Probabilistic Predictions from Deterministic Atmospheric River Forecasts with Deep Learning

WILLIAM E. CHAPMAN,[a] LUCA DELLE MONACHE,[a] STEFANO ALESSANDRINI,[b] ANEESH C. SUBRAMANIAN,[c] F. MARTIN RALPH,[a] SHANG-PING XIE,[a] SEBASTIAN LERCH,[d] AND NEGIN HAYATBINI[a]

[a] *Scripps Institution of Oceanography, La Jolla, California*
[b] *National Center for Atmospheric Research, Boulder, Colorado*
[c] *University of Colorado Boulder, Boulder, Colorado*
[d] *Karlsruhe Institute of Technology, Karlsruhe, Germany*

(Manuscript received 7 May 2021, in final form 21 September 2021)

ABSTRACT: Deep-learning (DL) postprocessing methods are examined to obtain reliable and accurate probabilistic forecasts from single-member numerical weather predictions of integrated vapor transport (IVT). Using a 34-yr reforecast, based on the Center for Western Weather and Water Extremes West-WRF mesoscale model of North American West Coast IVT, the dynamically/statistically derived 0–120-h probabilistic forecasts for IVT under atmospheric river (AR) conditions are tested. These predictions are compared with the Global Ensemble Forecast System (GEFS) dynamic model and the GEFS calibrated with a neural network. In addition, the DL methods are tested against an established, but more rigid, statistical–dynamical ensemble method (the analog ensemble). The findings show, using continuous ranked probability skill score and Brier skill score as verification metrics, that the DL methods compete with or outperform the calibrated GEFS system at lead times from 0 to 48 h and again from 72 to 120 h for AR vapor transport events. In addition, the DL methods generate reliable and skillful probabilistic forecasts. The implications of varying the length of the training dataset are examined, and the results show that the DL methods learn relatively quickly and ~10 years of hindcast data are required to compete with the GEFS ensemble.

KEYWORDS: Atmospheric river; Error analysis; Numerical analysis/modeling; Uncertainty; Probability forecasts/models/distribution; Short-range prediction; Numerical weather prediction/forecasting; Artificial intelligence; Deep learning; Machine learning; Neural networks; Regression; Other artificial intelligence/machine learning

1. Introduction

Deterministic numerical weather prediction (NWP) systems are momentous forecast tools but are fatedly flawed in that they represent a single plausible realization of a possible weather future. Because of initial condition uncertainty, NWP deficiencies (e.g., subgrid parameterization approximations), and nonlinear error growth associated with the chaotic nature of the atmosphere, initially small forecast errors eventually result in weather predictions that are as skillful as random forecasts (Lorenz 1963). Dynamic ensembles prediction systems (EPS) are utilized to represent the evolution of multiple likely weather trajectories. Although multiple methods for creating dynamic ensembles exist (e.g., Epstein 1969; Hacker et al. 2011; Kirtman et al. 2014); most modern EPS systems create ensembles by running many realizations of the atmospheric state evolution, initializing each ensemble with slightly varied starting conditions or using varied model physics (e.g., Toth and Kalnay 1993). A range of possible weather scenarios results, providing probabilistic bounds for future weather.

Ensemble systems have greatly advanced in the modern era, yet raw EPS forecasts still suffer from significant systematic model bias that must be corrected with statistical postprocessing methods (Hemri et al. 2014). Often the systematic bias is particularly projected into the spread of the ensemble members, and under/over dispersive forecasts are common. This leads to a low correlation between the raw ensemble uncertainty and the forecast error; reducing the value of the model spread for forecast uncertainty quantification. Recent advances in deep-learning (DL) and machine-learning (ML) techniques have provided a significant step forward in calibrating statistical ensembles (e.g., Rasp and Lerch 2018).

The current study investigates ML's algorithmic ability to provide uncertainty quantification from single-member NWP model realizations, providing a valuable probabilistic measure of uncertainty, at a significantly lower real-time computational cost. This study leverages the methods developed in Rasp and Lerch (2018) (henceforth; RL2018) for ensemble calibration, but tailors them for the generation of probabilistic predictions from a historical set of single-member deterministic forecasts. Additionally, this study adds further algorithmic spatial awareness through vision-based DL methods (convolutional neural networks; CNN).

Recently, there has been a surge of interest in DL-based NWP postprocessing systems (see, Haupt et al. 2021; Vannitsem et al. 2021). Similar to RL2018, Ghazvinian et al. (2021), developed an NN-based scheme that minimizes the continuous ranked probability score (CRPS) from a prescribed parametric forecast distribution (censored, shifted

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-21-0106.s1.

Corresponding author: William E. Chapman, wchapman@ucsd.edu

DOI: 10.1175/MWR-D-21-0106.1

gamma) for rainfall prediction. Additionally, more flexible, distribution-free methods, have also been developed that leverage quantile-based probabilities transformed to a full predictive distribution (Scheuerer et al. 2020) or create direct approximations of the quantile function via regression based on Bernstein polynomials (Bremnes 2020).

Traditional ensemble model output statistics (EMOS) postprocessing schemes fit parameters of prescribed distributions (Gneiting et al. 2005). Here we retain the parametric distribution prediction framework but leverage multiple NN architectures to statistically link the CRPS loss function to the NWP system and train the networks through stochastic gradient descent. NNs offer some ready advantages over more established EMOS methods. For example, EMOS is rigid with respect to feature selection and requires explicit prescription of predictor–predictand relationships in their implementation. Alternatively, NNs offer extreme flexibility in incorporating and ingesting ancillary weather variables as predictors. NNs can quickly encode spatial information through convolution (e.g., Chapman et al. 2019), and temporal information with recurrent NNs or attention-network systems (e.g., Li et al. 2020; Theocharides et al. 2020). NNs allow the postprocessing system to readily encode predictor–predictor variable interactions and capture nonlinear variable interaction (Nielsen 2015). Additionally, Modern DL training schemes (i.e., dropout, regularization, early training stopping) have been implemented that systematically prevent algorithmic overfitting (Krogh and Hertz 1992; Srivastava et al. 2014).

Though many prominent probabilistic ensemble regression calibration methods exist (e.g., Gneiting et al. 2005; Raftery et al. 2005; Scheuerer and Hamill 2015) most leverage ensemble mean and spread characteristics rather than single-member deterministic models. Still some established postprocessing methods operate solely on deterministic fields, or can be adapted to operate on deterministic hindcasts (e.g., Lerch and Thorarinsdottir 2013; Robertson et al. 2013; Scheuerer and Hamill 2015; Wilks 2009; Wu et al. 2011), though most of these methods have been tested with the mean of a dynamic ensemble.

Analog-based techniques, in which historical stores of similar forecasts are used to estimate uncertainty, have been similarly formulated to provide statistically developed uncertainty in forecasts starting from a dynamical ensemble (e.g., Hamill and Whitaker 2006), or from single-member deterministic predictions (Delle Monache et al. 2013). Here we use the latter approach, the analog ensemble, modified for optimal rare event prediction (Alessandrini 2019) as a state-of-the-art baseline to assess the DL methods.

For this study, a newly developed 34-yr deterministic hindcast is leveraged. This long training dataset provides near unprecedented opportunity to correct for systematic forecast error. High-impact, 0–5-day (at 6-h intervals) probabilistic integrated vapor transport (IVT) prediction for landfalling North American West Coast (NAWC) atmospheric river (AR) events is the focus. Vertically integrated IVT is the characteristic metric that defines the strength of an AR (Ralph et al. 2018). IVT is a combined thermodynamic and momentum metric that integrates specific humidity and zonal and meridional components of the wind from 1000 to 300 hPa. Though we train our postprocessing systems on every value forecast, the study focuses on verifying IVT events above 250 $\mathrm{kg\,m^{-1}\,s^{-1}}$ (~85th percentile of observed IVT), because events below this threshold rarely result in extreme precipitation and are thus less societally impactful.

This study aims to test computationally efficient and flexible DL methods to estimate forecast uncertainty from a single-member NWP system for probabilistic, AR associated, IVT prediction. Uncertainty quantification is explored with DL methods by leveraging a distributional regression framework—which aims to develop the conditional distribution of the weather given a deterministic set of variables. The DL methods train by optimizing CRPS—a mathematically principled loss function for probabilistic forecasts (Camporeale and Carè 2021; Gneiting et al. 2005; Matheson and Winkler 1976). We pit the developed statistical uncertainty methods against modern state-of-the-art dynamic ensembles (calibrated and not) to test their skill. Additionally, we use feature permutation (McGovern et al. 2019) to explore the variable importance in the NN-based systems. Finally, we test the length of training data required to develop skillful forecasts, in order to determine the length of hindcast required to train a prediction system.

The remainder of the paper is structured as follows. Section 2 presents the dynamic forecast systems, the statistical ensemble and dynamic ensemble postprocessing methods used, and the ground truth data. Section 3 discusses the resulting forecast skill, examines input variable importance, and explores the required length of training data to quantify uncertainty reliably. A discussion of possible extensions follows, and we present conclusions in section 4.

2. Data, methods, and metrics

a. Data

1) Region of interest

Forecasting land-falling AR events over the NAWC is crucial (Ralph et al. 2020a; Wilson et al. 2020), and several contributions on AR forecast skill assessment are present in the literature (DeFlorio et al. 2018; DeHaan et al. 2021; Nardi et al. 2018; Nayak et al. 2014; Wick et al. 2013). ARs bring valuable precipitation to this drought prone region (Fish et al. 2019; Lamjiri et al. 2017) while simultaneously being the dominant driver of flooding across the NAWC (Corringham et al. 2019; Ralph et al. 2020b). One forecast product is a series of ensemble-based forecast imagery that shows a forecast lead time–latitude framework spanning the west coast of North America with illustrated IVT data from the National Centers for Environmental Prediction Global Ensemble Forecast System (NCEP-GEFS) ensemble, known as the AR landfall tool (ARLT, Cordeira et al. 2017; Cordeira and Ralph 2021) on the Center for Western Weather and Water Extremes (CW3E) web portal. ARLT shows NCEP-GEFS data in a pseudo-Hovmöller coastline-spanning framework, illustrating IVT data and providing situational awareness of the likelihood, intensity, duration, and timing of possible landfalling ARs. Because of the importance of forecasting landfalling ARs, this study examines the probabilistic forecast accuracy of landfalling grid points in every examined/postprocessed forecast system, and the two adjacent (moving westward) oceanic model points. Figure 1 shows the examined points in this study. All verification metrics and forecast assessment henceforth are diagnosed at these 144 landfalling locations.

2) GROUND TRUTH

IVT from the National Aeronautics and Space Administration's Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2), reanalysis is used as ground truth to diagnose forecast error and in ML training. MERRA-2 provides regularly gridded observations of the global atmosphere with assimilated satellite, upper air, remote sensing, and surface data. The MERRA-2 product is resolved on a 0.5° latitude × 0.625° longitude grid and interpolated to 21 pressure levels between 1000 and 300 hPa for IVT calculation (Gelaro et al. 2017; McCarty et al. 2016). Every forecast field in this study is regridded to the MERRA-2 grid using a first- and second-order conservative remapping scheme (Schulzweida et al. 2006) prior to ML training.

