A Novel Hybrid Artificial Neural Network - Parametric Scheme for Postprocessing Medium-Range Precipitation Forecasts

by

Mohammadvaghef Ghazvinian [a], Yu Zhang [a], Dong-Jun Seo [a], Minxue He [b], Nelun Fernando [c]

[a] Department of Civil Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA

[b] California Department of Water Resources, 1416 9th Street, Sacramento, CA 95814, USA

[c] Texas Water Development Board, 1700 North Congress Avenue, Austin, TX 78701, USA

Revised Version: March 25, 2021

Manuscript submitted to the Advances in Water Resources

Corresponding author: Mohammadvaghef Ghazvinian

Department of Civil Engineering, The university of Texas at Arlington,
427 Nedderman Hall-416 Yates Street, Arlington, TX 76019, USA
Email address: mohammadvaghef.ghazvinian@mavs.uta.edu

Abstract

Many present-day statistical schemes for postprocessing weather forecasts, in particular precipitation forecasts, rely on calibration using prescribed statistical models to relate forecast statistics to distributional parameters. The efficacy of such schemes is often constrained not only by prescribed predictor-predictand relation, but also by arbitrary choices of temporal window and lead time range for training. To address this limitation, we propose an end-to-end, computationally efficient hybrid postprocessing scheme capable of producing full predictive distributions of precipitation accumulation without explicit stratification of forecast-observation pairs by forecast lead time and season. The proposed framework uses the censored, shifted gamma distribution (CSGD) as the predictive distribution but uses an artificial neural network (ANN) to estimate the distributional parameters of CSGD through a unified approach. This approach, referred to as ANN-CSGD, allows for simultaneous estimation of distributional parameters over multiple lead times and seasons in a single model by incorporating the latter variables as predictors to the ANN. We test our proposed ANN-CSGD model for postprocessing of ensemble mean forecasts of 24-h precipitation totals over selected river basins in California, at one- to seven-day lead times, from the Global Ensemble Forecast System (GEFS). The probabilistic quantitative precipitation forecasts (PQPFs) from the ANN-CSGD, are more skillful overall than those from the benchmark CSGD and the Mixed-type meta-Gaussian distribution (MMGD) models. The ANN-CSGD PQPFs highly improve the performance of those from CSGD in predicting the probability of precipitation (PoP) and are also much sharper and reliable at higher precipitation thresholds. We demonstrate how the hybrid approach, by using the entire available training data and its modified formulation, efficiently represents interactions between GEFS forecasts and season/lead times, thus leading to enhanced predictive performance.

Keywords: Statistical postprocessing; Artificial neural networks; Probabilistic quantitative precipitation forecast; Predictive distribution

1. Introduction

Statistical postprocessing techniques are increasingly used to improve the reliability and skill of real time probabilistic quantitative precipitation forecasts (PQPFs) produced by numerical weather prediction (NWP) models. Broadly speaking, these techniques can be categorized as nonparametric and parametric ones. A prominent example of the former is the Analog approach (Hamill and Whitaker 2006; Hamill et al. 2015). The parametric techniques rely on prescribed parametric forms of conditional (predictive), joint and marginal distributions, and employ various techniques ranging from regression to the method of moments, and their variants, for estimating distributional parameters. Many of the modern parametric approaches fall under the broad umbrella of Ensemble Model Output Statistics (EMOS; Gneiting et al. 2005), also known as nonhomogeneous regression. As the name implies, the EMOS approaches use prescribed predictive distributions and relate distributional parameters to ensemble statistics through a set of regression equations (Scheuerer and Hamill 2015; Zhang et al. 2017; Stauffer et al. 2017).

The extent to which postprocessing techniques have improved forecast skill has varied in practice (Li et al. 2017; Wilks 2018; Vannitsem et al. 2020). There are several common limitations in postprocessing methods adopted to date. Among the frequently cited are the inflexible and subjective way of selecting predictors, structural rigidity that makes it difficult to integrate ancillary predictors, and the ad hoc way of determining spatial-temporal training domains (see related discussions in Rasp and Lerch 2018). The advent of machine learning techniques offers many new opportunities to address these limitations. Relative to the parametric approaches, EMOS techniques included, some of the recent machine learning techniques offer flexibility in identifying predictors, in integrating ancillary information, and in capturing complex, nonlinear predictor-predictand relationships that are difficult to characterize parametrically (see, e.g., Taillardat et al. 2019). Particularly promising are the various artificial neural networks (ANNs) which have been known for their ability to model nonlinear dependencies. Recent years have seen an explosion of ANN-based prediction paradigms (Liu et al. 2016; Brenowitz and Bretherton 2018; Gentine et al. 2018; Rasp et al. 2018; Chapman et al. 2019; Cloud et al. 2019; Gagne et al. 2019; Lagerquist et al. 2019). Yet, the use of these techniques in the context of postprocessing remains relatively limited. Rasp and Lerch (2018) is perhaps the first attempt of this nature. The authors explored a hybrid scheme that retains a parametric form of the predictive distribution of 2-m temperature but relies on ANNs to estimate the distribution parameters from the ensemble statistics of 2-m temperature as well as ancillary variables. Scheuerer et al. (2020), in a similar vein, developed an ANN-based scheme for producing 7-day accumulated PQPFs at subseasonal range (2–4 weeks) from NWP ensemble forecasts, and showed that the PQPFs thus generated broadly outperforms climatology. Other studies of note include Bremnes (2020) where ANN was used for postprocessing wind speed forecasts. Collectively, these studies indicate that embedding local information and incorporating ancillary forecast variables can lead to larger improvements in forecast skills. They further suggest that ANN models, contrary to the common perception of being black boxes, can help uncover, and offer physical insights to the meteorological processes that underpin the links between predictors and predictands.

Inspired by the successes of recent ANN-based postprocessing approaches, and motivated by the broader need for improving the skill of PQPF while circumventing limitations inherent in existing EMOS schemes, we propose a hybrid ANN-nonhomogeneous regression-based scheme capable of postprocessing precipitation forecasts at multiple lead times and seasons in a unified way. The proposed scheme retains the parametric form of the predictive distribution of precipitation proposed by Scheuerer and Hamill (2015) and Baran and Nemoda (2016), but departs from the conventional EMOS by using ANNs to relate NWP forecasts to the distributional parameters. The potential advantages of the proposed scheme, which we will henceforth refer to as ANN-CSGD are three-fold. First, this scheme does not require an explicit prescription of predictor-predictand relationships as is currently done in EMOS models - it can discover and integrate arbitrary nonlinear relationships through training. Second, the training of the model can be done using the entire data archive and thereby obviate the need for explicit treatment of lead time-based and seasonally varying NWP forecast errors. Third, it can account for seasonal variations in the interaction between NWP forecasts and temporal predictors.

In this paper we describe and evaluate the proposed scheme which relies only on the ensemble mean of NWP forecasts as the major predictor. The evaluation is conducted for sub-basins within three selected river basins in California. The proposed scheme is applied to postprocess Global Ensemble Forecast System (GEFS; Hamill et al. 2013) precipitation reforecasts along with two benchmark schemes. The first is the single predictor version of the censored, shifted gamma distribution (CSGD; Scheuerer and Hamill 2015). The second is the Mixed-type Mata-Gaussian Distribution (MMGD; Wu et al. 2011), which has been the standard method in the U.S. National Weather Service (NWS) Hydrologic Ensemble Forecast Service (HEFS; Demargne et al. 2014). Our overarching hypothesis is that the flexibility accorded by the ANN-based model in establishing complex predictor-distributional parameter relationships, in determining temporal training windows, and in lumping forecasts for different lead times, will help the proposed scheme attain superior predictive performance relative to the benchmarks.

The reminder of this paper is organized as follows. Section 2 describes the proposed ANN-CSGD scheme as well as the benchmark methods, data, and experimental setup. Section 3 presents the outcomes of the experiments and section 4 summarizes the findings and discusses future possible extensions.

2. Materials and methods

2.1. Proposed model

The censored, shifted gamma distribution (CSGD) introduced by Scheuerer and Hamill (2015), has been a popular choice to represent the right skewed, mixed-type dichotomous-continuous nature of the predictive distribution of precipitation (Scheuerer and Hamill 2015; Baran and Nemoda 2016; Zhang et al. 2017; Scheuerer et al. 2020). Let $F_{k,\theta}$ denote the cumulative distribution function (CDF) of the gamma distribution with shape parameter $k > 0$ and scale parameter $\theta > 0$. The CDF at realized precipitation value y, and quantile functions of CSGD for any $0 \leq p < 1$ are defined by (Scheuerer and Hamill 2015; Baran and Nemoda 2016):

$$F^0{}_{k,\theta,\delta}(y) = \begin{cases} F_{k,\theta}(y-\delta), & y \geq 0 \\ 0, & y < 0 \end{cases} \tag{1}$$

$$q_p = max\left[0, \delta + F_{k,\theta}^{-1}(p)\right] \tag{2}$$

where the additional parameter, $\delta < 0$ shifts the gamma distribution to the negative values. To form the CSGD, the shifted gamma distribution is left censored at zero by assigning the mass probability $F_{k,\theta}(-\delta)$ to the origin to account for non-negativity of precipitation amounts. To relate the mean $\mu = k\theta$, standard deviation $\sigma = \sqrt{k}\theta$, and shift parameter δ of predictive CSGDs to the predictors, we propose a fully connected (dense) feed forward neural network where each node receives a linear combination of weighted outputs from nodes in the previous layer, adjusts it by adding a bias quantity, and applies an activation function to the result. Our proposed ANN-CSGD structure (Fig. 1) consists of the following elements:

- Input layer, where covariates are introduced to the network.
- One hidden layer; we use the exponential linear unit (ELU) with $\alpha = 1$ as the activation function to introduce nonlinearity to the network

$$f(x) = \begin{cases} x, & x > 0 \\ \alpha[\exp(x) - 1], & x \leq 0 \end{cases} \qquad (3)$$

  ELUs are known to provide more precise and faster learning compared to the other activation functions in deep learning experiments (see, Clevert et al. 2015).
- Layer normalization (Ba et al. 2016) which normalizes each sample output from hidden nodes to maintain the mean and standard deviation of node outputs within each example close to 0 and 1, respectively. Recent studies (see, e.g., Xu et al. 2019) show that Layer normalization helps stabilize the training process by enabling smoother gradients and yields faster training convergence.
- Output layer with a linear activation function. We set three CSGD parameters as functions of the network outputs O_i to constrain the values of these parameters to reasonable ranges (i.e., $\mu, \sigma > 0$ and $\delta < 0$). Therefore, we set $\delta = -sqrt({O_1}^2)$, $\mu = \exp(O_2)$, and $\sigma = \exp(O_3)$. These additional functions can be interpreted as inverse link functions used in conventional distributional regression or generalized additive models for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos 2005) (see, also, Cannon 2012; Rasp and Lerch 2018).

