.org/10.1175/JHM-D-20-0019.1>.

727 Ghazvinian, M., Zhang, Y., Seo D.-J., He, M., Fernando, N., 2021: A novel hybrid artificial
728 neural network - Parametric scheme for postprocessing medium-range precipitation
729 forecasts. *Advances in Water Resources*, Volume 151,103907,
730 <https://doi.org/10.1016/j.advwatres.2021.103907>.

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

734 Goodfellow, I., Y. Bengio, and A. Courville, 2016: Deep Learning. MIT Press, 775 pp.

735 Hamill, T.M., J.S. Whitaker, and X. Wei, 2004: Ensemble Reforecasting: Improving
736 Medium-Range Forecast Skill Using Retrospective Forecasts. *Mon. Wea. Rev.*, 132,
737 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).

738 Hamill, T.M., G.T. Bates, J.S. Whitaker, D.R. Murray, M. Fiorino, T.J. Galarneau, Y. Zhu,
739 and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble
740 Reforecast Dataset. *Bull. Amer. Meteor. Soc.*, 94, 1553–1565,
741 <https://doi.org/10.1175/BAMS-D-12-00014.1>.

742 Hamill, T. M., M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation
743 forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon.*
744 *Wea. Rev.*, 143, 3300–3309, <https://doi.org/10.1175/MWR-D-15-0004.1>.

745 Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts
746 based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, 134, 3209–3229,
747 <https://doi.org/10.1175/MWR3237.1>.

748 Hamill, T. M., Engle, E., Myrick, D., Peroutka, M., Finan, C., & Scheuerer, M. 2017: The
749 U.S. National Blend of Models for Statistical Postprocessing of Probability of
750 Precipitation and Deterministic Precipitation Amount, *Mon. Wea. Rev.*, 145(9), 3441-
751 3463, <https://doi.org/10.1175/MWR-D-16-0331.1>.

752 Hamill, T. M., & Scheuerer, M. 2018: Probabilistic Precipitation Forecast Postprocessing
753 Using Quantile Mapping and Rank-Weighted Best-Member Dressing, *Mon. Wea. Rev.*,
754 146(12), 4079-4098, <https://doi.org/10.1175/MWR-D-18-0147.1>.

755 Hamill, T. M., 2018: Practical Aspects of Statistical Postprocessing. In S. Vannitsem, D. S.
756 Wilks, & J. Messner (Eds.), *Statistical postprocessing of ensemble forecasts*,
757 <https://doi.org/10.1016/B978-0-12-812372-0.00007-8>.

758 Herman, G. R., & Schumacher, R. S. 2018: Money Doesn't Grow on Trees, but Forecasts Do:
759 Forecasting Extreme Precipitation with Random Forests, *Monthly Weather Review*,
760 146(5), 1571-1600, <https://doi.org/10.1175/MWR-D-17-0250.1>.

761 Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales:
762 Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, 15,
763 2542–2557, <https://doi.org/10.1175/JHM-D-11-0140.1>.

764 Hughes, M. D., 1993: Regression dilution in the proportional hazards model. *Biometrics*, 49,
765 1056–1066, <https://doi.org/10.2307/2532247>.

766 Ioffe, S. and Szegedy, C. 2015: Batch normalization: Accelerating deep network training by
767 reducing internal covariate shift. In *Proceedings of the 32nd International Conference on*
768 *Machine Learning*, pages 448–456. <http://proceedings.mlr.press/v37/ioffe15.pdf>.

769 Jolliffe, I. T., and D. B. Stephenson, Eds., 2012: *Forecast Verification: A Practitioner's Guide*
770 *in Atmospheric Science*. 2nd ed. John Wiley & Sons, 292 pp.,
771 doi:10.1002/9781119960003.

772 Jozaghi, A., and Coauthors, 2021: Multi-model streamflow prediction using conditional bias-
773 penalized multiple linear regression. *Stoch Environ Res Risk Assess* 35, 2355–2373.
774 <https://doi.org/10.1007/s00477-021-02048-3>.

775 Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. *Third Int.*
776 *Conf. for Learning Representations*, San Diego, CA, ICLR, 1–15,
777 <https://arxiv.org/abs/1412.6980>.

778 Krzysztofowicz, R., 2008: January. Bayesian Processor of Ensemble: concept and
779 development. In *Proc. 19th Conf. Probability and Statistics in the Atmospheric*
780 *Sciences (Vol. 4)*. Seattle: American Meteorological Society.

781 Li, W., and Coauthors, 2022: Convolutional neural network-based statistical post-processing
782 of ensemble precipitation forecasts, *Journal of Hydrology*, Volume 605, 127301,
783 <https://doi.org/10.1016/j.jhydrol.2021.127301>.

784 Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability
785 distributions. *Manage. Sci.*, 22, 1087–1096, <https://doi.org/10.1287/mnsc.22.10.1087>.

786 Pappenberger, F., and R. Buizza, 2009: The skill of ECMWF precipitation and temperature
787 predictions in the Danube basin as forcings of hydrological models. *Wea.*
788 *Forecasting*, 24, 749–766, <https://doi.org/10.1175/2008WAF2222120.1>.