3) DYNAMIC MODEL FOR DETERMINISTIC FORECAST

Uncertainty quantification is generated for a version of the Weather Research and Forecasting Model that has been tuned specifically to western U.S. extreme precipitation (West-WRF, Martin et al. 2018). West-WRF is a near real-time model developed at CW3E that was run retrospectively to generate a 34-yr (1984–2019) hindcast spanning December through March of each year. In addition to providing a long training dataset, the model's consistency with the operational version provides an unprecedented opportunity for training machine-learning models on historical forecasts. The model is operationally run at a 9-km resolution, but we use first- and second-order conservative remapping to regrid these data to the common MERRA-2 grid, the model domain spans 25°–60°N and 150°–115°W. In this study, the December–March season is referred to as a water year (WY) with the year specified as the March of that year. For example, December 2018–March 2019 is referred to as WY2019. West-WRF is evaluated in the last three years of the dataset (WY2017, WY2018, and WY2019). threefold cross validation is leveraged in which the previous year is used as validation data and each of the three evaluated WYs is held out as testing data.

b. *Machine-learning-generated forecast uncertainty*

Four ML methods for uncertainty quantification are evaluated and compared against a dynamical ensemble's raw model output and a dynamical ensemble calibrated with a neural network. The computational cost of developing ML-based probabilistic predictions compared with dynamical ensembles is significantly less, both in real-time forecasting and for hindcast generation. Each method is described below. For each postprocessing system, the inputs are described in Table 1. Multiple deep-learning (DL) models are trained, with their architecture shown in Table 2, and model architecture diagrams shown in Fig. 1s in the online supplemental material. An extensive, although not exhaustive, hyperparameter search was conducted on two forecast lead times (48 and 96 h) to select model parameters by minimizing the model loss (described below) on the validation dataset. To aide future DL postprocessing development, the hyperparameter search method and model architecture intuition are described in the online supplemental material.

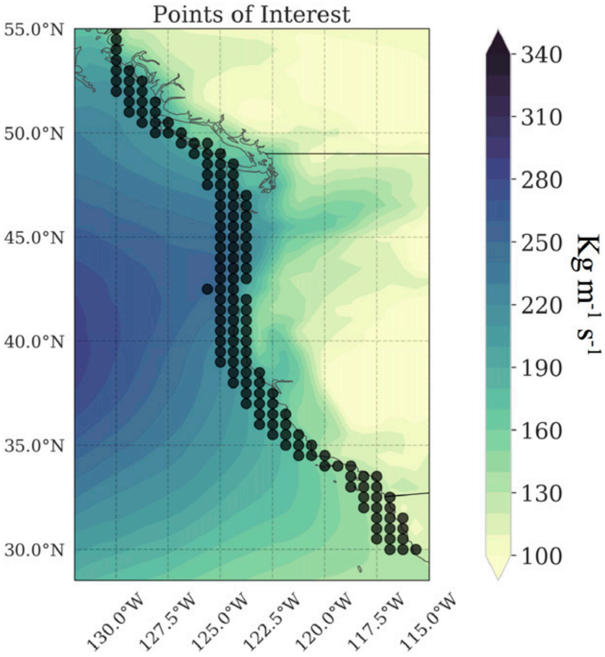

FIG. 1. Coastal evaluation locations and climatological (December–March 1984–2019) IVT (color fill).

1) NEURAL NETWORK WITH LOCATION EMBEDDINGS

Here the neural network (NN) is described, with a focus on the architecture, hyperparameters, and training routine utilized in this study. For a more complete exploration of the topic of NNs, the reader is referred to Nielsen (2015). The DL functionality is developed in python using the Tensorflow 2.0 (Abadi et al. 2016) library with the embedded Keras (Chollet et al. 2018) implementation. Following Chapman et al. (2019), an independent NN is trained for every forecast lead time. The input for this method is described in Table 1, the output is the mean and standard deviation of a probability distribution representing a probabilistic forecast for IVT at a specified (to the NN) coastal location.

Neural networks approximate nonlinear functions and processes (Nielsen 2015) through a series of feed forward matrix operations. NNs pass input predictor variables through a succession of "hidden" layers, resulting in a specified output layer. Each layer is described by the number of nodal points in that layer with the initial layer being the number of input variables. Prior to input, each predictor variable is standardized using the global (every point in the examined domain)

TABLE 1. Abbreviations and descriptions of all input variables.

Feature	Description	Model (input)
IVT	Integrated vapor transport (kg m^{-1} s^{-1})	CNN/NN/FCN/ GEFSnn/AnEn
P_{sfc}	Surface pressure (hPa)	CNN/NN/FCN/AnEn
U_{500}	500-hPa zonal wind (m s^{-1})	CNN/NN/FCN/AnEn
V_{500}	500-hPa meridional wind (m s^{-1})	CNN/NN/FCN/AnEn
Z_{500}	500-hPa geopotential height (m s^{-1})	CNN/NN/FCN/AnEn
IWV	Integrated water vapor (mm)	CNN/NN/FCN/AnEn
locID	Location identifier No.	NN/FCN/GEFSnn/AnEn

mean and standard deviation. In this work, a simple model with 2 hidden layers containing 30 and 40 nodes, respectively, is used. Nodes from adjacent layers are connected via model weights. The hidden nodal point values are determined by the sum of the product of associated model weights and the input values from the previous layer. Each nodal point is then "activated" by a nonlinear function before passing the variables to the following layer. We use a rectified linear unit (ReLU) activation function (Nair and Hinton 2010). The task of training an NN is to learn the optimal nodal weights, computed iteratively through backward optimization and gradient descent. In particular, each iteration seeks to minimize the cost of a specified loss function, by determining the gradient field of the weights and taking a small step in the direction opposite to this gradient. The NN leverages an Adam optimizer (Kingma et al. 2014) with a 0.005 training step that reduces by 10% on a validation plateau of 5 epochs (learning cycles). After 8 epochs of no decrease in validation error, training is ended. This typically resulted in ~40 training epochs.

The output model parameters (μ_{IVT}, σ_{IVT}) are estimated by minimizing the prescribed loss function of the continuous ranked probability score (CRPS) of a Gaussian distribution truncated at zero (as the magnitude of IVT cannot be negative). This loss function has been used in several notable EMOS postprocessing studies, largely in applications of wind speed prediction (e.g., Baran and Lerch 2015; Thorarinsdottir and Gneiting 2010; Thorarinsdottir and Johnson 2012). CRPS is discussed in section 3 and the analytical expression of the CRPS Gaussian distribution is provided in the appendix. Multiple CRPS loss functional families were attempted (logistic, lognormal, gamma, and Gaussian) [see Jordan et al. (2019) for these formulations], and a Gaussian truncated at zero provided the best fit, as determined by evaluation of the threshold-weighted CRPS and the shape of the stratified rank histogram, evaluated on the validation dataset–motivating the final loss function choice.

TABLE 2. CNN/NN parameters by category. The network layer components and their abbreviations are convolutional layer (Conv), max pooling (MP), addition (Add), concatenation (Concat), zero padding (Pad), crop (Crop), dense (Dense), input (Input), input embedding layer (Input Embed), and embedding vector (embedding). Here X is the batch size and N represents number of predictors (see Table 1).

Layer	Parameters	Activation	Norm	Shape
		Convolutional neural network		
Input	—	—	—	[X, 71, 57, N]
Pad	[1, 3]	—	—	[X, 72, 60, N]
Conv0	[3, 3, 16], 1, 1	LeakyReLU	BatchNorm	[X, 72, 60, 16]
Conv1	[3, 3, 16], 1, 1	LeakyReLU	BatchNorm	[X, 72, 60, 16]
MP	2	—	—	[X, 36, 30, 16]
Conv2	[3, 3, 32], 1, 1	LeakyReLU	BatchNorm	[X, 36, 30, 32]
Conv3	[3, 3, 32], 1, 1	LeakyReLU	BatchNorm	[X, 36, 30, 32]
Conv4	[3, 3, 32], 1, 1	LeakyReLU	BatchNorm	[X, 36, 30, 32]
Add	[Conv2, Conv4]	—	—	[X, 36, 30, 32]
Conv2dT	[2, 2, 16], 1, 1	LeakyReLU	BatchNorm	[X, 72, 60, 16]
Concat	[Conv2dT, Conv0]	—	—	[X, 72, 60, 32]
Conv5	[3, 3, 16], 1, 1	LeakyReLU	—	[X, 72, 60, 16]
Conv6	[3, 3, 1, 2], 1, 1	Linear	—	[X, 72, 60, 2]
Crop	[1, 3]	—	—	[X, 71, 57, 2]
Conv7	[3, 3, 32], 1, 1	LeakyReLU	—	[X, 71, 57, 32]
Output	[3, 3, 1, 2], 1, 1	Linear	—	[X, 71, 57, 2]
		Neural Network (NN and GEFS$_{nn}$)		
Input Embed	—	—	—	[X, 1]
Embedding	2	—	—	[X, 1, 2]
Input Main	—	—	—	[X, N]
Concat	[Input, embedding]	—	—	[X, N+2]
Dense1	[30]	ReLU	—	[X, 30]
Dense2	[40]	ReLU	—	[X, 40]
Dense3	2	Linear	—	[X, 2]
Dense4	[32]	ReLU	—	[X, 32]
Dense5	2	Linear	—	[X, 2]

Additionally, location embeddings are used as an input to the NN. Embeddings are responsible for encoding a vectorized version of discrete information, in this case, an ID number specified for each of the 144 locations (1–144). These vectors are learned and updated during training, but do not correspond to any real variable. This allows the network to learn customized nodal weights for each lat/lon location while still benefitting from the relationships learned at every location (Guo and Berkhahn 2016; RL2018). The vector length is specified as part of the network architecture. By conducting a hyperparameter search, it was determined that two latent variables (vector length) provided the greatest model performance without adding additional model parameters. Thus, one NN can be trained for the entire domain and the bias specific to each location (e.g., topographically or latitudinally driven NWP biases (Gowan et al. 2018)) can be corrected.

Because of the dataset spanning multiple decades, we have noticed oscillations in NWP model skill and bias. To mitigate the potential effects of secular climate change, or slowly varying decadal variability, we institute a customized training regime similar to model transfer learning (Torrey and Shavlik 2010). The model is first trained on the full 32-yr training set (34 years, minus 1 year of validation and 1 year of test). Next those model weights are saved and frozen from updating, a final layer is concatenated to the network and we "fine-tune" on just 1 WY, two years prior to the testing dataset (one year prior to the validation). For example, an NN that is tested on WY2019 is initially trained on WY1985-WY2017, then frozen, a new layer is concatenated, and it is then tuned on WY2017. The training schedule is exactly similar to that described above with identical criterion for ending the training. While the mean prediction is relatively unaffected by fine-tuning, this was found to significantly improve predicted spread statistics (not shown).

2) Convolutional neural network

The convolutional neural network (CNN) architecture is shown in Table 2 and online supplemental Fig. 1s. The architecture is adapted from a U-NET (Long et al. 2015). The U-NET architecture is ideal for this task as it passes less abstracted information from shallow layers in the CNN to deep layers [see Ronneberger et al. (2015) for more detail]. Additionally, versions of this architecture have been shown to significantly reduce IVT deterministic forecast error (Chapman et al. 2019). The computational details are similar to those described in the above NN. Again, a Gaussian distribution truncated at zero provided the best skill on our cross-validated dataset and was selected as the loss function. The CNN utilizes an Adam optimizer (Kingma et al. 2014) with a 0.0001 training step that reduces by 40% on a validation plateau of 2 epochs. After 8 epochs of no decrease in the validation error, training is ended. This typically resulted in ~50 training epochs. The CNN uses identical predictors to the NN, the largest difference being that CNNs operate on images by updating weights associated with convolutional kernels that are slid across input image fields and trained to highlight salient forecast features. In the CNN, the entire spatial domain is fed to the model at training for each independent forecast rather than independent training data for each location (as in the NN). The goal of model training is thus to learn the optimal weights in the convolutional kernels that minimize CRPS by best predicting μ_{IVT} and σ_{IVT} for every pixel in the image. The network is trained to optimize predictions in the entire model domain; however, in the following analysis, the CNN is evaluated only at the aforementioned coastal locations. The implications of this choice are discussed in section 4. The reader is referred to Zhang et al. (2021) for a theoretical description of CNNs and convolutional kernel training.

The same training regimen is utilized for the CNN with one additional convolutional layer concatenated to the end of the network after freezing all previous layers. The CNN is then fine-tuned on the WY 2 years previous to the testing WY.