We incorporate the ensemble mean forecast, forecast lead time (1 to 7 days), and month of the year of the verifying observations (1 to 12) as predictors to the ANN. Using the latter two predictors enables us to train a single model to postprocess forecasts from multiple lead times and months. Lead time values are normalized by dividing each quantity by the maximum value ($i.e., day/7$). To account for seasonal cycle, we use the cosine term $[cos(2\pi(month - 1)/12)]$ to both introduce the cyclical nature of the month of the year to the network and to enforce the network to encode the annual cycle of precipitation over the study area (see, Liu et al. 2018; Scheuerer et al. 2017).

We retain the average value of continuous ranked probability score (CRPS) of predictive CSGDs as the loss function for training the weights and biases of the ANN-CSGD. The ANN is trained by minimizing the CRPS computed using collocated and coincidental forecast-observation pairs over training data (see the appendix B for the mathematical definition of CRPS)

$$\text{CRPS} = \frac{1}{N}\sum_{i=1}^{N} \text{crps}\left(F_{k_i\theta_i\delta_i}, y_i\right) \tag{4}$$

The analytical expression of CRPS for a paired CSGD predictive distribution and verifying observation was proposed by Scheuerer and Hamill (2015). Similarly, we implement

$$\text{crps}\left(F_{k_i\theta_i\delta_i}, y_i\right) = (y_i - \delta_i)\left[2F_{k_i,\theta_i}(y_i - \delta_i) - 1\right] - \frac{\theta_i k_i}{\pi} B\left(\frac{1}{2}, k_i + \frac{1}{2}\right)\left[1 - F_{2k_i,\theta_i}(-2\delta_i)\right] + \theta_i k_i\left[1 + 2F_{k_i,\theta_i}(-\delta)F_{k_i+1,\theta_i}(-\delta_i) - F_{k_i,\theta_i}(-\delta_i)^2 - 2F_{k_i+1,\theta_i}(y_i - \delta_i)\right] + \delta F_{k,\theta}(-\delta)^2 \tag{5}$$

where $B(0,0)$ is the beta function, and $(k_i, \theta_i, \delta_i)$ are three parameters of *ith* predictive CSGD with y_i being the corresponding verifying observation. To minimize the loss function, we use the Adam stochastic gradient descent-based optimization algorithm (Kingma and Ba 2014) and update model parameters based on small batches randomly sampled from the training dataset. One major challenge in applying ANNs is to constrain the complexity of the model while attaining optimal predictions. Overfitting can occur if a very complex structure is used. Several regularization techniques to reduce generalization errors in ANNs are available as reviewed by

Goodfellow et al. (2016). Among them, we use early stopping, which is one of the most popular and widely used regularization techniques in ANNs.

In our work, we leave 20% of the available training data as the validation set and do not include them in training process. This practice enables us to reduce overfitting by monitoring the average loss value over the validation set while we train the model, and return the best possible training parameters (weights and biases) at the time when the lowest CRPS for the validation set is achieved. We terminate training when no further decrease in validation set loss is seen after 15 iterations through all training batches or the entire training data (epochs), with up to 1000 epochs.

We train ANNs using the previously described process, with all possible combinations of different settings, using the early stopping technique for the following hyperparameters

- *Number of nodes in the hidden layer:* $\{5,10,15\}$
- *Batch size:* $\{2048, 4096, 8192\}$
- *Learning rate of the Adam optimization algorithm:* $\{0.01, 0.005\}$

All networks are trained with the same random number generator (seed) and are evaluated based on the average loss value in the validation set. The ANN configuration with the lowest validation loss is chosen for out-of-sample predictions. Individual tested ANNs have $O\{7n+3\}$ trainable parameters where n refers to the number of nodes in the hidden layer. We used a simple (non-trained) layer as the normalization layer. Our assessments showed that training Layer normalization parameters (beta and gamma) does not yield significant improvement over the non-trained one and possibly increases the risk of overfitting due to the increased number of overall network parameters.

2.2. Benchmark models

2.2.1. CSGD

To generate postprocessed precipitation forecasts at a given location, for each forecast lead time and month of the year, Scheuerer and Hamill (2015), first fit three climatological CSGD parameters (μ_{cl}, $\sigma_{cl}$and δ_{cl}) to locally observed training precipitation data using a 91-day temporal window centered around the 15th of each month. In the second step these parameters are included in nonlinear, nonhomogeneous regression equations to relate monthly parameters of predictive CSGDs to statistics of spatially smoothed ensemble of forecasts.

In this study, we use the regression equations that incorporate only the ensemble mean:

$$\mu = \mu_{cl}/a_1\, log\{1 + [(\exp(a_1) - 1)(a_2 + a_3\, \overline{f}/\overline{f_{cl}})]\} \tag{6}$$

$$\sigma = a_4 \sigma_{cl} \sqrt{\mu/\mu_{cl}} \tag{7}$$

$$\delta = \delta_{cl} \tag{8}$$

where $\overline{f}$ and $\overline{f_{cl}}$ correspond to the raw ensemble mean forecasts and their climatological mean in training data, respectively. In the version of CSGD described in Scheuerer and Hamill (2015), the predictive shift parameter δ is kept identical to the climatological shift to ensure that the predictive CSGD reverts to climatology as a limiting case when the forecast becomes less skillful (e.g., at longer lead times) (see related discussion in Scheuerer and Hamill 2015).

The four regression coefficients a_1, a_2, a_3, a_4 are estimated by minimizing the CRPS using the closed form expression proposed by Scheuerer and Hamill (2015) (see sec. 2.1) as a function of CSGD parameters over training data.

Past studies (Scheuerer and Hamill 2015; Baran and Nemoda 2016; Zhang et al. 2017; Baran and Lerch 2018; Taillardat et al. 2019) show that CSGD method and its variants perform well in comparison with other modern postprocessing techniques. Recent exploratory analyses (see, Ghazvinian et al. 2020, Fig. 1) showed that the climatological CSGD shift parameter, derived by CRPS minimization approach, tends to be inflated and this leads to an underestimation of a probability of precipitation (PoP). This bias directly affected the performance of predictive CSGD, primarily in predicting PoP and, to a degree, the predicted magnitude of precipitation. This was particularly evident at shorter lead times and in rainy seasons where the predictive distribution of precipitation deviates widely from climatology.

2.2.2. MMGD

The MMGD (Herr and Krzysztofowicz 2005; Wu et al. 2011) was developed by the U.S. NWS as a component of the Meteorological Ensemble Forecast Processor (MEFP) of the operational HEFS (Demargne et al. 2014). This mechanism is routinely used to generate calibrated PQPF from single-valued precipitation forecasts (ensemble mean) at river basin scales and at temporal aggregation scales ranging from 6-h to 3-months and for lead times up to 9-months (Wu et al. 2018; Demargne et al. 2014). In contrast to the CSGD, where PoP and the probability of magnitude of precipitation are estimated using the same predictive distribution, MMGD uses a Bayesian approach to break down the predictive distribution to explicitly account for the dichotomous-continuous nature of precipitation.

Let X and Y denote the random variables of a single-valued quantitative precipitation forecast and the observed precipitation amount, respectively. The conditional distributions of observed precipitation, given a current forecast of no precipitation and positive precipitation, are given as follows (details of this derivation can be found in Wu et al. 2011; Ghazvinian et al. 2020):

$$\begin{aligned} F_{Y|X}(y|x, x=0) &= P(Y=0|X=0) + P(0 < Y \leq y|X=0) \\ &= a + (1-a)G_Y(y) \end{aligned} \quad (9)$$

$$F_{Y|X}(y|x, x > 0) = P(Y \leq y|X = x, X > 0) \tag{10}$$
$$= c(x) + (1 - c(x))D_{Y|X}(y|x)$$

where a and $c(x)$ represent mass probabilities of observed precipitation being equal to zero, and are combined with the continuous conditional distributions $G_Y(y) = P(Y \leq y|X = 0, Y > 0)$ and $D_{Y|X}(y|x) = P(Y \leq y|X = x, X > 0, Y > 0)$ to construct the predictive distributions. To estimate $D_{Y|X}(y|x)$, its marginal continuous variates $[X|X > 0, Y > 0]$ and $[Y|X > 0, Y > 0]$ undergo normal quantile transformation (NQT), yielding standard normal variates $U = \Phi^{-1}[D_X(x)]$ and $V = \Phi^{-1}[D_Y(y)]$. Following the meta-Gaussian distribution theorem of Kelly and Krzysztofowicz (1997), $D_{Y|X}(y|x)$ assumes the following form

$$D_{Y|X(y|x)} = \Phi\left[\frac{\Phi^{-1}[D_Y(y)] - \rho\Phi^{-1}[D_X(x)]}{\sqrt{1-\rho^2}}\right] \tag{11}$$

where $\Phi(\,)$ and $\Phi^{-1}(\,)$ denote the standard normal CDF and quantile function of standard normal distribution, respectively; and ρ is the Pearson's product correlation coefficient between $U$and V.

The performance of MMGD has been evaluated in a number of studies (see, e.g., Wu et al. 2011; Brown et al. 2014a; Demargne et al. 2014; Kim et al. 2018; Seo et al. 2015; Ghazvinian et al. 2019). While conclusions indicate that overall, MMGD produces reliable PQPFs and is capable of preserving the skill in the raw forecast, its PQPFs underestimate heavy-to-extreme precipitation amounts (low reliability for higher thresholds). The latter finding was also corroborated by Zhang et al. (2017), where the authors compared the performances of MMGD and CSGD over the Mid-Atlantic region in U.S. Their results pointed to the superior performance of CSGD. In that study, CSGD's ability to ingest additional ensemble statistics as predictors was shown to play a key role in its outperformance. Further performance comparisons by Ghazvinian et al. (2020), which relied on only the ensemble mean predictor and were conducted over the American River Basin in California, pointed to the clear outperformance of MMGD, particularly in predicting PoP. The authors confirmed that the use of a two-part scheme helped improve the representation of the predictive distribution.

We select MMGD as the second reference model to further address these discrepancies in the findings of previous studies. This enables us to determine whether our unified ANN-CSGD model improves upon the operational paradigm (MMGD), especially in situations where CSGD underperforms the latter, and helps us identify possible factors that contribute to the differential performance of the three schemes.