789 Python Software Foundation, 2018: Python Language Reference, version 3.7. Available at
790 <http://www.python.org>.

791 R Core Team, 2017: R: A language and environment for statistical computing. R Foundation
792 for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>.

793 Rasp, S., and S. Lerch, 2018: Neural Networks for Postprocessing Ensemble Weather
794 Forecasts. *Mon. Wea. Rev.*, 146, 3885–3900, <https://doi.org/10.1175/MWR-D-18-0187.1>.

795 Reggiani, P., and O. Boyko, 2019: A Bayesian Processor of Uncertainty for Precipitation
796 Forecasting Using Multiple Predictors and Censoring. *Mon. Wea. Rev.*, 147, 4367–4387,
797 <https://doi.org/10.1175/MWR-D-19-0066.1>.

798 Robertson, D. E., D. L. Shrestha, and Q. J. Wang, 2013: Post-processing rainfall forecasts
799 from numerical weather prediction models for short-term streamflow forecasting. *Hydrol.*
800 *Earth Syst. Sci.*, 17, 3587–3603, <https://doi.org/10.5194/hess-17-3587-2013>.

801 Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation
802 forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, 143, 4578–
803 4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.

804 Scheuerer, M., T. M. Hamill, B. Whitin, M. He, and A. Henkel, 2017: A method for
805 preferential selection of dates in the Schaake shuffle approach to constructing
806 spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.*, 53,
807 3029–3046, <https://doi.org/10.1002/2016WR020133>.

808 Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using Artificial
809 Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over

810 California. *Mon. Wea. Rev.*, 148, 3489–3506, <https://doi.org/10.1175/MWR-D-20->
811 [0096.1](https://doi.org/10.1175/MWR-D-20-0096.1).

812 Schulz, B., and Lerch, S. 2021: Machine learning methods for postprocessing ensemble
813 forecasts of wind gusts: A systematic comparison. *ArXiv Preprint ArXiv:2106.09512*.
814 <https://arxiv.org/abs/2106.09512>.

815 Sloughter, JM., AE. Raftery, T. Gneiting, and C Fraley, 2007: Probabilistic quantitative
816 precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, 135:3209 –
817 3220, <https://doi.org/10.1175/MWR3441.1>.

818 Taillardat, M., A. Fougères, P. Naveau, and O. Mestre, 2019: Forest-Based and
819 Semiparametric Methods for the Postprocessing of Rainfall Ensemble Forecasting. *Wea.*
820 *Forecasting*, 34, 617–634, <https://doi.org/10.1175/WAF-D-18-0149.1>.

821 Valdez, E. S., Anctil, F., and Ramos, M.-H. 2021: Choosing between post-processing
822 precipitation forecasts or chaining several uncertainty quantification tools in hydrological
823 forecasting systems, *Hydrol. Earth Syst. Sci. Discuss.* [https://doi.org/10.5194/hess-2021-](https://doi.org/10.5194/hess-2021-391)
824 [391](https://doi.org/10.5194/hess-2021-391).

825 Vannitsem, S., and Coauthors, 2020: Statistical Postprocessing for Weather Forecasts:
826 Review, Challenges and Avenues in a Big Data World. *Bulletin of the American*
827 *Meteorological Society*, 102(3), E681-E699, <https://doi.org/10.1175/BAMS-D-19-0308.1>.

828 Vaswani, A., and coauthors, 2017: Attention is all you need. In *Advances in neural*
829 *information processing systems* (pp. 5998-6008). Available online at
830 [https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-](https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
831 [Paper.pdf](https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)

832 Veldkamp, S., Whan, K., Dirksen, S. and Schmeits, M. 2021: Statistical postprocessing of
833 wind speed forecasts using convolutional neural networks. *Mon. Wea. Rev.*, 149, 1141–
834 1152, <https://doi.org/10.1175/MWR-D-20-0219.1>.

835 Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution
836 MOS forecasts. *Meteor. Appl.*, 6,361–368, <https://doi.org/10.1002/met.134>.

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

839 WPC, 2019: 2019 Flash Flood and Intense Rainfall Experiment: Findings and Results,
840 National Centers for Environmental Prediction, Weather Prediction Center,
841 https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2019_FfaIR.pdf.

842 WPC, 2020: 2020 Flash Flood and Intense Rainfall Experiment: Findings and Results,
843 National Centers for Environmental Prediction, Weather Prediction Center,
844 https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2020_FFaIR_Experiment_Nov13.pdf.

845 Wu, L., D.J. Seo, J. Demargne, J. Brown, S. Cong, and J. Schaake, 2011: Generation of
846 ensemble precipitation forecast from single-valued quantitative precipitation forecast for
847 hydrologic ensemble prediction. *J. Hydrol.*, 399, 281–298,
848 <https://doi.org/10.1016/j.jhydrol.2011.01.013>.

849 Zhang, Y., L. Wu, M. Scheuerer, J. Schaake, and C. Kongoli, 2017: Comparison of
850 Probabilistic Quantitative Precipitation Forecasts from Two Postprocessing Mechanisms.
851 *J. Hydrometeor.*, 18, 2873–2891, <https://doi.org/10.1175/JHM-D-16-0293.1>.

852



Click here to access/download
Supplemental Material
DL_supplemental.docx