3) Fully connected distributional regression

We include, as an additional baseline, a parametric prediction method that is conceptually similar to traditional distributional regression performed via EMOS systems. We implement a fully connected neural network (FCN) with no hidden layers, trained using CRPS estimated from a Gaussian distribution truncated at zero. The FCN, without inputting ancillary predictor fields, is conceptually equivalent to a global EMOS scheme, but differs in the parameter estimation approach. Here, as it is easily implemented and also demonstrated improvement in minimizing CRPS, we include all of the predictor variables that are supplied to the NN and CNN, and the same location embedding vector supplied only to the NN (see Table 2). The FCN leverages an Adam optimizer (Kingma et al. 2014) with a 0.005 training step that reduces by 10% on a validation plateau of 5 epochs (learning cycles). After 8 epochs of no decrease in validation CRPS, training is ended. To create as similar training conditions to the CNN and NN we fine-tune the FCN system by loading the model weights from the model trained on the 34 years of data, reducing the learning rate to 0.000 01 and training again on data from 2 years prior to the testing data (thus, 1 year prior to the validation data). The FCN serves to assess the value of the nonlinear predictor–predictand relationships in both the NN and CNN. A local FCN implementation was also tested but showed poorer forecast performance and calibration for high-threshold IVT events (250, 350, and 500 $\mathrm{kg\,m^{-1}\,s^{-1}}$).

4) The analog ensemble for rare events

To compare the DL-based statistical ensemble to a state-of-the-art ML-based ensemble method, an analog ensemble (AnEn; Delle Monache et al. 2013) coupled with a recent bias correction innovation for rare events (Alessandrini 2019) is constructed. The AnEn generates an ensemble by exploiting an issued NWP forecast, a history of forecasts made by the same model, and the corresponding resultant observed weather. When a forecast is issued, the AnEn is tasked with searching for analogous forecasts in the historical record, it then uses the corresponding resultant observations for the analogous forecasts as the ensemble prediction.

Let $f(y|\mathbf{x}^f)$ be the probability distribution of the observed value y of some predicted quantity given a model prediction $\mathbf{x}^f$; $\mathbf{x}^f$ is a vector of k predictor variables issued from the NWP forecast $(\mathbf{x}^f = x_1^f, x_2^f, x_3^f, \ldots, x_k^f)$ that includes the desired forecast variable (IVT) and a suite of other relevant predictor variables (Table 1). AnEn uses a distance function to then identify the closest analogs to $\mathbf{x}^f$ from a database of previously issued forecasts $\mathbf{x}^j$. The ground truth observations y_j from the previously issued forecasts form the ensemble. Like the NN, the analog ensemble is a point-based method, and only forecasts at a given location are used to form this ensemble. The distance function is given by

$$d(\mathbf{x}^f, \mathbf{x}^j) = \sum_{i=1}^{k} \frac{w_i}{\sigma_i} \sqrt{\sum_{r=-t}^{\tilde{t}} (x_{i,t+r}^f - x_{i,t'+r}^j)^2},$$

where the current NWP forecast $\mathbf{x}^f$ is valid at time t at a given forecast location; $\mathbf{x}^j$ is the analog at the same location with the same forecast lead time but valid at a past time t', k defines the number of predictor variables weighted by w_i, σ_i is the standard deviation of the time series of past forecasts of a given variable at the same location and forecast lead, and $\tilde{t}$ is equal to one-half of the number of additional times over which the metric is computed. Accounting for a forecast window ($\tilde{t}$) ensures that the trend of the examined variables is considered and has been shown to be valuable for minimizing forecast error (Alessandrini et al. 2015). This temporal trend gives the AnEn a potential predictor advantage over the CNN and NN.

The AnEn is run and optimized (through w_i selection) at every location individually. By leveraging small predictor sets k, a full brute-force optimization search can be conducted by trying every permutation of w_i, and examining mean-square error on an independent validation dataset subject to $\sum_{i=1}^{k} w_i = 1$, where $w_i \in [0, 0.1, 0.2, \ldots, 1]$, which has been shown to improve predictions in several past studies (e.g., Alessandrini et al. 2015; Junk et al. 2015). Figure 3s in the online supplemental material shows the distribution of predictor variable weights across every location. The predictor variable of interest (IVT) is dominantly weighted, as expected, with the remaining variables accounting for ~10% of variability each. To match the dynamic ensemble, we specify the return of 21 ensemble members. The numbers of ensembles were varied, but the results showed little sensitivity between 10 and 50 members.

The AnEn has a tendency to introduce a conditional negative bias when predicting events in the right tail of the forecast (extreme or rare events), which are the focus of this study. To ensure that the AnEn is optimized to correctly forecast rare events (as our target is to most accurately forecast ARs), and to set the best baseline possible for the DL/ML methods, we leverage the modifications to the AnEn as presented in Alessandrini (2019) for conditional bias correction. The proposed method is based on a linear regression analysis between forecast and observations performed independently at each lead time and location. Each member is adjusted by adding a factor proportional to the difference between the target forecast and the mean of the past analog forecasts multiplied by the coefficient obtained after the linear regression analysis. We refer the reader to Alessandrini (2019) to examine additional details of the bias adjustment algorithm. A threshold value of 300 kg m^{-1} s^{-1} (or units) of IVT (~90th percentile of station observations) is used to enact the bias correct. This value was determined by incrementing the value of the threshold from 250 to 500 (in 50-unit increments) and minimizing the CRPS on the validation datasets.

5) Raw and calibrated Global Ensemble Forecast System

We assess the probabilistic skill of the NN-based methods (FCN, NN, and CNN) against a state-of-the-art dynamical ensemble: the operational 0.5° latitude × 0.5° longitude NCEP-GEFS, version 11.0.0, from 1 December to 31 March of WYs ending in 2017, 2018, and 2019. The GEFS includes 21 members (20 perturbed initial conditions and 1 control member). These data were obtained from The Interactive Grand Global Ensemble (TIGGE) data portal at the European Centre for Medium-Range Weather Forecasts. WY17 and WY18 contained 70 missed forecasts in the TIGGE system so these were then calculated from 1° latitude × 1° longitude GEFS data obtained from the National Centers for Environmental Information Data Archive and were simply interpolated to the 0.5° grid spacing. We additionally apply the same NN postprocessing to the GEFS ensemble system as described in RL2018, by minimizing CRPS while using the ensemble IVT mean and standard deviation as predictors and leveraging the embedded forecast location. This algorithm was shown to outperform the best traditional postprocessing methods (RL2018). The NN applied to the GEFS system (GEFS$_{\mathrm{nn}}$ henceforth) is subject to the same train/test split as described above in which the network is trained in a threefold cross-validation manner in which the previous year is used as validation data and each of the three evaluated WYs are held out as testing data. Table 1 describes the input variables. Table 2 describes the utilized network.

3. Results

In this section, we evaluate the predictive performance of the postprocessing systems and raw dynamic ensemble, all based on the cross-validated testing data from WY17, WY18, and WY19. For an introduction to the evaluation methods and underlying theory, see the appendix. We use skill scores [$\mathrm{SS} = 1 - (S/S_{\mathrm{ref}})$, $-\infty < \mathrm{SS} \leq 1$], where positive or negative values are shown respectively to be more or less skillful than the reference forecast S_{ref}. Python code for reproducing the results and models is available online (https://github.com/WillyChap/ARML_Probabilistic).

This analysis evaluates six forecast systems, termed: AnEn, FCN, CNN, NN, GEFS, and GEFS$_{\mathrm{nn}}$, evaluated from 0 to 120 h (in 6-h intervals). The AnEn, FCN, NN and CNN systems are built from a historical dataset including a single deterministic forecast (based on the dynamical model West-WRF), whereas the GEFS is built from the raw GEFS EPS forecast. Additionally, the results focus on high-impact IVT

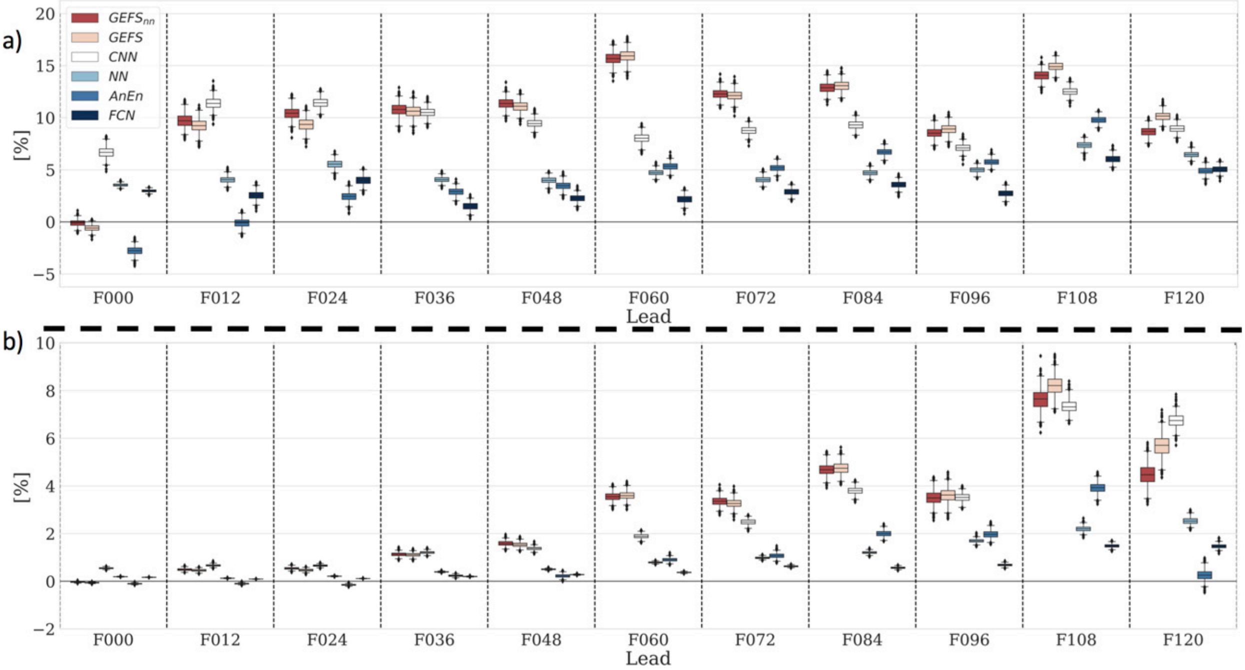

FIG. 2. (a) Root-mean-square error and (b) Pearson correlation skill scores against the forecast lead time for the ensemble mean or predicted mean from each forecast system. The West-WRF reforecast is used as the reference forecast, with positive values showing percent improvement. Shown predictions include GEFSnn (dark red), GEFS (light red), CNN (white), NN (light blue), AnEn (blue), and FCN (dark blue). The error bars indicate the 95% bootstrap confidence intervals (where $n = 1000$).

events that are likely to cause NAWC precipitation. We threshold at 250, 350, and 500 kg m^{-1} s^{-1} (units) of IVT, which guarantees local AR conditions. This represents percentile values of ~85th, ~93rd, ~97th, respectively. Right-tailed events are traditionally more difficult for postprocessing methods to improve upon. Although higher-impact events exist (500+ threshold), their rarity prevents robust probabilistic statistical comparisons (Wilks 2010), thus the above stated thresholds are evaluated. All results shown henceforth are for the independent testing data years (WY2017, WY2018, and WY2019).

The GEFS and AnEn consist in a set of ensemble members while the FCN, NN, CNN, and GEFS$_{nn}$ include the mean and standard deviation of a truncated Gaussian distribution. To ensure a fair assessment, the following verification is conducted by computing the mean and standard deviation for every individual forecast and randomly sampling from that distribution to create pseudoensembles. The exact ensembles for the GEFS and AnEn were also assessed, but the resulting analysis was not significantly changed.

a. *Deterministic predictions*

Figure 2 shows the root-mean-square error (RMSE) (Fig. 2a) and Pearson correlation (PC) (Fig. 2b) of the deterministic forecasts (ensemble mean) using the West-WRF raw reforecast as the reference forecast in 12-h increments. Although the primary focus of this work is to evaluate the probabilistic skill of the ensemble forecast methods, we first demonstrate the deterministic skill of the forecast systems. For each method, this is taken as the mean of the predictive distribution/ensemble. The authors realize that this is not a direct comparison because the postprocessing methods were not applied to the same forecast baseline (i.e., West-WRF vs GEFS).