2.3. Data and experimental setup

The experiments focus on 24-h mean areal precipitation (MAP) totals over sub-basins of three major river basins in the service area of the NWS California-Nevada River Forecast Center (CNRFC; https://www.cnrfc.noaa.gov).

We use ensemble mean precipitation forecasts from January 1985 through December 2016 (32 years) for lead times 1 to 7 days. These data were obtained from the Global Ensemble Forecast System (GEFS; version 10) reforecast dataset (Hamill et al. 2013) and were processed by the CNRFC at 1-degree spatial resolution and 6-h accumulation intervals issued daily at 00 universal time (UTC). As ground truth, we use the basin MAP data generated by the CNRFC. The MAP data were created using the so-called Mountain Mapper tool, which relies on the Parameter-elevation Regressions on Independent Slopes Model (PRISM; Daly et al. 2008) to group gauges and interpolate gauge reports onto the domain of each watershed. The CNRFC MAP series are at 6-h increments and are available for the period between October 1948 and September 2017. The

MAP data were temporally aggregated to 24-h accumulation and paired with coincidental reforecasts.

Postprocessing experiments are performed over sub-basins in the American River Basin (NFDC1, FOLC1), the Russian River Basin (WSDC1, GUEC1), and the Eel River Basin (DOSC1, FTSC1) (Fig. 2), and separately for upper/lower elevation zones when applicable. Sub-basin names and corresponding NWS IDs are presented in Table 1. The CNRFC runs HEFS routinely to produce postprocessed PQPFs and ensemble streamflow forecasts for many of the sub-basins.

For each river basin, we selected one headwater and one downstream sub-basin for the hindcast experiment to examine the potential elevation dependence in forecast skills. The selected basins have been recognized for their importance in water resources management and flood control, as noted in past hydrometeorological forecast postprocessing/verification studies (see, e.g., Wu et al. 2011; Brown et al. 2012; Seo et al. 2015; He et al. 2016; Scheuerer et al. 2017; Ghazvinian et al. 2020).

The climate of the region is characterized by very dry summers, with most of its annual precipitation falling during the cool season (October – April), and the highest monthly averaged precipitation typically recorded in January. The American River originates from the Tahoe and El Dorado national forests of the Sierra Nevada and is one of the major water supply sources for California. Streamflow in the American River is mainly (2/3) supplied from wintertime rainfall and snowmelt runoff, with a small portion (1/3) from spring to early summer snowmelt runoff (Dettinger et al. 2014). On the other hand, the Russian, and Eel River Basins are coastal basins where snowmelt runoff is much less important (Scheuerer et al. 2017). To be consistent with the

CNRFC operations, we use the nearest neighbor interpolation (Brown et al. 2014a; Seo et al. 2015; Ghazvinian et al. 2020) to pair forecasts-observations.

For generating PQPFs and evaluating the performances of ANN-CSGD relative to the two benchmark models, we adopt an 8-fold cross validation approach. In this approach, for a given basin, we divide the data to 8 consecutive 4-year length folds. Predictions for each fold are produced using each postprocessing mechanism trained with the data of remaining 7 folds (28 years). Postprocessed out of sample forecasts from all models are verified against observations in individual months of the year in verification years and separately for each sub-basin and forecast lead time. This leads to 32 years of verified forecasts for each sub-basin and lead time. While the ANN-CSGD uses the entire available training data (i.e., covering all lead times and seasons) for training and hyperparameter tuning, the benchmark models are trained using subsamples representing each forecast lead time and a month/season of the year. To gain insights on how increasing the length of training record and using different seasonal windows for training can affect the predictions of benchmark models, we train each model with different training window sizes and regulations. A summary of training schemes for ANN-CSGD and benchmark models is provided as follows:

- Unified approach (*ANN-CSGD*) uses forecast-observation pairs of all months and lead times of training years for training and hyperparameter tuning, resulting in a training sample size of up to $7\ \text{lead times} \times 28\ \text{years} \times 365\ \text{days} = 71540$, 20% of which is dedicated for hyperparameter tuning and not used in training.
- MMGD and CSGD with 61 days and 91 days training windows (*MMGD-61, CSGD-61*) and (*MMGD-91, CSGD-91*) use 61 and 91 training days around the 15$^{\text{th}}$ of each month across training years for generating PQPF for out of sample data of that month, yielding

training sample size up to $28\,\text{years} \times 61\,\text{days} = 1708$ and $28\,\text{years} \times 91\,\text{days} = 2548$ for each lead time and month, respectively. 61 days and 91 days training windows have been used in several past studies (e.g., Hamill et al. 2015; Scheuerer and Hamill 2015; Scheuerer et al. 2017; 2018; Wu et al. 2018).

- MMGD seasonal training scheme (*MMGD-seasonal*), where forecasts in out of sample data from the cool (October-April) and dry (May-September) seasons are postprocessed by a model trained using the data in each season. Thus, a single model is trained for each season and each lead time.
- CSGD seasonal training scheme (*CSGD-seasonal*) (Scheuerer et al. 2020) where the climatological CSGD parameters (μ_{cl}, σ_{cl} and δ_{cl}) as well as the climatological mean forecast $\overline{f_{cl}}$ are derived using a 61-day window around the 15th of each month, but the same regression coefficients are used across cool and dry seasons to increase the training sample size.

The latter two training schemes yield a sample size of up to 5942 and 4284 for the cool and dry seasons, respectively.

3. Results

In this section we present verification results using different metrics (see appendix B for mathematical definitions and details). We first use the continuous ranked probability skill score (CRPSS) to assess the overall predictive performance of PQPFs from ANN-CSGD relative to those from the benchmark models with different training scenarios. Subsequently, we analyze ANN-CSGD's performance relative to the benchmark models with a 61-day training window, using Brier skill score (BSS), reliability diagrams, and mean squared error skill score (MSESS).

3.1. Overall predictive performance of PQPFs

Fig. 3 compares CRPSS of PQPFs from ANN-CSGD and those from the benchmark models with different training scenarios and for the three river basins. The results are computed using cross validated-forecasts from all months and are aggregated over sub-basins of each river basin with *MMGD-61* as the reference forecast. To assess whether differences in predictive performances shown are statistically significant, we perform one-sided Diebold-Mariano test (Diebold and Mariano 1995) for all possible pairs of model comparisons (see appendix B for details). These results are provided in tables S1–S3 in the supplemental material to this article.

Overall, ANN-CSGD generates the most skillful PQPFs across lead times. In the American River (Fig. 3a), ANN-CSGD outperforms its baseline CSGD with different training scenarios by a wide margin. The improvement upon each CSGD scheme is statistically significant at all lead times. Nevertheless, performance differences between ANN-CSGD and each of MMGDs are not statistically significant. In the Russian River Basin (Fig. 3b), ANN-CSGD significantly outperforms each of benchmark models in a large number of cases. In the Eel River Basin (Fig. 3c), ANN-CSGD outperforms both MMGDs and CSGDs, though its difference with MMGD-61 is not statistically significant. It is apparent that the relative performance of MMGD and CSGD varies by river basin and at different lead times. Except for the American River Basin, where most differences are not statistically significant, the seasonal version of MMGD trails behind those calibrated with 61- and 91-day moving windows.

For all three river basins, the performance differences of CSGD-61 and CSGD-91 are not statistically significant across the lead times. Interestingly, unlike MMGD-seasonal, CSGD-seasonal tends to considerably improve its performance at longer lead times and for all river basins. The training strategy used in CSGD-seasonal was recently introduced by Scheuerer et al. (2020) in their subseasonal forecast scheme (+ 2 week ahead). This scheme presumes that NWP forecast error characteristics change on a season scale when the forecast has very limited skill. Our result confirms the hypothesis that performance is enhanced through the use of wider seasonal windows. Expanding the seasonal window potentially reduces the risk of overfitting of nonlinear CSGD regression model coefficients at longer lead times when the signal to noise ratio is rather poor.

The results corroborate our postulation that different temporal data pooling methods for training statistical postprocessing models exert influences on the accuracy of postprocessed PQPFs. The use of MMGD as an alternative scheme serves to further illustrate the significance of ANN-CSGD model. EMOS methods such as CSGD are deemed inflexible in that the response variable in these models is assumed to follow a single unimodal parametric distribution (see, e.g., Taillardat et al. 2016; Wu et al. 2019; Baran and Lerch 2018), which potentially limits their performance. As such, why does ANN-CSGD retain its superior performance relative to CSGD across lead times and study basins while both use the same predictive distribution? This is most likely due to the fact that ANN-CSGD uses the entire training dataset and encodes nonlinear lead time- and seasonal-error dependencies in forecasts in an adaptable manner. Thus, it can preserve the skill of raw forecast, particularly at longer lead times, where postprocessing via CSGD-seasonal offers marginal benefit, or even degrades forecast skill. Another advantage of the proposed scheme is that it reduces the risk of overfitting due to the early stopping algorithm implemented as a part of its training.

3.2. Brier skill score and reliability

Fig. 4 shows the results of BSS for three thresholds > 0.25, > 30 and 60 mm/24h and for the three river basins. While both ANN-CSGD and CSGD underperform MMGD in predicting events > 0.25 mm (i.e., PoP), ANN-CSGD, interestingly, conspicuously outperforms CSGD

(Figs. 4a-c). As pointed out by Ghazvinian et al. (2020), CSGD performs poorly in predicting the PoP due to its reliance on the climatological shift parameter (see also sec. 2.2.1 for further details). When the forecast is very skillful, the predictive CSGD departs from climatology, so does the optimal shift parameter. At longer forecast lead times, the forecast skill declines and the predictive CSGD tends to approach the unconditional climatological one. This feature is reflected in the improvement in CSGD's performance across the lead times. ANN-CSGD, on the other hand, directly estimates the shift parameter of the predictive CSGD as an arbitrary function of predictors, thus eliminating the need for a climatological shift parameter. This results in large and statistically significant improvements relative to the CSGD in predicting the PoP. As for the outperformance of MMGD relative to the ANN-CSGD, we hypothesize that the flexible two-part structure of MMGD is likely a major contributor. A detailed discussion on this matter can be found in Ghazvinian et al. (2020).

At the middle threshold of 30 mm/day, ANN-CSGD outperforms both schemes in the American River Basin (Fig. 4d). In the Russian River Basin and the Eel River Basin (Figs. 4e and f), the relative performance of ANN-CSGD and CSGD is mixed but both manage to outperform MMGD, except at Day 7 in the Russian River basin where CSGD slightly underperforms, though it is not statistically significant (not shown here). At the highest threshold, namely 60 mm/day (Figs 4g-i), ANN-CSGD outperforms all other schemes. CSGD mostly outperforms MMGD in the Russian River Basin (Fig. 4h) but underperforms the latter in American River and Eel River basins (Fig. 4g, i).