At all lead times, the GEFS ensemble mean (red) forecast is more skillful than the West-WRF deterministic model from which the AnEn, FCN, NN, and CNN are developed. The GEFS$_{nn}$ (red) and Raw GEFS (light red) systems ensemble mean performance differences are not statistically significant and the GEFS$_{nn}$ ensemble calibration is largely just influencing the ensemble spread statistics (discussed further below) to improve the forecast skill. The CNN (white) is resulting in the largest improvements of West-WRF reforecast when compared to AnEn, FCN, and NN at all lead times in both PC and RMSE, with a stable improvement of ~10% at every lead time for RMSE while improving the correlation from 1% to 7% with greater improvements at the longer lead times. The NN (light blue) also improves the forecast at every lead time at ~5% for RMSE and improves correlation from 0% to 3% across lead times. The NN generally outperforms the FCN, showing the value of the adding nonlinear activations, with statistically significant improvement at lead times 0, 12, 36, 48, 72, and 96 h for RMSE and at every lead time past 12 h for PC. The NN is run locally with embedded location identifier (ID) information, and therefore does not have the benefit of a global field view (like the CNN), this additional spatial feature helps to quantify the difference in mean statistics. The CNN corrections result in forecasts that significantly outperform

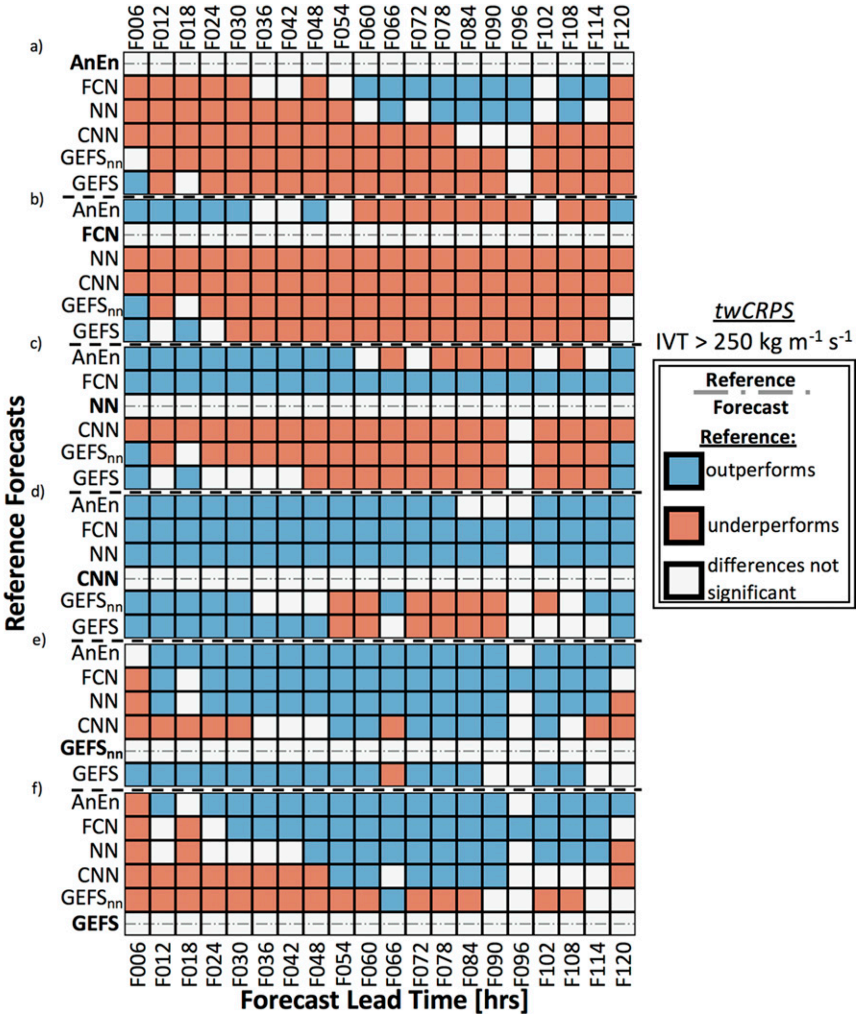

FIG. 3. Two-sided Diebold–Mariano test using twCRPS (threshold = 250 kg m^{-1} s^{-1}) for the five forecast systems (a) AnEn, (b) FCN, (c) NN, (d) CNN, (e) GEFSnn, and (f) GEFS. The reference forecast is indicated with a gray dash–dot line. Blue shading indicates that the reference forecast significantly outperforms the row's forecast, red indicates that the reference forecast significantly underperforms the row's forecast, and white indicates that the reference forecast and the row's forecast differences are not statistically significant. Significance is determined by examining the test p values after controlling for the false discovery rate at the level $\alpha_{\mathrm{FDR}} = 0.05$.

those from the FCN and NN at every lead time for both PC and RMSE metrics. The AnEn (blue) initially negatively impacts the analysis forecast (F000) but improves the skill of the deterministic forecast from 24 to 120 h with similar statistics as the NN.

b. *Diebold–Mariano test (under AR conditions)*

For comparative model assessment, proper scoring rules are leveraged to simultaneously evaluate the calibration and sharpness of forecasts (Gneiting and Raftery 2007). Proper scoring rules assign a numerical score to pairs of probabilistic forecasts and observations such that the expected score is optimized if the true distribution of the observation is issued as a forecast. Here, two negatively oriented (a smaller value is better) proper scoring metrics are examined, the Brier skill (BS; Brier 1950) and twCRPS (Gneiting and Ranjan 2011).

Figure 3 shows the results of the two-sided Diebold–Mariano (DM, Diebold and Mariano 2002) test calculated on the basis of mean threshold-weighted CRPS (twCRPS, Gneiting and Ranjan 2011) over all the samples at each lead time (0–120 h) as the determining metric, with a threshold set to 250 kg m^{-1} s^{-1} (units) of IVT. Simultaneous interpretation of the test results across lead times requires that we account for test multiplicity. We do so by controlling the false discovery rate at α_{FDR} = 0.05 (see the appendix for details)

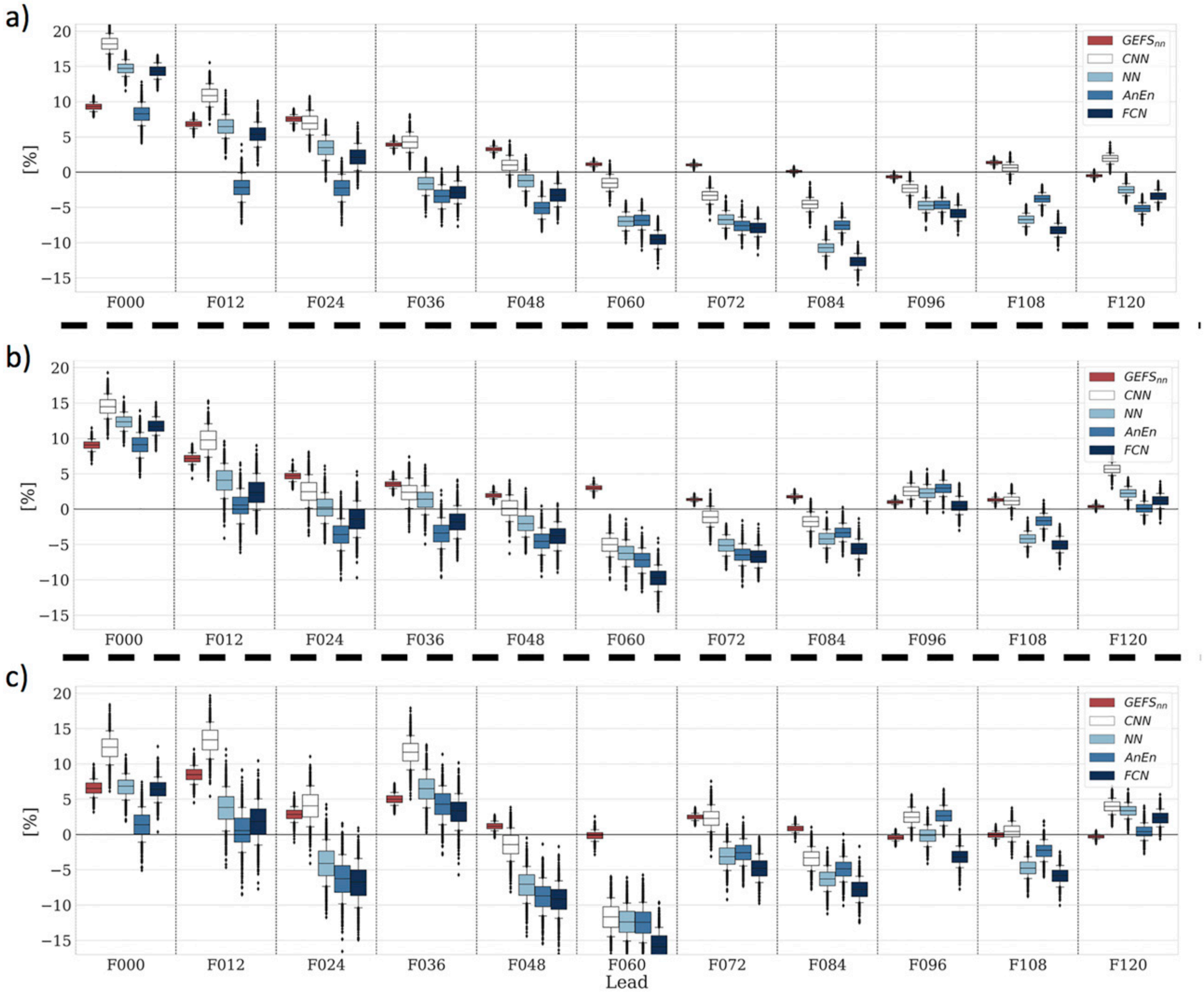

FIG. 4. Brier skill score at thresholds of (a) 250, (b) 350, and (c) 500+ kg m^{-1} s^{-1} forecast units of IVT against the forecast lead time for the ensemble mean or predicted mean from each forecast system. The GEFS ensemble is used as the reference forecast, with positive values showing percent improvement. Shown predictions include GEFSnn (dark red), GEFS (light red), CNN (white), NN (light blue), AnEn (blue), and FCN (dark blue). The error bars indicate the 95% bootstrap confidence intervals (where n = 1000).

(Benjamini and Hochberg 1995; Wilks 2016). Figures 3a–e leverage a separate reference forecast (AnEn, FCN, NN, CNN, GEFS$_{nn}$, and GEFS, respectively) to compare with each other forecast. The reference is shown in boldface type, and a gray dash–dotted line is used to delineate the reference further. Red panels indicate that the reference underperforms the compared forecast, blue panels indicate that the reference forecast outperforms compared forecast, and white panels show that the difference is not statistically significant between the two systems. Panels with large blue swaths are better forecast systems than the compared postprocessing system.

It is apparent that, when comparing individual forecast systems built from West-WRF (AnEn, FCN, NN, CNN: Figs. 3a–d, respectively), the forecasts from the CNN generally outperform the other systems. Forecasts from the CNN significantly outperform forecasts from the AnEn at all lead times except from 84 to 96 h in which the differences are not statistically significant. The forecasts from the CNN significantly outperform the GEFS ensemble system or the differences are statistically significant not at all lead times except the 66-h lead forecasts. Additionally, at 0–48 and 96–120 h the CNN is competitive with the calibrated GEFS forecast (GEFS$_{nn}$). The GEFS$_{nn}$ systematically significantly outperforms the GEFS system for all short lead times (0–60 h) and outperforms or is not significantly different from GEFS from 72 to 120 h. We show similar figures for the 350 and 500 kg m^{-1} s^{-1} IVT thresholds in Figs. 4s and 5s in the online supplemental material. While still generally outperforming each forecast system from a Brier skill score (BSS; Brier 1950) and twCRPS perspective, the CNN struggles more to improve over the GEFS with high-impact events at the longer lead times (3–5 days); this is discussed further in section 4. The NN is shown to significantly outperform the FCN for the 250 kg m^{-1} s^{-1} threshold at all lead times and is generally more skillful (although not always significantly) at the 350 and 500 kg m^{-1} s^{-1} thresholds (supplemental Figs. 4s and 5s).

c. Brier skill score and CRPS

Figure 4 shows the BSS at three threshold levels (250, 350, and 500 kg m^{-1} s^{-1}) of IVT for forecasts from 0 to 120 h using the GEFS forecast BS as a reference metric. The GEFS$_{\mathrm{nn}}$ may leave the ensemble mean forecast relatively unaffected (Fig. 2), but it improves the GEFS forecast by calibrating its probabilistic skill (Fig. 4, dark red). Within the first 36 h the CNN outperforms or shows insignificant differences from the GEFS$_{\mathrm{nn}}$ forecast system for every threshold value (Fig. 4: white vs red). Between 96 and 120 h the NN-based ensembles again compete with or outperform the GEFS forecast systems. The NN is able to outperform the AnEn at most lead times out to 48 h and the two methods have similar performance from 60 to 120 h. At lead times between 96 and 120 h and higher-impact events, the AnEn and NN show similar skill to the CNN (Figs. 4b and 4c). Although the differences are not always statistically significant, the NN generally outperforms the FCN at most lead times. To complement Figs. 3 and 4, Fig. 2s in the online supplemental material shows the twCRPS skill score at the same three threshold levels (250, 350, and 500 kg m^{-1} s^{-1}) of IVT for forecasts from 0 to 120 h using the GEFS forecast as a baseline metric. The twCRPS tells a very similar story to the BSS and DM test.

In both BSS and twCRPSS, between 60 and 84 h, we note a drop in skill between the AnEn or the NN-derived ensembles that are built on a deterministic prediction and the dynamic ensemble system. Figure 2 shows that this skill is largely derived from a comparative discrepancy in deterministic forecast skill between the two forecast systems used to build these ensembles (GEFS and West-WRF). Again, we stress that each method is built from different dynamic forecast models and that this does not represent a detriment added by the postprocessing methods (see Fig. 2). It appears that this comparatively larger forecast skill difference (see Fig. 2, hours 0–48 vs hours 60–84) is responsible for the difference of skill in the interim forecast window.

d. Spread/skill

Figure 5 shows the binned spread–skill plots of the evaluated models partitioned into 0–48 h and 54–120 h for forecasts of IVT. In the first 48 h (Figs. 5a–f) the GEFS model (light red) is severely overconfident (Fig. 5a). The AnEn faces the opposite problem and appears to overestimate values of forecast uncertainty (Fig. 5c). The remaining models (CNN, NN, and GEFS$_{\mathrm{nn}}$) provide statistically consistent forecasts and indicate that they are able to capture the flow-dependent forecast uncertainty because their spread dependably reflects the forecast error variance. The CNN and the NN are virtually indistinguishable and perfectly calibrated while the GEFS$_{\mathrm{nn}}$ does reflect small conditional bias toward the highest binned events. Across all tested models, forecasts from 54 to 120 h (Figs. 5g–l) are less calibrated, but still represent a good flow-dependent forecast uncertainty relationship. The GEFS, FCN, NN and CNN forecasts are overconfident and contain a slight low bias. The AnEn is the best calibrated forecast for the right-tailed forecast error events, followed closely by the NN and GEFS$_{\mathrm{nn}}$ forecast systems, showing that these systems capture the flow-dependent forecast uncertainty since the spread dependably reflects the forecast error variance.

e. Stratified rank histograms

Figure 6 shows the stratified rank histograms of the evaluated models partitioned into the 0–48 h and 54–120 h. The histograms are stratified into three categories: [250–350), [350–500), and 500+ kg m^{-1} s^{-1}. For the analog ensemble and GEFS system we use the 21 ensemble members generated by each system. To be consistent we sample 21 random pulls from the distribution described by each individual forecast to form a pseudo ensemble and build the stratified rank histogram from those forecasts. Bröcker and Smith (2007) and Siegert et al. (2012) demonstrated that when stratifying on the ensemble forecast mean (or other ensemble derived statistics), a uniform rank histogram distribution is not necessitated to show a calibrated forecast ensemble system. Bellier et al. (2017) offered a graphical test to check the true calibration shape through random sampling of ensemble members that serve as pseudo-observations to determine the shape of a perfectly calibrated forecast ensemble. After conducting this test for the prescribed IVT thresholds [250–350), [350–500), and 500+ kg m^{-1} s^{-1}, it was determined that a uniform stratified distribution is optimal (not shown). To aid in interpretation, Table 3 shows the reliability index (RI; Delle Monache et al. 2006) for the stratified rank histograms. Here,

$$\mathrm{RI} = \sum_{i=1}^{K+1} \left| f_i - \frac{1}{K+1} \right|,$$

where f_i is the frequency of observations in the ith rank and K is the number of ensembles that were forecast.

We first examine the 0–48-h forecasts. The most apparent errors are in the GEFS forecast ensemble system, which is highly underdispersive/overconfident, and a general lack of statistical consistency. Applying a neural network with location embeddings to this dynamic ensemble (GEFS$_{\mathrm{nn}}$) results in a very well calibrated forecast for AR events (Fig. 6b). This confirms that GEFS$_{\mathrm{nn}}$ is largely correcting the forecast spread while leaving the ensemble mean relatively unchanged (See Fig. 2). The AnEn is overdispersive/underconfident from 0 to 48 h. Despite developing all of the spread characteristics from data alone (unlike the GEFS$_{\mathrm{nn}}$), the NN and the CNN (Figs. 6d and 6e) both represent well calibrated probabilistic distributions. There is a small indication of under prediction for both of these systems, exacerbated further in the CNN. The FCN struggles to calibrate the right-tailed events, showing a high bias. This demonstrates the important nonlinear information in the predictor fields as the NN shows a very well calibrated ensemble, with the same input predictors. The RI values in Table 2 indicate that the NN is largely more calibrated than the CNN system though all postprocessing methods outperform the raw GEFS calibration.

The 54–120-h forecasts struggle more with statistical consistency. The GEFS ensemble again shows signs of over

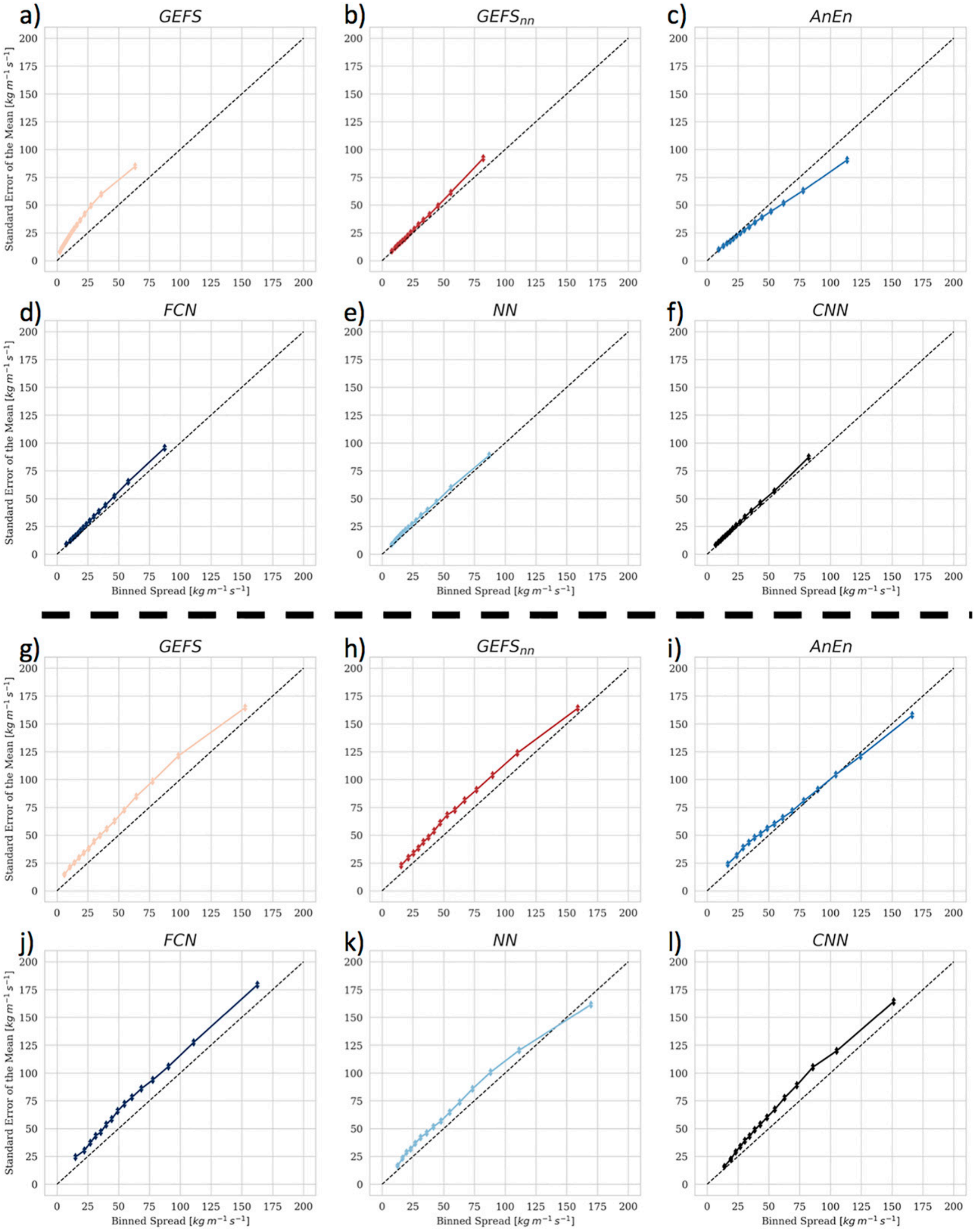

FIG. 5. Binned spread–skill plots for forecasts over (top),(top middle) 0–48 and (bottom middle),(bottom) 56–120 h. Error bars indicate the 95% bootstrap confidence interval (n = 1000), and the 1:1 dotted line indicates a perfect spread–skill line. For each plot, ensemble spread is binned into 15 equally populated class intervals. Shown prediction systems include (a),(g) GEFS (light red); (b),(h) GEFSnn (dark red); (c),(i) AnEn (blue); (d),(j) FCN (dark blue); (e),(k) NN (light blue); and (f),(l) CNN (black).