To compare the calibration of PQPFs produced through each scheme, we plot reliability diagrams for the same events and evaluate the contribution of reliability and resolution to the Brier score (Figs. 5,6, and 7). To attain a large enough sample size to better study larger thresholds, we lump cross-validate forecasts at all lead times, and divide forecast probabilities [0,1] into 15 evenly distributed probability categories to discern the differential performance of schemes under higher probability categories. The major findings for each river basin are summarized as follows:

- American River Basin: In predicting positive precipitation events (> 0.25 mm/day) (Figs. 5a-c), ANN-CSGD's outperformance relative to CSGD is attributed to improvements in both reliability (lower REL) and resolution (higher RES). ANN-CSGD mitigates to a great extent the underforecast issue of CSGD. ANN-CSGD generates PQPFs that are more reliable than MMGD but are characterized with lower resolution, yielding an overall inferior predictive performance. At higher thresholds (Figs. 5d-i), ANN-CSGD clearly outperforms both CSGD and MMGD in terms of both reliability and resolution. As shown in the histograms embedded in each subplot, ANN-CSGD generates PQPFs that are able to issue high probabilities in predicting mid-to-heavy precipitation with higher frequencies, and this points to improved sharpness (Figs. 5f and i).
- Russian River Basin: Similar to the American River Basin, at the lowest threshold (Figs. 6a-c), ANN-CSGD produces forecasts with higher reliability (lower REL) than MMGD but with lower resolution and overall lower predictive skill (higher BS). In > 30 mm/day ANN-CSGD performs better than CSGD in terms of both reliability and resolution (Figs. 6e, f). At the highest threshold (Figs. 6h, i), the lack of reliability in ANN-CSGD PQPFs relative to those from CSGD is compensated by the higher resolution, and this leads to a superior predictive performance of the former as evidenced by the lower BS. MMGD at both thresholds (Figs. 6d, g) produces less

reliable PQPFs with lowest sharpness. At the 30 mm/day threshold (Fig. 6d), MMGD PQPFs' resolution is somewhat higher but is compensated by lower reliability.

- Eel River Basin: At the lowest threshold (> 0.25mm/day) (Figs. 7a-c), the relative performance of schemes is quite similar to that for the other two river basins, with ANN-CSGD outperforming MMGD in terms of reliability but not resolution. At higher thresholds (Figs. 7d-i), PQPFs from ANN-CSGD are more reliable and sharper and, overall, more skillful (lowest BS). Though at the highest threshold (i.e., > 60 mm/day), the former exhibit slightly lower resolution than those from MMGD, but this is compensated by superior reliability.

3.3. Evaluation of deterministic forecasts

Finally, we compute mean squared error skill score (MSESS) to evaluate the performance of the distribution mean of PQPF produced using each scheme relative to the GEFS ensemble mean forecast (Fig. 8). These results are accompanied by the results of the Diebold-Mariano test based on the squared error of mean PQPFs (see Tables S4–S6 in the supplemental material). The relative performance varies among the river basins. For the American River Basin (Fig. 8a), all postprocessed PQPFs outperform the GEFS ensemble mean in terms of MSESS. ANN-CSGD PQPFs perform favorably against MMGD PQPFs for all three river basins (the performance differences are not statistically significant). For both the Russian and Eel River Basins (Figs. 8b and c), MSESS values are generally lower relative to those for the American River Basin. This, as we posit, is attributable to location-dependent biases in the GEFS ensemble mean forecast. For example, GEFS is more skillful in the Russian and Eel River Basins according to the MSESS results relative to climatological forecasts (the results are shown in Fig. S1 of supplemental materials). For the Russian River Basin (Fig. 8b), underperformance of postprocessed PQPF relative to the GEFS ensemble mean is seen; however, the performance differences are not statistically significant. Unlike the benchmarks, mean PQPF from ANN-CSGD for Russian River Basin significantly outperforms GEFS ensemble mean forecast in all lead times. For both the Russian and Eel River Basins (Figs. 8b, c), ANN-CSGD tends to outperform the other two schemes, though the performance differentials are not statistically significant when comparing with MMGD.

4. Discussion and conclusions

We propose a unified, univariate, hybrid neural network-parametric PQPF postprocessing scheme capable of producing postprocessed forecasts for lead times at least up to 7 days (medium-range). This scheme retains the use of parametric predictive distribution, but employs ANN to estimate distribution parameters from forecast-observation pairs. The predictors explored in this study include ensemble mean forecast, forecast lead time, and month of the year, whereas the predictands are three parameters of the predictive censored, shifted gamma distribution (CSGD). The ANN-CSGD model parameters were obtained by minimizing a loss function that is the closed-form expression of CRPS for CSGD (Scheuerer and Hamill 2015), with the Adam stochastic gradient descent algorithm (Kingma and Ba 2014) as the optimization approach. To test the performance of our model, we conducted cross-validation experiments to generate medium-range (lead times 1–7 days) daily accumulated PQPFs over selected river basins in the service area of the CNRFC. We used two benchmarking processing schemes in this study, namely the CSGD EMOS (Scheuerer and Hamill 2015) with a single-predictor formulation and the NWS operational postprocessor mixed-type Meta-Gaussian distribution (MMGD). These benchmark models were calibrated based on different seasonal data pooling scenarios to investigate the possible impacts of training window size and strategies on the performance of postprocessed PQPFs.

Verification results showed that ANN-CSGD, in general, outperform the baseline CSGD and MMGD in terms of overall calibration, and significantly so in some cases. Interestingly, ANN-CSGD mainly impacts (improves) BSS of PQPF from CSGD at the lowest threshold, which has disproportionate impacts on CRPSS. ANN-CSGD manages to address the CSGD's poor performance in predicting PoP as noted in Ghazvinian et al. (2020). While the ANN-CSGD performance comparison results are mixed in predicting 30 mm/day thresholds, it outperforms both benchmark models in predicting large-extreme events (> 60mm/day). On average, the proposed method generates high probability forecasts for heavy precipitation more frequently than benchmarks as assessed by sharpness histograms (higher sharpness). This is particularly useful to CNRFC's operational precipitation and flood forecasting practice and, thus, could benefit real-time reservoir operations (e.g., determining reservoir release schedules) in California. In its current practice, CNRFC relies on HEFS to produce ensemble PQPFs from NWP precipitation forecasts and then generates ensemble streamflow forecasts, which are used to guide real-time flood management and control practices. The MMGD model, embedded in HEFS, has shown to systematically underestimate heavy precipitation amounts, leading to negative biases in subsequent flood forecasts (Demargne et al. 2014; Brown et al. 2014b). The superior performance of the proposed ANN-CSGD on heavy precipitation estimation makes it a viable tool to address limitations in the forecast skills for extreme precipitation and floods. These improvements in forecasts will, in turn, serve to aid real-time reservoir operations and flood risk management.

In contrast to the CSGD version of Scheuerer and Hamill (2015), the proposed method directly estimates predictive CSGD's shift parameter given each set of predictors. In doing so, it circumvents the need of invoking climatology, and thereby alleviates the bias issue in estimating the PoP in the existing CSGD scheme. Furthermore, the use of ANN allows for representations of complex interactions between three predictive CSGD parameters. Together, these new features help the scheme produce sharper (narrower) predictive distributions than the benchmark CSGD. Moreover, ANN-CSGD is able to use much larger training data with extra high forecast-observation values, and efficiently translate this to predictive skill at the highest threshold.

The new scheme also has a distinct practical advantage in that it eliminates the need for more computationally expensive and operationally labor-intensive approach used in most contemporary statistical postprocessing schemes. Whereas the benchmark models need to be re-trained for every forecasting lead time and month/season, ANN-CSGD does not, and it can simultaneously utilize forecast-observation pairs across all lead times, months, and seasons. Our results support our hypothesis that the fixed size seasonal window training schemes for current postprocessing methods may not be sufficient for generating consistently skillful PQPFs across all lead times. In other words, the performance of existing schemes may be improved by identifying an *optimal* seasonal training window specific for each lead time, depending on the study area and the statistical model at hand. For example, it was shown that a seasonal CSGD tended to improve the performance benchmark 61-day and 91-day CSGDs at longer lead times but not in shorter lead times. ANN-CSGD, on the other hand, automatically adapts to the changes in raw forecasts-observations errors along with all lead times and seasons, and hence, is capable of producing PQPFs with consistently higher skills.

A major limitation of nonhomogeneous regression or GAMLSS techniques is that their performance is dependent on the robustness of user-prescribed regression relationships. Moreover, they are typically limited in digesting ordinal temporal covariates such as those used in the ANN-CSGD model. The proposed model, by contrast, can freely learn to characterize arbitrary nonlinear predictor-distribution parameters relationships and among-predictors interactions efficiently.

A well-known challenge in training ANN models is model configuration (hyperparameter tuning) to achieve the best validation score. Generally, it is very difficult to find the best possible ANN configuration in a very large parameter space. As pointed out by Scher (2018), there is a trade-off between robustness, which depends on the depth and thoroughness of grid search, and computational expenses. For example, our initial assessment showed that maintaining the architecture but expanding the number of layers does not significantly improve the model performance. Other regularization techniques such as L1 could be used in combination with early stopping to further reduce generalization errors. However, these techniques could require deeper search for hyperparameters and, therefore, increase computational complexity. We also experimented with training embedding layers with different sizes $\{2,3,4,5,6,7\}$ to project discrete lead times onto a larger vector of inputs but only found very marginal improvements in the validation score. Therefore, we decided not to include embedding layers in our final model.

In future work, we aim to extend the current approach to create a spatially adaptable scheme for postprocessing medium-range ensemble precipitation forecasts on a gridded basis. We expect to achieve this by incorporating geographical information into the network as shown by Scheuerer et al. (2020) in their subseasonal forecasting approach. For example, entire ensemble members or their statistics at a grid point, in addition to those from a specific radius of surrounding grid points, can be direct inputs to the model as the predictors. Such a model potentially eliminates the need for generating a local superensemble to address the issue of displacement errors in gridded precipitation forecasts.