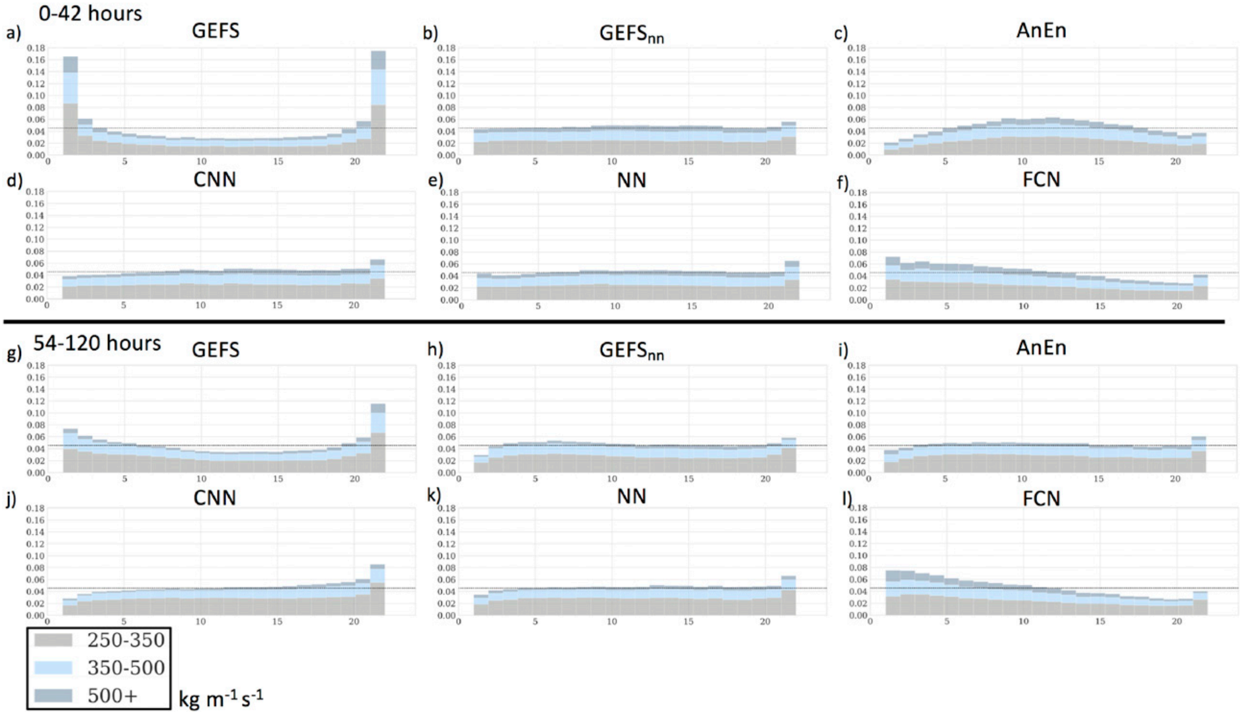

FIG. 6. Stratified rank histograms. The x axis is number of ensembles + 1, and the y axis is fractional occurrence of rank. A perfectly calibrated forecast is uniform with amplitude at the shown horizontal dotted line for forecasts over (a)–(f) 0–48 and (g)–(l) 56–120 h, stratified on forecasts [250–350), [350–500), and 500+. A total of 21 ensembles are drawn from the distribution representing the mean and Gaussian spread from individual forecast systems (as labeled).

dispersion, coupled with a high bias (Fig. 6g), but is less affected by the overconfidence than at shorter lead times. The GEFS$_{nn}$ acts to fix the dispersion and produces a relatively calibrated ensemble though there are signs of a low bias. The NN-based methods and AnEn again produce fairly calibrated ensembles. The CNN struggles prominently with a low bias (Fig. 6j and Table 2). The AnEn and NN are relatively indistinguishable and produce a well calibrated statistical ensemble. The FCN again shows a severe high bias and offers a good contrast to the NN and CNN methods.

f. Variable importance

To investigate rankings of input variable importance in the FCN, NN, and CNN we use a single-pass permutation-based measure introduced by Breiman (2001). The goal is to determine the level of twCRPS deterioration when the statistical link between forecast field F_j and the target observation y_j is broken by randomly permuting each F_j, one at a time, over all forecast samples. We use the mean twCRPS(τ = 250 kg m^{-1} s^{-1}) of the nonpermuted input features as a relative reference baseline. The reference twCRPS baseline is recalculated at each lead time to prevent skewing the variable importance via dependence on model lead-time forecast skill. If performance deteriorates significantly (high values in Fig. 7) the variable is considered to be important. The single-pass permutation algorithm is described in detail in the appendix. Figures 7a–f respectively show the relative variable permutation importance at forecast lead times of 12, 24, 48, 72, 96, and 120 h for a twCRPS with threshold τ = 250 kg m^{-1} s^{-1} (units) of IVT.

The most important variable across the three systems, at all lead times, is the NWP model output IVT. IVT's importance, relative to other variables, diminishes at longer forecast horizons. The CNN considers integrated water vapor (IWV) as the second most important variable, accounting for model degradation of 6%–10% across all forecast lead times. IWV is an integrated component of specific humidity calculated at the same model levels as IVT. Its relative variable importance indicates that the CNN is learning some error dependence that is contained solely in the thermodynamic component of IVT. The CNN shows minor dependence on the remaining forecast variables. We note that the CNN does not leverage location ID as a predictor (see Table 2). The CNN does not show much sensitivity to the meridional or zonal components of the 500-hPa wind, though these variables have been shown to impact forecast error state in other NWP systems (Stone et al. 2020).

TABLE 3. Stratified rank histogram reliability index by method and for forecasts aggregated over lead times 0–48 h and 54–120 h. Boldface type indicates the best reliability index score.

	AnEn	FCN	NN	CNN	GEFS$_{nn}$	GEFS
F00–F048	0.23	0.24	0.07	0.097	**0.056**	0.50
F054–F120	**0.081**	0.28	0.084	0.15	0.09	0.25

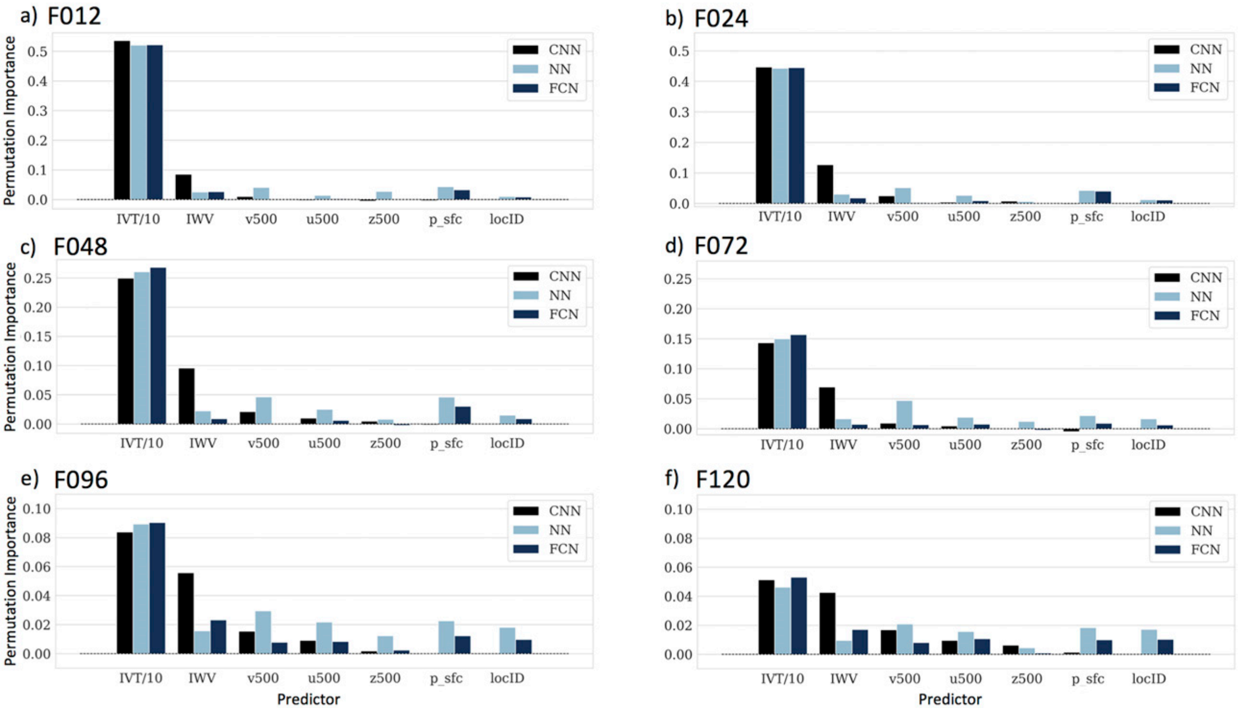

FIG. 7. Relative permutation importance for the input predictors (defined in Table 1) in the CNN (black), NN (light blue), FCN (dark blue) postprocessing systems for lead times (a) 12, (b) 24, (c) 48, (d) 72, (e) 96, and (f) 120h, using twCRPS with threshold set to 250 kg m^{-1} s^{-1}. Note the changing scale on the y axis. The CNN system does not leverage a locID predictor. IVT relative predictor importance is divided by 10.

The FCN and the NN leverage the same input predictors and the difference in variable importance is an indication of the important nonlinear predictor-predictand relationships learned by the NN. In general, the addition of the nonlinearity, spreads predictor sensitivity more evenly across multiple variables, leading to greater importance of several variables. The second most important variable for the FCN and NN, across most lead times, is the meridional component of the wind at 500 hPa V_{500}. This is intuitive as modulation to the V_{500} variable is leveraged to diagnose storm track variability (Chang and Yu 1999; Wirth et al. 2018) and is an indication of amplification in the synoptic scale control (via large-scale troughing or ridging) over the AR system. Meridional-oriented ARs tend to be stronger in magnitude (higher IVT) and result in greater precipitation (Cobb et al. 2021; Hecht and Cordeira 2017). Interestingly, the FCN and NN systems both show a sensitivity to surface pressure, which is not learned in the CNN. The location ID (input via an embedding layer) accounts for a 1%–2% model degradation across all lead times and is twice as important in the NN than the FCN, indicating some nonlinear dependence on spatially dependent information.

g. *Length of training*

The NN-based methods and AnEn, run in real-time, have a significantly lower cost compared to the GEFS system as only a single deterministic forecast is required to produce this probabilistic prediction. However, a longer training dataset was used compared to the stable GEFS system (version 11.0.0 of the GEFS model was only stable for 3 years). To test the impact of the length of training (deterministic hindcast years that are required) needed to achieve comparative skill we retrain the CNN, NN, and AnEn holding out forecast years one year at a time counting backward in time. For example, the methods are trained solely on WY2016 and skill is determined on the testing dataset. Then the methods are trained on WY2015-WY2016, and so on until the entire dataset is utilized. Figure 8 shows forecast 48-h twCRPS(τ = 250 kg^{-1} m^{-1} s^{-1}) of AnEn, CNN, and NN by the number of years in the training dataset using the GEFS$_{nn}$ as a reference forecast.

The NN trains well with a single year of data and plateaus in skill quickly afterward. As one NN is trained for the entire domain, the NN is able to learn the forecast IVT error relationship from every point in the field domain, effectively multiplying the length of the training data by the number of points used (144, although these are not necessarily independent forecasts). The AnEn learns most quickly within the first 10 years but continues to learn as the lengths of training data are extended. This can be explained by the fact that, as more similar analogs are added with each year and the AnEn is unable to extrapolate forecast information but must rely on the past forecast record (except for the right tail of the distribution when the bias correction for rare events is applied). Although we truncate the figure at 23 yr, the AnEn continues to learn for the 34-yr period (though marginally; not shown).

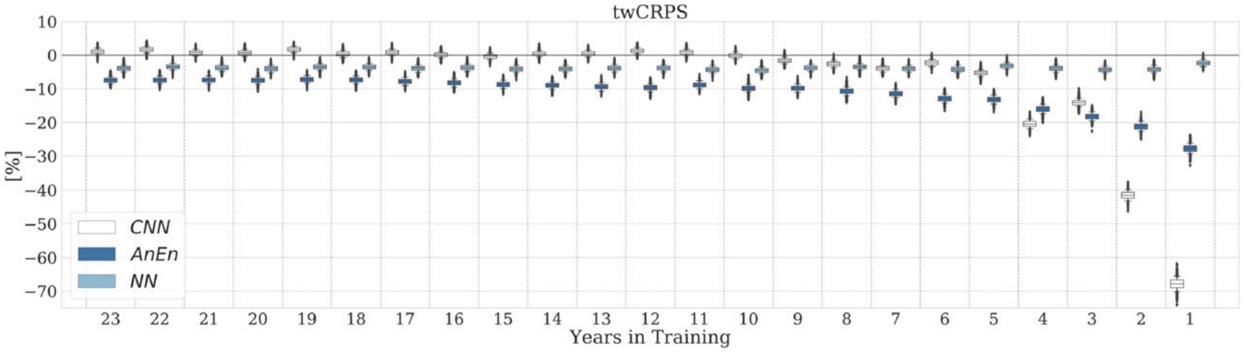

FIG. 8. Forecast 48-h threshold-weighted (250 kg m^{-1} s^{-1}) continuous ranked probability skill score against the number of water years included in the training dataset. GEFS$_{nn}$ is used as a reference forecast. Shown prediction systems include CNN (white), NN (light blue), and AnEn (dark blue). The error bars indicate the 95% bootstrap confidence intervals (where $n = 1000$).

The CNN is the worst performing forecast for the first 2 years and does not significantly outperform a simple NN until 9 yr of data are utilized. The associated cost of producing a hindcast, if a CNN is desired is thus high (although again, only a deterministic hindcast is required). The CNN appears to plateau after 11 years of training and only very marginal skill is added in the remaining 20 yr of training data.

4. Discussion and conclusions

Integrated water vapor transport is postprocessed to derive a forecast uncertainty quantification. There has been a recent surge of interest and method development of machine learning and deep learning for numerical weather prediction postprocessing (Baran and Baran 2021; Kirkwood et al. 2021; McGovern et al. 2017; Meech et al. 2021; Schulz and Lerch 2021; Vannitsem et al. 2021). The examined DL methods (NN and CNN) are flexible and easy to implement with modern DL toolboxes. The ML methods can compete well with dynamic model ensemble due to the severe underdispersion of the GEFS ensemble, and the DL ability to adjust the deterministic (mean) score to be competitive with the Global Ensemble Forecast System mean (Fig. 2). At lead times when the GEFS mean forecast skill significantly outperforms the deterministic dynamical forecast skill, the probabilistic methods have trouble competing (cf. Fig. 2 and Fig. 4). The GEFS$_{nn}$ does not adjust the GEFS ensemble mean prediction, but simply calibrates the ensemble spread. A well calibrated dynamically generated ensemble would be more difficult to outperform.

During the first 48 h the convolutional neural networks developed from a long running deterministic forecast system shows the best performance compared to each of the tested forecast systems, including a calibrated dynamic ensemble system. This represents a significant computational cost saving as a single deterministic model run is required as an input variable. At longer lead times (3–5 days) the CNN again is the best-performing forecast system for AR conditions (250 kg m^{-1} s^{-1} of IVT) but struggles more than the other systems with predicting the highest-impact events (350 and 500+ kg m^{-1} s^{-1} of IVT). From the perspective of the Brier skill score and threshold-weighted continuous rank probability score, the CNN is the best-performing forecast at nearly all lead times and for every threshold when compared with each other forecast system. However, the BSS can be broken down into components of reliability and resolution (Murphy 1973). Figures 6s and 7s in the online supplemental material show these components. For the longer lead times and high-impact events, it is clear that the CNN is favoring resolution at the expense of reliability. The online supplemental material also contains reliability diagrams (Bröcker and Smith 2007) to demonstrate this issue; while still reliable, the CNN is marginally less reliable than the other methods (Figs. 8s–13s in the online supplemental material). The CNN is clearly reliable for IVT events with magnitudes greater than 250 kg m^{-1} s^{-1} at all lead times but struggles slightly with reliability at the longer forecast leads for the highest-impact events. Therefore, if a user wants to know if AR conditions are probable, the CNN is the best West-WRF based forecast available among the NN-based methods and AnEn and is competitive or better than the dynamic ensemble methods. For AR events greater than 500 kg m^{-1} s^{-1} the AnEn or NN systems are more reliable, but much less resolved (Fig. 6s in the online supplemental material). Our results show that an NN trains extremely quickly and with a single year of hindcast data can create a very reliable probabilistic forecast.

Challenges remain for this DL postprocessing systems. The demonstrated DL methods are distinctly disadvantaged in that they fit unimodal parametric distributions, and variables that are not described well by a simple distribution will yield poor probabilistic forecast skill. In addition, these are highly parameterized models and significant computational time was required to find the prescribed model hyperparameters. The presented neural networks do not offer a seamless forecast system, with individual networks trained at each lead time. The FCN and NN embed location information in their forecasts, which was shown to effect forecast skill, this could easily be extended to embed temporal information (by embedding representations of forecast lead) that would unify the forecast system into a single neural network rather than training individual networks at every forecast lead (e.g., Ham et al. 2021). Also, the stability of these networks has not been proven under changing climate scenarios, and the relative nonstationarity of the training data could affect future long hindcast projects. Although a method for addressing this issue through fine-tuning was presented in this study, more work needs to be done to see if this offers a robust solution.

Work is under way to improve the neural network–based forecasts for high-impact events. The shape of Fig. 6j indicates a slight dry model bias, suggesting that simple postprocessed conditional bias correction may improve CNN model skill further. Success could be found in simply developing the loss function to act on twCRPS rather than CRPS alone. Additionally, focusing loss just on the coastal landfalling points, training with a greater percentage of high IVT events in the training set, adding AR/no AR discriminator networks to the CNN, or adding metrics that specifically target calibration all have offered positive results in initial testing.

This study uses neural networks and a CNN for distributional regression to quantify the prediction uncertainty from deterministic numerical weather forecast systems. The networks compete with or outperform state-of-the-art dynamic models, even when calibrated with the most modern postprocessing methods. The model's parameters are estimated by optimizing continuous ranked probability score, a standard metric in evaluating probabilistic weather forecasts, but one that is rarely used in ML communities. The models are flexible, fast, and can be readily trained with a few years of hindcast data.

Acknowledgments We acknowledge ample code sharing and collaboration and encourage the machine-learning/atmospheric science community to continue this open practice. We thank the West-WRF team led by Daniel Steinhoff for generating the deterministic forecast. This work was supported by USACE FIRO Grant W912HZ1520019 and CDWR AR Program Grant 4600013361. Sebastian Lerch gratefully acknowledges support by the Deutsche Forschungsgemeinschaft through SFB/TRR 165 "Waves to Weather," and by the Vector Stiftung through the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting." The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Data availability statement. Python code for reproducing the results and models is available online (https://github.com/WillyChap/ARML_Probabilistic). West-WRF simulations are archived at the Center for Western Weather and Water Extremes and on the National Center for Atmospheric Research servers are readily available upon request. GEFS data can be retrieved through the TIGGE archive (https://www.ecmwf.int/en/research/projects/tigge). MERRA2 data can also be retrieved online (https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/data_access/).

APPENDIX

Evaluation Metrics

We provide a summary of the methods used for forecast evaluation. In the following we will refer to a forecast by F, the random variable of the observation by Y, and a realization of Y by y, that is, observed IVT. General skill metrics are referred to as S. To avoid the known issues in evaluating stratified forecasts (e.g., Bellier et al. 2017; Lerch et al. 2017) careful consideration is taken to ensure that skill metrics remain proper when stratified, additionally all thresholding criteria is performed on the forecast ensemble mean (Hamill and Colucci 1997; Siegert et al. 2012).

a. Deterministic metrics

While the primary goal of this work is to determine the quality of the NN-derived uncertainty quantification. However, we additionally evaluate how the postprocessing methods affect the deterministic ensemble mean predictions. This is done largely because the skill of the deterministic prediction greatly affects the efficacy of the probabilistic methods. For example, CRPS reduces to mean absolute error for a deterministic forecast. For skill scores, we show root-mean-square error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(F_j - y_j)^2}$$

and Pearson correlation:

$$\mathrm{PC} = \frac{\sum_{j=1}^{N}(F_j - \overline{F})(y_j - \overline{Y})}{\sqrt{\sum_{j=1}^{N}(F_j - \overline{F})^2}\sqrt{\sum_{j=1}^{N}(y_j - \overline{Y})^2}}.$$

We aggregate forecasts over every location and demonstrate the methods' comparative skill score (SS). Each of the above scores S may be converted to SS by comparison with the same metric evaluated for a reference forecast S_{ref} through $\mathrm{SS} = 1 - (S/S_{\mathrm{ref}})$, $-\infty < \mathrm{SS} \leq 1$, where positive or negative values are shown to be more or less skillful, respectively, than the reference forecast.

b. Probabilistic predictions

1) PROPER SCORING RULES

Brier skill (BS; Brier 1950) is used to assess the prediction of binary events:

$$\mathrm{BS}_{\tau}(F_j, y_j) = [F_j(\tau) - I\{y_j \geq \tau\}]^2,$$

where τ is a prescribed threshold value, $I\{\}$ is the indicator (step) function that takes the value 1 if the jth verifying observation exceeds τ and is 0 otherwise, and $F_j(\tau)$ is the probability of event occurrence, which is forecast. The BS is particularly useful to check how skillful probabilistic IVT forecasts are in predicting different events across various established AR thresholds (e.g., Guan and Waliser 2015; Ralph et al. 2020a).

Continuous ranked probability score (CRPS; Matheson and Winkler 1976) is a measure of overall predictive performance that integrates the squared difference between cumulative probability distribution functions of the forecast F and observation Y:

$$\mathrm{CRPS}(F_j, y_j) = \int_{-\infty}^{\infty}[F_j(x) - I\{y_j \leq x\}]^2\,dx,$$

where $I\{\}$, again, is the indicator (step) function that takes the value of 1 if $x \geq y$ and 0 elsewhere. The integral in CRPS can be computed analytically for ensemble forecasts (Hersbach 2000) and a suite of continuous forecast distributions (Jordan et al. 2019). To remain proper, CRPS must be tailored to forecast extreme events; Gneiting and Ranjan (2011) defined the threshold-weighted CRPS (twCRPS):

$$\text{twCPRS}(F_j, y_j) = \int_{-\infty}^{\infty} w(x)[F_j(x) - I\{y_j \leq x\}]^2\, dx,$$

where w is a nonnegative weight function and when $w = 1$ twCRPS reduces to CRPS. To examine extreme events (right tail of distribution) we can set $w(x) = I\{x \geq \tau\}$, where again τ is a prescribed threshold value. The twCRPS integral can be computed numerically, and we leverage this method for our verification. Again, for this metric, we aggregate the forecasts over every station location and demonstrate the methods' comparative skill.

To compare the relative performance of each scheme we evaluate the Brier skill score,

$$[\text{BSS} = 1 - (\overline{\text{BS}}/\overline{\text{BS}_{\text{ref}}})],$$

and continuous ranked probability skill score,

$$[\text{CRPSS} = 1 - (\overline{\text{CRPS}}/\overline{\text{CRPS}_{\text{ref}}})],$$

which are shown in Fig. 4 and supplemental Fig. 2s, respectively. Positive values indicate a skill improvement.

The integral in the CRPS equation can be computed analytically for ensembles, and for many continuous forecast distributions (see, Jordan et al. 2019). In this work we use the exact Gaussian CRPS solution to train our neural networks. Although this rarely occurs, because the focus of this study is on IVT events that are above the $250\,\text{kg}\,\text{m}^{-1}\,\text{s}^{-1}$ threshold, we truncate predictions at 0 because negative values of integrated vapor transport are nonphysical. The exact solution of the Gaussian CRPS with mean value μ, standard deviation σ, and observation y is

$$\text{CRPS}(F_{\mu\sigma}, y) = \sigma\left\{\frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right] + 2\varphi\frac{y-\mu}{\sigma} - \frac{1}{\sqrt{\pi}}\right\},$$

where Φ and φ denote the CDF and PDF of a standard Gaussian distribution, respectively (Gneiting et al. 2005).

2) Diebold–Mariano test for statistical test of equal predictive performance

We utilize a two-sided Diebold–Mariano test to assess whether differences in forecast performances are statistically significant (Diebold and Mariano 2002). Consider two forecasts F_1 and F_2, with respective mean scores

$$\overline{S}(F_i) = \frac{1}{n}\sum_{j=1}^{n} S(F_j^i, y_j)$$

for i =1, 2 over a test $j = 1, \ldots, n$, where the forecast F_j^i was issued k time steps before the observation y_j. The DM test assumes that, under standard regularity conditions and the forecast cases being independent,

$$t_n = \sqrt{n}\frac{\overline{S}(F^1) - \overline{S}(F^2)}{\hat{\sigma}_n},$$

where

$$\hat{\sigma}_n = \frac{1}{n}\sum_{j=1}^{n}[S(F_j^1, y_j^1) - S(F_j^2, y_j^2)]^2$$

follows the standard Gaussian distribution under the null hypotheses of equal predictive performance of two forecast sources. The null hypothesis is rejected for large values of $|t_n|$, by obtaining the corresponding p value, where forecast F^1 outperforms F^2 if t_n is negative and underperforms F^2 if t_n is positive.