Additionally, the current study focuses on 24-hour accumulated precipitation. In operations, CNRFC produces 6-hourly PQPFs and updates their forecasts every 6 hours during major storm events. To align with CNRFC operations, we also plan to explore the performance of the proposed ANN-CSGD in generating 6-hourly PQPFs in our future work. Finally, stacked convolution or Long Short-Term Memory (LSTM) layers applied on top of embedding vectors, appear to be very effective in object detection (Krizhevsky et al. 2012), in computer vision, and in Natural Language Processing (Collobert et al. 2011), including Machine Translation and Question Answering (Devlin et al. 2018). We envision investigating similar techniques to possibly improve the skill of postprocessed forecast at longer lead times.

Acknowledgements

The authors thank the editor and reviewers for their valuable comments that helped improve the article. The first author was financially supported by the faculty startup fund for Dr. Yu Zhang provided by UT Arlington, NOAA Grant NA18OAR4590370-01, Texas Water Development Board Contract No. 1800012276, and NSF grant 1909367. These supports are duly acknowledged here. The authors would also like to thank Michael Scheuerer at Norwegian Computing Center (NR) whose comments and suggestions led to the development of the scheme, and CNRFC staff for providing the forecast and analysis archive.

Appendix A

Implementation details

We implemented our ANN codes in python (Python Software Foundation 2018) using Google's deep learning platform, Tensorflow (Abadi et al. 2016) and Keras API (Chollet et al. 2015). For fitting CSGD climatological and predictive distributions, R (R Core Team 2018) scripts provided by Dr. Michael Scheuerer were used. To calibrate NWS postprocessor, mixed-type meta-Gaussian distribution (MMGD), a research version, very similar to the operational one was implemented in R.

Appendix B

Verification metrics used in this study

A. Mean squared error skill score (MSESS)

The mean squared error skill score (MSESS; Jolliffe and Stephenson 2003) measures the reduction in mean squared error (MSE) of deterministic forecast (mean PQPF/ensemble mean) and verifying observations relative to the reference forecast.

$$MSESS = 1 - \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - y_i)^2 \Big/ \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i^{ref} - y_i)^2 \qquad \text{(A1)}$$

Positive values of MSESS indicates improvement in skill of deterministic forecast relative to the reference forecast.

B. Brier skill score (BSS)

The Brier score (BS; Brier 1950) is equivalent to mean squared error of probabilistic forecast exceeding a given threshold over n pairs of forecast and observations

$$BS(\tau) = \frac{1}{n}\sum_{i=1}^{n}[F_i(\tau) - I\{y_i \geq \tau\}]^2 \qquad \text{(A2)}$$

where $F_i(\tau)$ is the probability of probabilistic forecast exceeding the threshold value τ , and $\mathbf{I}(.)$ is the indicator (step) function that takes the value 1 if the ith verifying observation exceeds the threshold value and 0 otherwise. BS is negatively oriented and ranges from zero to one. To assess the improvement in BS relative a reference forecast, we compute Brier skill score

$$BSS = 1 - BS/BS_{ref} \tag{A3}$$

Positive values of BSS indicate improvement of BS over that of reference forecast. Brier score can be decomposed to three terms: *reliability or Type-I conditional bias*, *resolution,* and *uncertainty* (Murphy 1973; Wilks 2011)

$$BS(\tau) = Reliability(\tau) - Resolution(\tau) + Uncertainity(\tau) \tag{A4}$$

$$= \frac{1}{n}\sum_{i=1}^{K} N_i[F_i(\tau) - \bar{o}_i(\tau)]^2 - \frac{1}{n}\sum_{i=1}^{K} N_i[\bar{o}_i(\tau) - \bar{o}(\tau)]^2$$

$$+ \bar{o}(\tau)[1 - \bar{o}(\tau)]$$

where K indicated the number of categories, forecast are aggregated to, N is the number of cases in each category, $\bar{o}_i(\tau)$ is the average climatological probability (ACP) exceeding the threshold τ in that category and $\bar{o}(\tau)$ is the overall ACP. It should be noted that uncertainty term as seen is independent of the forecast source. Probabilistic forecasts with lower/higher reliability/resolution values are desirable.

C. Continuous ranked probability score (CRPS)

The continuous ranked probability score (CRPS; Matheson and Winkler 1976) measures the integral of squared differences between the cumulative distribution function (CDF) of probabilistic forecast and verifying observation. It is a popular metric to assess the overall predictive performance of probabilistic forecasts (sharpness and reliability; see Gneiting et al. 2007 for further details). CRPS averaged over the sample of forecast-observations with size of n is given by

$$CRPS = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}[F_i(x) - I\{y_i \le x\}]^2 dx \tag{A5}$$

where F_i denotes the CDF of PQPF at the ith forecast instance and $y_i$is the verifying observation. **I (.)** is the indicator (step) function which takes the value of 1 if $x \ge y_i$ and 0 elsewhere. Continuous ranked probability skill score (CRPSS) is routinely used to assess the performance of probabilistic forecast relative to a reference forecast

$$CRPSS = 1 - CRPS/CRPS_{ref} \tag{A6}$$

D. Reliability diagrams and sharpness histograms

The reliability and resolution of a probabilistic forecast for exceeding some specific thresholds (τ) can be assessed graphically using reliability diagrams. The reliability diagram consists of a plot of the average values of forecast probabilities exceeding τ, against that of observed relative frequencies over each defined probability category. In a reliable probabilistic forecast, the reliability diagram should be a close 1:1line. Interested readers are referred to Brocker and Smith (2007) and Wilks (2011) for details on how to interpret the deficiencies in probabilistic forecasts using reliability diagrams. To assess the sharpness of PQPF for specific thresholds, we use sharpness histograms to investigate the frequency of forecast probabilities for different probability bins. Note, a sharp forecast is characterized by higher frequencies for the forecast probabilities close to either 0 or 1.

E. The Diebold-Mariano test

To assess statistical significance of verification score differences between two forecast methods, we use the Diebold-Mariano statistical test of the null hypothesis of equal predictive performance (Diebold and Mariano 1995). Let $\Delta = S_{F1} - S_{F2}$ denote the vector of verification score S differences from two competing forecast methods F_1 and F_2 over verification sample with length n, $\overline{\Delta} = 1/n \sum_{i=1}^{n} \Delta_i$, and $\hat{\sigma}_{\Delta}$ a suitable estimator of asymptotic standard deviation of Δ. Under standard regularity conditions, the test statistic $t_n = \sqrt{n}\frac{\overline{\Delta}}{\hat{\sigma}_{\Delta}}$ asymptotically follows a standard Gaussian distribution under the null hypothesis of no difference in predictive performances of two competing forecast methods. Following the past studies (Baran and Lerch 2016, 2018; Rasp and Lerch 2018) $\hat{\sigma}_{\Delta}$ can be estimated by square root of sample autocovariance up to lag $k-1$ for the k step-ahead forecasts to account for temporal dependencies in forecast errors. We use one-sided Diebold-Mariano tests. The alternative hypothesis is that forecast method F_2 underperforms forecast method F_1 and the statistical significancy of the test's statistic can be assessed by obtaining corresponding *p-value.* we perform the tests based on both CRPS and squared error of mean PQPF (on a limited basis) and for each lead time and separately for each river basin. To address spatial dependence of forecast errors, scores are averaged across sub-basins in each river basins (M. Scheuerer 2021, personal communication). Further, we adjust the test results by accounting for test multiplicity (i.e., simultaneously analyzing test results of multiple lead times) using false discovery rate (FDR) method (Benjamini and Hochberg 1995) by controlling the FDR at the level α_{FDR}= 0.05. Note that, this procedure was discussed by Wilks (2016) in spatial context where test results are interpreted simultaneously across multiple grid points but also was suggested to be applied whenever the results of simultaneous several hypothesis tests are reported or interpreted.

References

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

Ba, J. L., J. R. Kiros, and G. E. Hinton, 2016: Layer normalization. arXiv preprint arXiv:1607.06450, https://arxiv.org/abs/1607.06450.

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.

Baran, S., and S. Lerch, 2016: Mixture EMOS model for calibrating ensemble forecasts of wind speed. Environmetrics, 27, 116–130, https://doi.org/10.1002/env.2380.

Baran, S., 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.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B, 57, 289–300, https://doi.org/10.1111/J.2517-6161.1995.TB02031.X.

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.

Brenowitz, N. D., and C. S. Bretherton, 2018: Prognostic validation of a neural network unified physics parameterization. Geophys. Res. Lett., 45, 6289–6298, https://doi.org/10.1029/2018GL078510.

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.

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., D. Seo, and J. Du, 2012: Verification of Precipitation Forecasts from NCEP's Short-Range Ensemble Forecast (SREF) System with Reference to Ensemble Streamflow Prediction Using Lumped Hydrologic Models. J. Hydrometeor., 13, 808–836, https://doi.org/10.1175/JHM-D-11-036.1.

Brown, J. D., L. Wu, M. He, S. Regonda, H. Lee, and D.J. Seo, 2014a: 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.

Brown, J. D., M. He, S. Regonda, L. Wu, H. Lee, and D.J. Seo, 2014b: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. Hydrol, 519, 2869–2889, https://doi.org/10.1016/j.jhydrol.2014.05.030.

Cannon AJ., 2012: Neural networks for probabilistic environmental prediction: conditional density estimation network creation and evaluation (CaDENCE) in R. Comput. Geosci.41:126–35, https://doi.org/10.1016/j.cageo.2011.08.023.

Chapman, W. E., Subramanian, A. C., Delle Monache, L., Xie, S. P., and F. M. Ralph, 2019: Improving atmospheric river forecasts with machine learning. Geophysical Research Letters, 46, 10,627–10,635. https://doi.org/10.1029/2019GL083662.

Chollet, F., and Coauthors, 2015: Keras: The Python Deep Learning library. Accessed 2019, https://keras.io.

Clevert, D. A., T. Unterthiner, and S. Hochreiter, 2015: Fast and accurate deep network learning by exponential linear units (ELUs). Int. Conf. on Learning Representations, San Juan, Puerto Rico, ICLR, 1–14, https://arxiv.org/abs/1511.07289.

Cloud, K. A., B. J. Reich, C. M. Rozoff, S. Alessandrini, W. E. Lewis, and L. Delle Monache, 2019: A feed forward neural network based on model output statistics for short-term hurricane intensity prediction. Wea. Forecasting,34, 985–997, https://doi.org/10.1175/WAF-D-18-0173.1.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, 2011: Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12:2493-2537. Available online at: 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.

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

Dettinger, M. D., D. R. Cayan, M. K. Meyer, and A. E. Jeton, 2014: Simulated hydrologic responses to climate variations and change in the Merced, Carson, and American river basins, Sierra Nevada, California, 1900–2099, Clim. Change, 62, 283–317.

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

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. J. Bus. Econ. Stat., 13, 253–263, https://doi.org/10.1080/07350015.1995.10524599.

Gagne II, D. J., S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms. Mon. Wea. Rev., 147, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1.

Gentine, P., M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis, 2018: Could machine learning break the convection parameterization deadlock? Geo-phys. Res. Lett., 45, 5742–5751, https://doi.org/10.1029/2018GL078202.

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

Ghazvinian, M., Seo, D. J., and Y. Zhang, 2019: Improving Medium-range Probabilistic Quantitative Precipitation Forecast for Heavy-to-extreme Events through the Conditional Bias-penalized Regression. In AGU Fall Meeting 2019. AGU. https://agu.confex.com/agu/fm19/meetingapp.cgi/Paper/517742.

Gneiting, T., A.E. Raftery, A.H. Westveld, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. Mon. Wea. Rev., 133, 1098–1118, https://doi.org/10.1175/MWR2904.1.

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

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

Hamill, T.M., 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.

Hamill, T. M., 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.

Hamill, T. M., 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.

He, M., and Coauthors, 2016: Verification of ensemble water supply forecasts for Sierra Nevada watersheds. Hydrology, 3, 35, https://doi.org/10.3390/hydrology3040035.

Herr, H. D., and R. Krzysztofowicz, 2005: Generic probability distribution of rainfall in space: The bivariate model. J. Hydrol., 306, 234–263, https://doi.org/10.1016/j.jhydrol.2004.09.011.

Kelly, K. S., and R. Krzysztofowicz, 1997: A bivariate meta-Gaussian density for use in hydrology. Stochastic Hydrol. Hydraul., 11, 17–31, https://doi.org/10.1007/BF02428423.

Kim, S., and Coauthors, 2018: Assessing the Skill of Medium-Range Ensemble Precipitation and Streamflow Forecasts from the Hydrologic Ensemble Forecast Service (HEFS) for the Upper Trinity River Basin in North Texas. J. Hydrometeor., 19, 1467–1483, https://doi.org/10.1175/JHM-D-18-0027.1.

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

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012: Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105). Available online at: http://www.csri.utoronto.ca/~hinton/absps/imagenet.pdf.

Lagerquist, R., A. McGovern, and D. J. Gagne II, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. Wea. Forecasting, 34, 1137–1160, https://doi.org/10.1175/WAF-D-18-0183.1.

Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di, 2017: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. WIREs Water, 4: e1246, https://doi.org/10.1002/wat2.1246.

Liu, Y., E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, 2016: Application of deep convolutional neural networks for detecting extreme weather in climate datasets.arXiv.org, https://arxiv.org/abs/1605.01156.

Liu, Y., P. Di, S. Chen, and J. DaMassa, 2018: Relationships of Rainy Season Precipitation and Temperature to Climate Indices in California: Long-Term Variability and Extreme Events. J. Climate, 31, 1921–1942, https://doi.org/10.1175/JCLI-D-17-0376.1.

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.

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

R Core Team, 2017: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 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.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. Proc. Natl. Acad. Sci. USA,115, 9684–9689, https://doi.org/10.1073/pnas.1810286115.

Rigby, R. A. and D. M. Stasinopoulos, 2005: Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54: 507-554, https://doi.org/10.1111/j.1467-9876.2005.00510.x.

Scher, S., 2018: Toward data-driven weather and climate fore-casting: Approximating a simple general circulation model with deep learning. Geophys. Res. Lett.,45, 12 616–12 622, https://doi.org/10.1029/2018GL080704.

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.

Scheuerer, M., T. M. Hamill, 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.

Scheuerer, M. and T.M. Hamill, 2018: Generating Calibrated Ensembles of Physically Realistic, High-Resolution Precipitation Forecast Fields Based on GEFS Model Output. J. Hydrometeor., 19, 1651–1670, https://doi.org/10.1175/JHM-D-18-0067.1.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California. Mon. Wea. Rev., 148, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Seo, D.-J., and Coauthors, 2015: On improving ensemble forecasting of extreme precipitation using the NWS Meteorological Ensemble Forecast Processor (MEFP). 2015 Fall Meeting, San Francisco, CA, Amer. Geophys. Union, Abstract H51P-08, https://agu.confex.com/agu/fm15/meetingapp.cgi/Paper/81958.

Stauffer, R., N. Umlauf, J.W. Messner, G.J. Mayr, and A. Zeileis, 2017: Ensemble Postprocessing of Daily Precipitation Sums Over Complex Terrain Using Censored High-Resolution Standardized Anomalies. Mon. Wea. Rev., 145, 955–969, https://doi.org/10.1175/MWR-D-16-0260.1.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. Mon. Wea. Rev., 144, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

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.

Vannitsem, S., and Coauthors, 2020: Statistical Postprocessing for Weather Forecasts -- Review, Challenges and Avenues in a Big Data World. arXiv preprint arXiv:2004.06582, https://arxiv.org/abs/2004.06582.

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

Wilks, D. S., 2016: ‘‘The stippling shows statistically significant grid points’’: How research results are routinely overstated and over-interpreted, and what to do about it. Bull. Amer. Meteor. Soc.,97, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1.

Wilks, D. S., 2018: Univariate ensemble postprocessing. In S. Vannitsem, D. S. Wilks, & J. Messner (Eds.), Statistical postprocessing of ensemble forecasts (pp.49–89). https://doi.org/10.1016/B978-0-12-812372-0.00003-0.

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.

Wu, L., Y. Zhang, T. Adams, H. Lee, Y. Liu, and J. Schaake, 2018: Comparative Evaluation of Three Schaake Shuffle Schemes in Postprocessing GEFS Precipitation Ensemble Forecasts. J. Hydrometeor., 19, 575–598, https://doi.org/10.1175/JHM-D-17-0054.1.

Wu, Y., X. Yang, X. Zhang, and Q. Kuang, 2019: Mixture probabilistic model for precipitation ensemble forecasting. Q J R Meteorol Soc. 2019; 145: 3516– 3534. https://doi.org/10.1002/qj.3637.

Xu, J., X. Sun, Z. Zhang, G. Zhao, and J. Lin, 2019: Understanding and improving layernormalization. *arXiv preprint arXiv:1911.07013,* https://arxiv.org/abs/1911.07013.

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

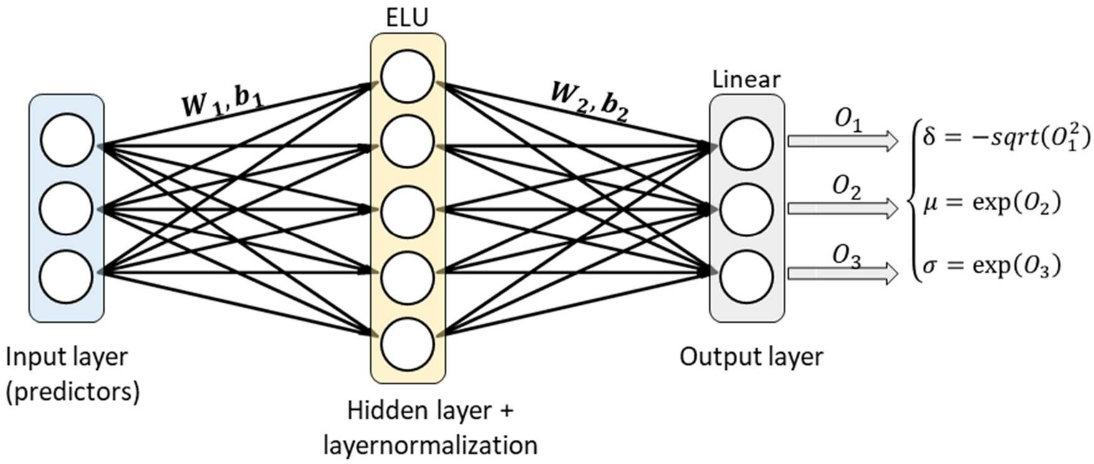

Fig. 1. Schematic of the ANN-CSGD structure. We illustrate hidden layer with 5 nodes for the sake of demonstration. Three parameters of predictive CSGDs are considered as additional functions of ANN outputs to constrain the values of these parameters to reasonable ranges.

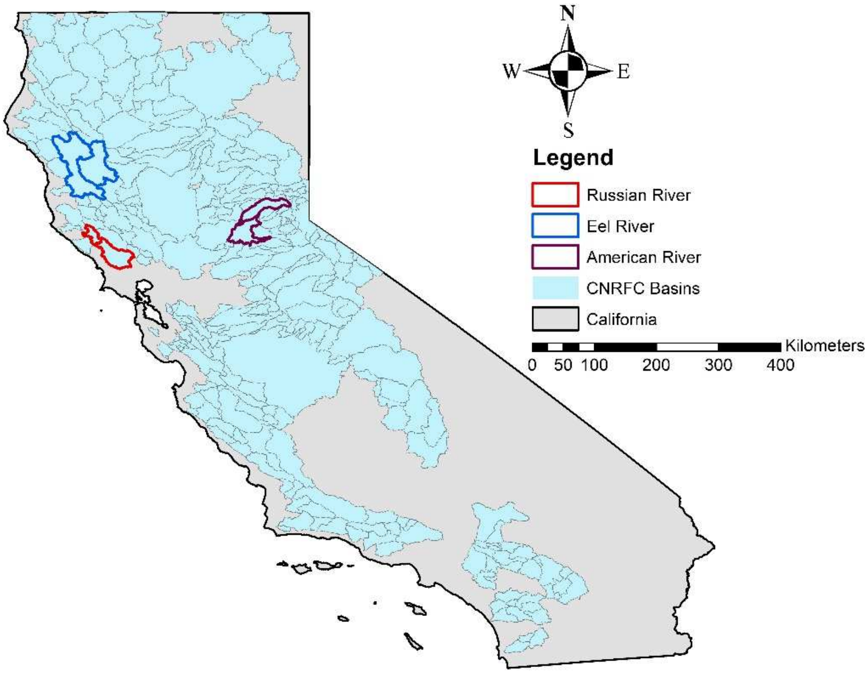

Fig. 2. Location map of the study basins as well as basins in the service area of CNRFC within the State of California.

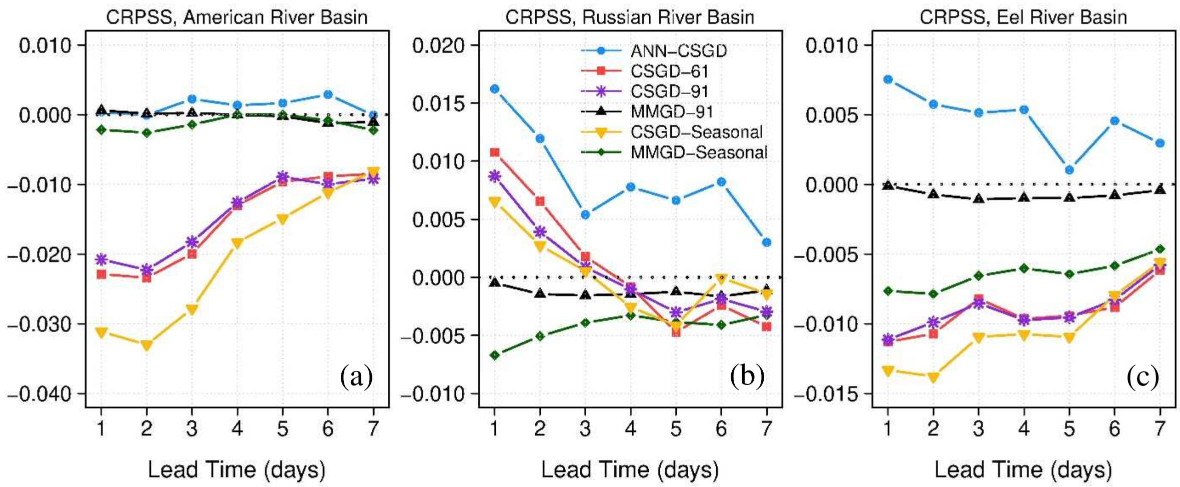

Fig. 3. CRPSS for ANN-CSGD and benchmark postprocessing models with different training scenarios (61-day, 91-day, and seasonal window). Displayed are cross-validated CRPSS computed by pooling CRPS values across study sub-basins in each river basins and all months as a function of lead time. MMGD PQPFs with 61-day training window serve as the reference.

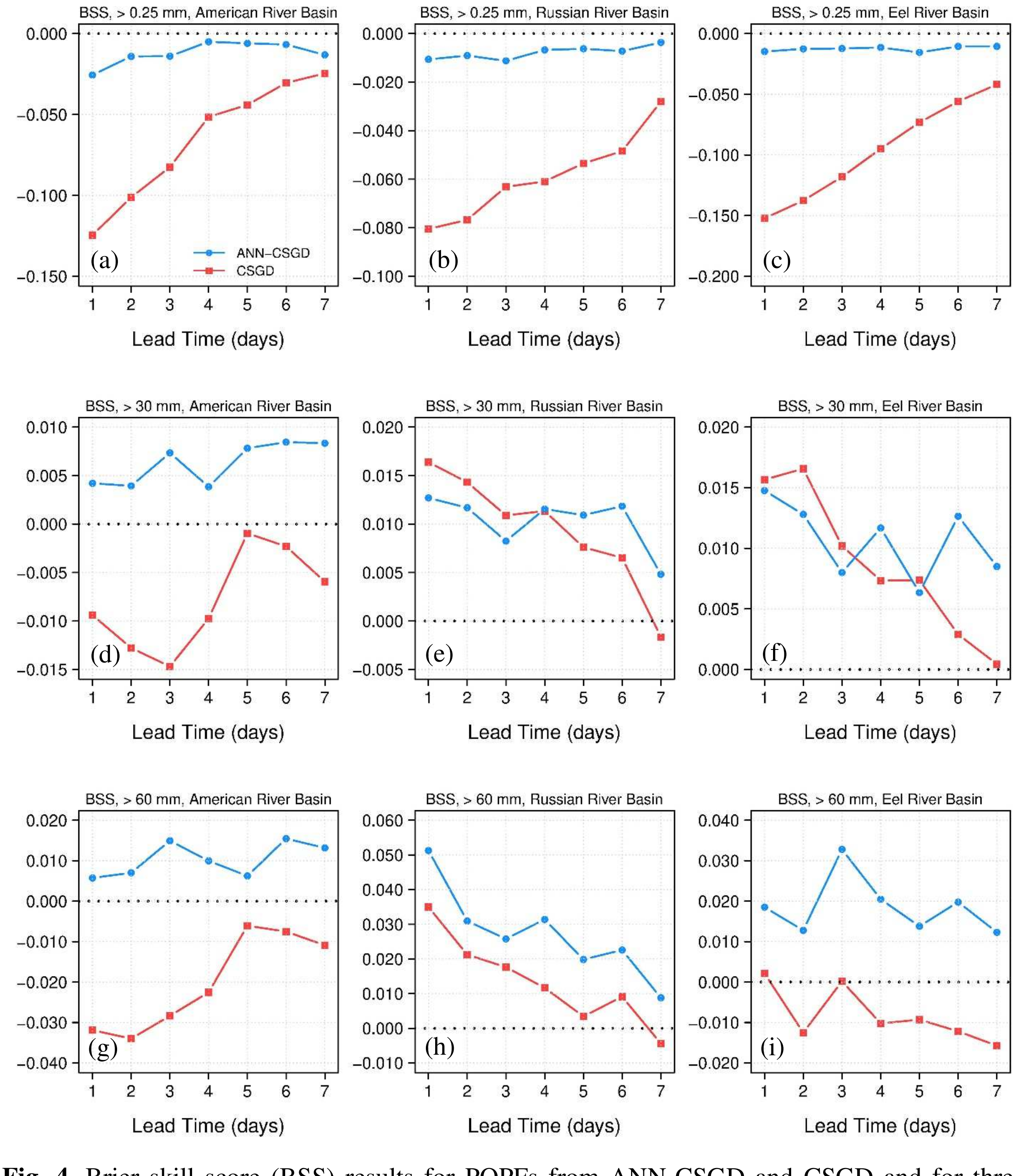

Fig. 4. Brier skill score (BSS) results for PQPFs from ANN-CSGD and CSGD and for three different thresholds: > 0.25, 30 and 60 mm, averaged over study sub-basins in each river basin and shown as a function of lead time, with MMGD-61 as the reference.

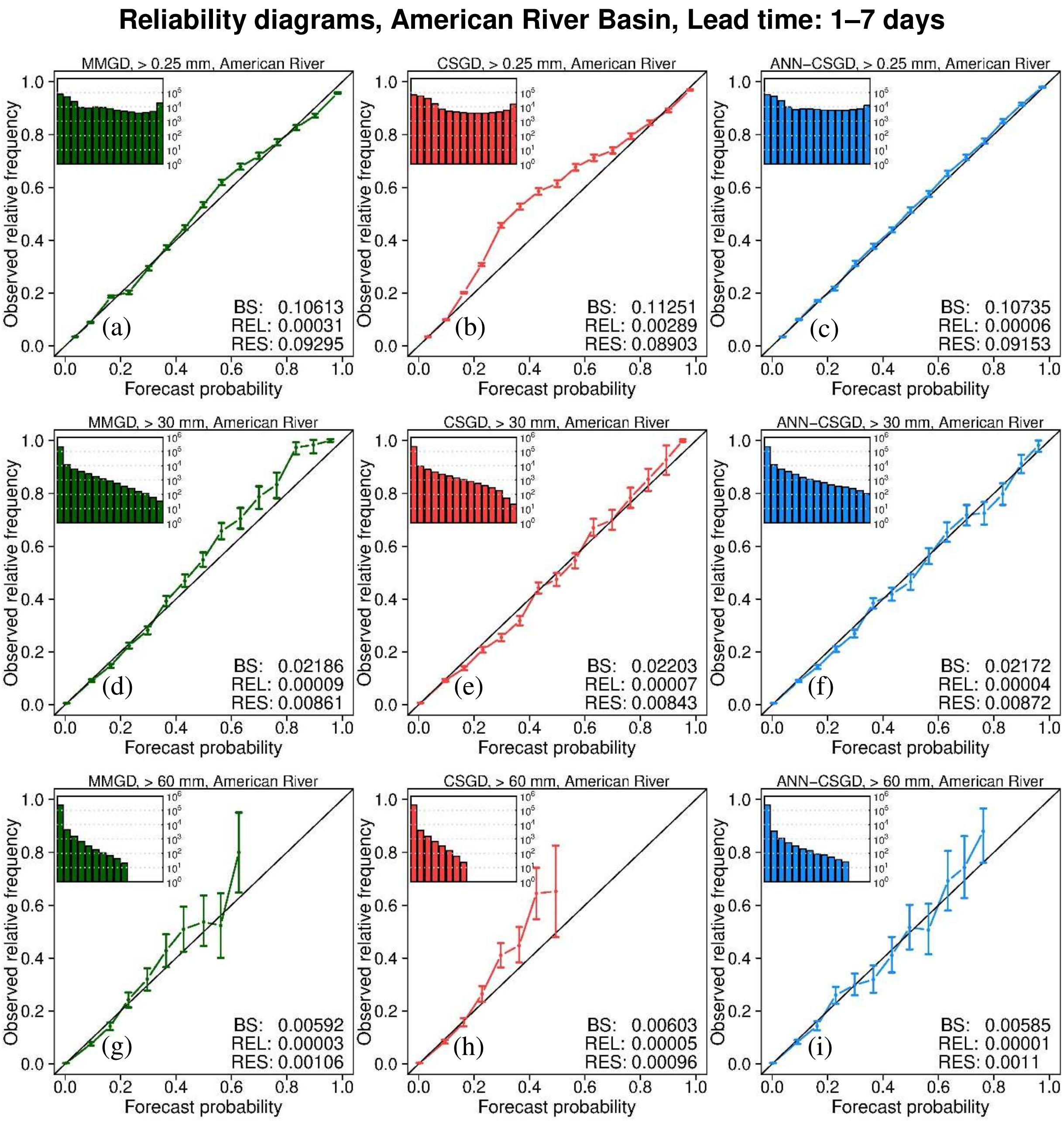

Fig. 5. Reliability diagrams for the three thresholds (> 0.25, 30 and 60 mm) and for sub-basins in the American River Basin were computed based on observations and cross-validated postprocessed forecasts pooled across study sub-basins and all forecast lead times. Brier score (BS), Reliability (REL) and Resolution (RES) values are shown in each panel. The insert histograms show the frequencies for each of 15 forecast probability bins in log10 scale for better visibility and the bars show 90% bootstrap confidence intervals of observed frequencies for estimated forecast probabilities. Benchmark models are trained using 61-day window centered around the 15th of each month.

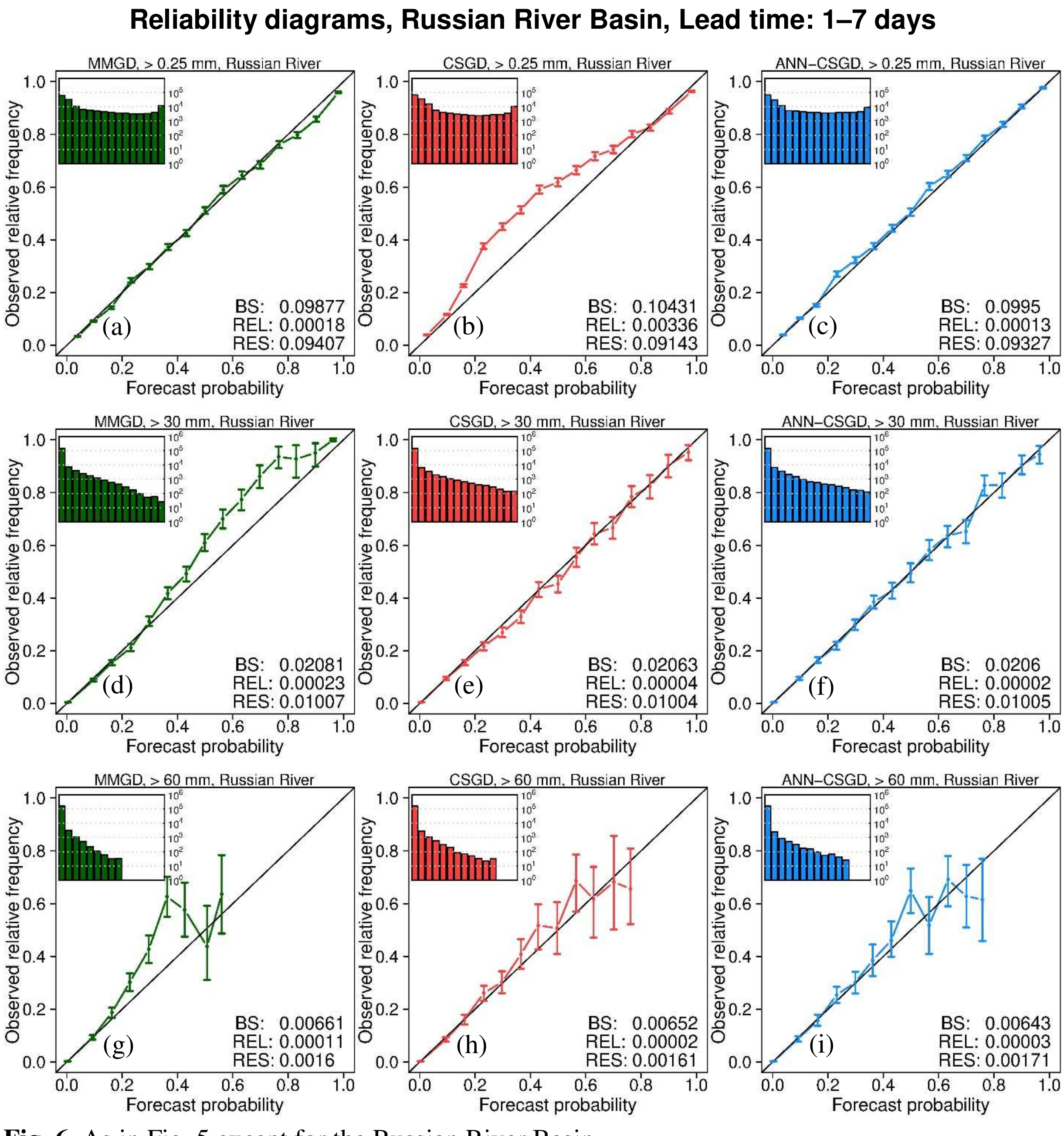

Fig. 6. As in Fig. 5 except for the Russian River Basin.

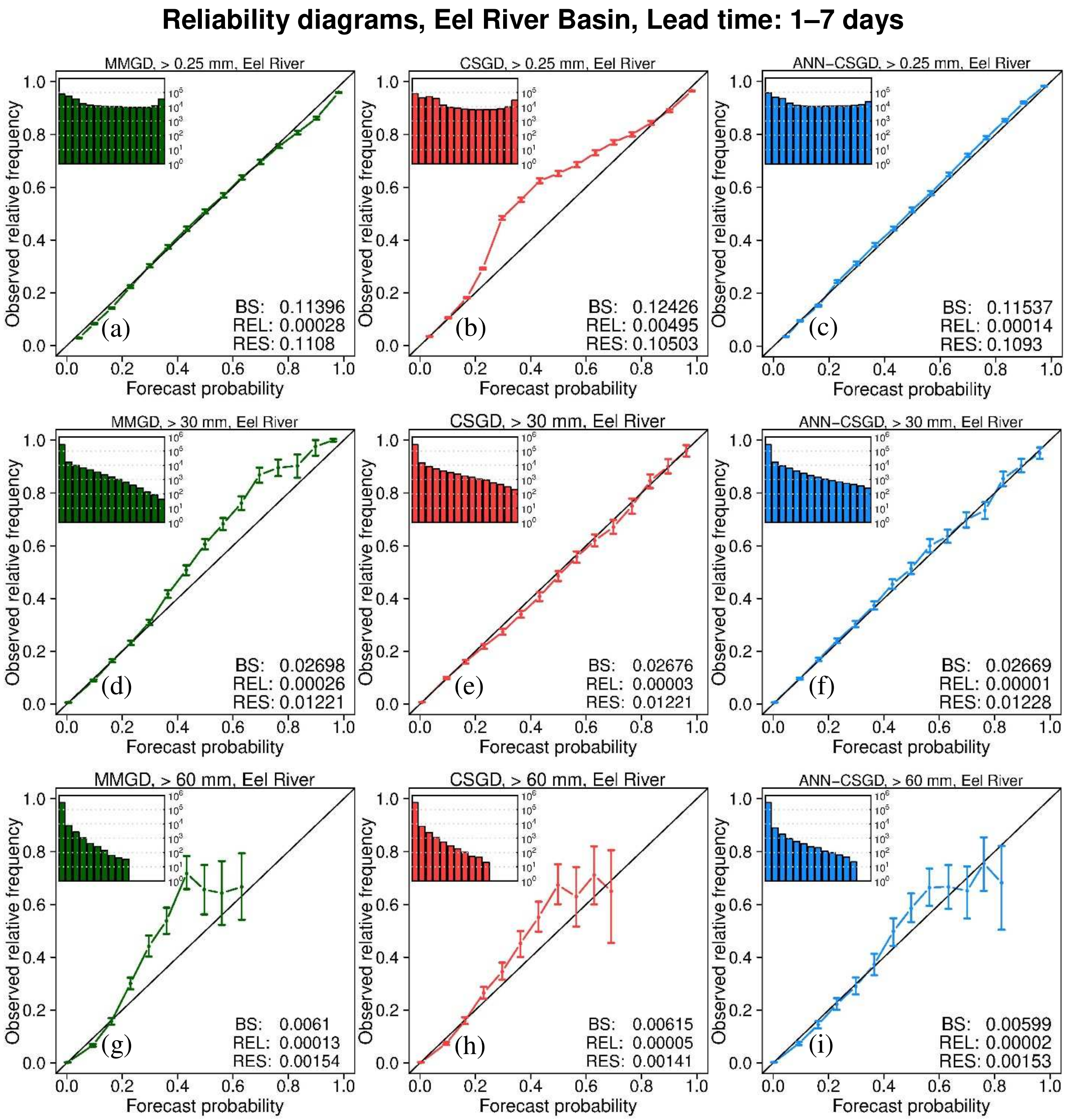

Fig. 7. As in Fig. 5 except for the Eel River Basin.

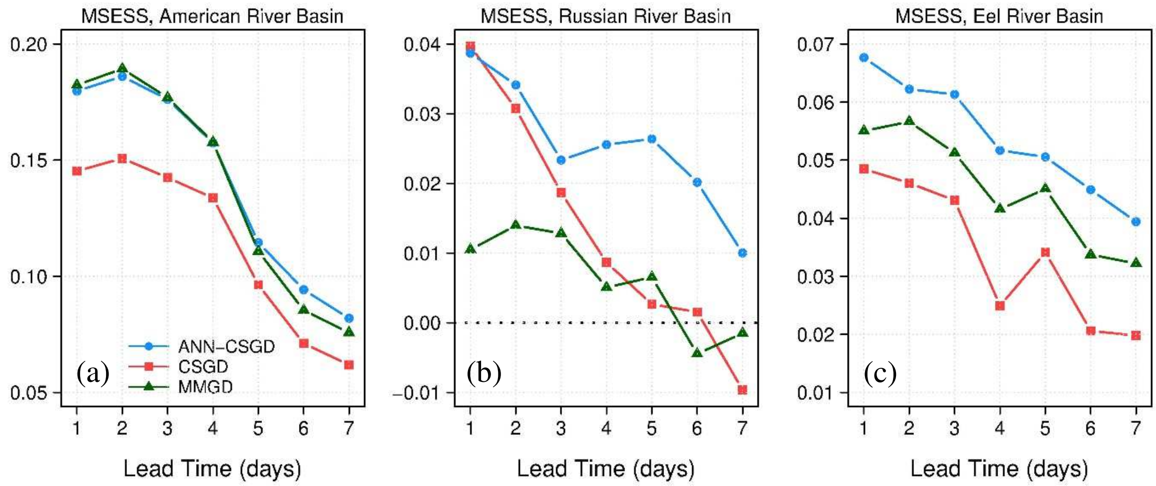

Fig. 8. As in Fig. 3 except for MSESS with benchmark models trained using 61-day window. GEFS ensemble mean forecast is considered as the reference.

	sub-basin ID	sub-basin name
American River Basin		
	NFDC1HUF	North Fork American River-North Fork Dam (upper)
	NFDC1HLF	North Fork American River-North Fork Dam (lower)
	FOLC1LOF	American River-Folsom Lake
Russian River Basin		
	WSDC1HOF	Dry Creek - Lake Sonoma
	GUEC1LOF	Russian River - Guerneville
Eel River Basin		
	DOSC1HUF	Middle Fork Eel River-Dos Rios (upper)
	DOSC1HLF	Middle Fork Eel River-Dos Rios (lower)
	FTSC1LUF	Eel River-Fort Seward (upper)
	FTSC1LLF	Eel River-Fort Seward (lower)

Table 1. Names and NWS IDs of study sub-basins of each river basin.