To account for test multiplicity when comparing methods across multiple forecast lead times, we follow Wilks (2016), and use the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995) to control for the false discovery rate. Given the ordered lead-time p values $p_1, \ldots, p_M$ of M hypothesis tests, a new threshold p value (p^*) is determined via $p^* = p_{(i^*)}$, where

$$i^* = \min\left[i = 1, \ldots, M : p_{(i)} \leq \alpha_{\text{FDR}}\Delta\frac{i}{M}\right],$$

and we choose $\alpha_{\text{FDR}} = 0.05$. We then reject the null hypothesis if the test p value $< p^*$. In Fig. 3, M is set to the number of tested lead times while comparing the twCRPS ($M = 22$).

3) Stratified rank histogram

Rank histograms are diagnostic tools that assess the calibration of a forecast ensemble (Hamill 2001). An ensemble is statistically consistent when ensemble members cannot be distinguished from observations. Therefore, an observation ranked among the corresponding ordered ensembles is equally likely to assume any position. If a significant amount of forecasts are assessed in this manner, a histogram of the observation ranks should show a perfectly uniform distribution [rank probability of $1/(n+1)$, where n = number of ensemble members]. Bellier et al. (2017) proposed the use and demonstrated the consistency of forecast-based stratified rank histograms, which show calibration between given forecast thresholds, and easily enables one to assess the contribution of each stratum to the overall rank histogram uniformity.

4) Spread skill diagram

The ability of a forecast systems to quantify uncertainty is examined with binned spread–skill plots, which compare ensemble spread (i.e., the standard deviation of the ensemble members) to RMSE of the ensemble mean over small class intervals of model spread, rather than considering the overall average spread as in the dispersion diagram (e.g., van den Dool 1989; Wang and Bishop 2003). The spread of a forecast perfectly describes the uncertainty of the system if the actual forecast error equals its spread (Leutbecher and Palmer 2008). The ability to quantify the prediction uncertainty thus requires the two metrics to match at all binned values of ensemble spread, resulting in a line that falls upon the 1:1 line.

5) PERMUTATION IMPORTANCE

Permutation importance has been explored for describing variable importance to DL models in multiple Earth system studies (Brenowitz et al. 2020; McGovern et al. 2019; Rasp and Lerch 2018). Permutation importance is here defined as

$$\Delta(x^*, x)_{F^1,F^2} = \overline{\mathrm{twCRPS}}[F_j^1(x), y_j(x)] - \overline{\mathrm{twCRPS}}[F_j^2(x^*), y_j(x)],$$

where x^* represents input variable space with a singly randomly permutated input variable selected from the set of input features The input features have length equal to the total number of forecasts j. Permutation importance PI is then set in reference against the nonpermuted forecast twCRPS:

$$\mathrm{PI} = \left\{\frac{\Delta(x^*, x)_{F^1,F^2}}{\overline{\mathrm{twCRPS}}[F_j^1(x), y_j(x)]}\right\}.$$

The random permutation process is repeated for each input variable shown in Fig. 7.

REFERENCES

Abadi, M., and Coauthors, 2016: Tensorflow: A system for large-scale machine learning. *Proc. 12th USENIX Symp. on Operating Systems Design and Implementation (OSDI'16)*, Savannah, GA, USENIX, 21 pp., https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Alessandrini, S., 2019: Improving the analog ensemble wind speed forecasts for rare events. *Mon. Wea. Rev.*, **147**, 2677–2692, https://doi.org/10.1175/MWR-D-19-0006.1.

——, L. Delle Monache, S. Sperati, and G. Cervone, 2015: An analog ensemble for short-term probabilistic solar power forecast. *Appl. Energy*, **157**, 95–110, https://doi.org/10.1016/j.apenergy.2015.08.011.

Baran, S., and S. Lerch, 2015: Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quart. J. Roy. Meteor. Soc.*, **141**, 2289–2299, https://doi.org/10.1002/qj.2521.

——, and Á. Baran, 2021: Calibration of wind speed ensemble forecasts for power generation. Preprint, 15 pp., https://arxiv.org/abs/2104.14910.

Bellier, J., I. Zin, and G. Bontron, 2017: Sample stratification in verification of ensemble forecasts of continuous scalar variables: Potential benefits and pitfalls. *Mon. Wea. Rev.*, **145**, 3529–3544, https://doi.org/10.1175/MWR-D-16-0487.1.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **57B**, 289–300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

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., T. Beucler, M. Pritchard, and C. S. Bretherton, 2020: Interpreting and stabilizing machine-learning parametrizations of convection. *J. Atmos. Sci.*, **77**, 4357–4375, https://doi.org/10.1175/JAS-D-20-0082.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

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.

Camporeale, E., and A. Carè, 2021: ACCRUE: Accurate and reliable uncertainty estimate in deterministic models. *Int. J. Uncertain. Quantif.*, **11**, 81–94, https://doi.org/10.1615/Int.J.UncertaintyQuantification.2021034623.

Chang, E. K. M., and D. B. Yu, 1999: Characteristics of wave packets in the upper troposphere. Part I: Northern Hemisphere winter. *J. Atmos. Sci.*, **56**, 1708–1728, https://doi.org/10.1175/1520-0469(1999)056<1708:COWPIT>2.0.CO;2.

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

Chollet, F., and Coauthors, 2018: Keras: The Python deep learning library. Astrophysics Source Code Library, ASCL-1806, accessed 10 April 2019, https://ascl.net/1806.022.

Cobb, A., A. Michaelis, S. Iacobellis, F. M. Ralph, and L. Delle Monache, 2021: Atmospheric river sectors: Definition and characteristics observed using dropsondes from 2014–20 CalWater and AR Recon. *Mon. Wea. Rev.*, **149**, 623–644, https://doi.org/10.1175/MWR-D-20-0177.1.

Cordeira, J. M., and F. M. Ralph, 2021: A summary of GFS ensemble integrated water vapor transport forecasts and skill along the U.S. West Coast during water years 2017–20. *Wea. Forecasting*, **36**, 361–377, https://doi.org/10.1175/WAF-D-20-0121.1.

——, ——, A. Martin, N. Gaggini, J. R. Spackman, P. J. Neiman, J. J. Rutz, and R. Pierce, 2017: Forecasting atmospheric rivers during CalWater 2015. *Bull. Amer. Meteor. Soc.*, **98**, 449–459, https://doi.org/10.1175/BAMS-D-15-00245.1.

Corringham, T. W., F. M. Ralph, A. Gershunov, D. R. Cayan, and C. A. Talbot, 2019: Atmospheric rivers drive flood damages in the western United States. *Sci. Adv.*, **5**, eaax4631, https://doi.org/10.1126/sciadv.aax4631.

Deflorio, M. J., D. E. Waliser, B. Guan, D. A. Lavers, F. Martin Ralph, and F. Vitart, 2018: Global assessment of atmospheric river prediction skill. *J. Hydrometeor.*, **19**, 409–426, https://doi.org/10.1175/JHM-D-17-0135.1.

DeHaan, L. L., A. C. Martin, R. R. Weihs, L. Delle Monache, and F. M. Ralph, 2021: Object-based verification of atmospheric river predictions in the northeast Pacific. *Wea. Forecasting*, **36**, 1575–1587, https://doi.org/10.1175/WAF-D-20-0236.1.

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. Stull, 2006: Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *J. Geophys. Res.*, **111**, D05308, https://doi.org/10.1029/2005JD006311.

——, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

Diebold, F. X., and R. S. Mariano, 2002: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **20**, 134–144.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759, https://doi.org/10.3402/tellusa.v21i6.10143.

Fish, M. A., A. M. Wilson, and F. M. Ralph, 2019: Atmospheric river families: Definition and associated synoptic conditions. *J. Hydrometeor.*, **20**, 2091–2108, https://doi.org/10.1175/JHM-D-18-0217.1.

Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *J. Climate*, **30**, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1.

Ghazvinian, M., Y. Zhang, D.-J. Seo, M. He, and N. Fernando, 2021: A novel hybrid artificial neural network-parametric scheme for postprocessing medium-range precipitation forecasts. *Adv. Water Resour.*, **151**, 103907, https://doi.org/10.1016/j.advwatres.2021.103907.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

——, and R. Ranjan, 2011: Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econ. Stat.*, **29**, 411–422.

——, A. E. Raftery, A. H. Westveld III, 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.

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, https://doi.org/10.1175/WAF-D-17-0144.1.

Guan, B., and D. E. Waliser, 2015: Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *J. Geophys. Res. Atmos.*, **120**, 12 514–12 535, https://doi.org/10.1002/2015JD024257.

Guo, C., and F. Berkhahn, 2016: Entity embeddings of categorical variables. Preprint, 9 pp., https://arxiv.org/abs/1604.06737.

Hacker, J. P., and Coauthors, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus*, **63A**, 625–641, https://doi.org/10.1111/j.1600-0870.2010.00497.x.

Ham, Y.-G., J.-H. Kim, E.-S. Kim, and K.-W. On, 2021: Unified deep learning model for El Niño/Southern Oscillation forecasts by incorporating seasonality in climate data. *Sci. Bull.*, **66**, 1358–1366, https://doi.org/10.1016/j.scib.2021.03.009.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

——, 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.

Haupt, S. E., W. Chapman, S. V. Adams, C. Kirkwood, J. S. Hosking, N. H. Robinson, S. Lerch, and A. C. Subramanian, 2021: Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. *Philos. Trans. Roy. Soc. London*, **A379**, 20200091, https://doi.org/10.1098/rsta.2020.0091.

Hecht, C. W., and J. M. Cordeira, 2017: Characterizing the influence of atmospheric river orientation and intensity on precipitation distributions over North Coastal California. *Geophys. Res. Lett.*, **44**, 9048–9058, https://doi.org/10.1002/2017GL074179.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden, 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205, https://doi.org/10.1002/2014GL062472.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Jordan, A., F. Krüger, and S. Lerch, 2019: Evaluating probabilistic forecasts with scoring rules. *J. Stat. Software*, **90**, 1–37, https://doi.org/10.18637/jss.v090.i12.

Junk, C., L. Delle Monache, S. Alessandrini, G. Cervone, and L. Von Bremen, 2015: Predictor-weighting strategies for probabilistic wind power forecasting with an analog ensemble. *Meteor. Z.*, **24**, 361–379, https://doi.org/10.1127/metz/2015/0659.

Kingma, P., B. Diederik, and J. Lei, 2014: Adam: A method for stochastic optimization. Preprint, 15 pp., https://arxiv.org/abs/1412.6980.

Kirkwood, C., T. Economou, H. Odbert, and N. Pugeault, 2021: A framework for probabilistic weather forecast post-processing across models and lead times using machine learning. *Philos. Trans. Roy. Soc.*, **A379**, 20200099, https://doi.org/10.1098/rsta.2020.0099.

Kirtman, B. P., and Coauthors, 2014: The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

Krogh, A., and J. A. Hertz, 1992: A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems 4 (NIPS'91)*, J. Moody, S. Hanson, and R. P. Lippmann, Eds., MIT Press, 950–957.

Lamjiri, M. A., M. D. Dettinger, F. M. Ralph, and B. Guan, 2017: Hourly storm characteristics along the U.S. West Coast: Role of atmospheric rivers in extreme precipitation. *Geophys. Res. Lett.*, **44**, 7020–7028, https://doi.org/10.1002/2017GL074193.

Lerch, S., and T. L. Thorarinsdottir, 2013: Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus*, **65A**, 21206, https://doi.org/10.3402/tellusa.v65i0.21206.

——, ——, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma. *Extreme Events Forecast Eval.*, **32**, 106–127.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Li, Y., J. Lang, L. Ji, J. Zhong, Z. Wang, Y. Guo, and S. He, 2020: Weather forecasting using ensemble of spatial-temporal attention network and multi-layer perceptron. *Asia-Pac. J. Atmos. Sci.*, **57**, 533–546, https://doi.org/10.1007/s13143-020-00212-3.

Long, J., E. Shelhamer, and T. Darrell, 2015: Fully convolutional networks for semantic segmentation. *2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, IEEE, https://doi.org/10.1109/CVPR.2015.7298965.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Martin, A., F. M. Ralph, R. Demirdjian, L. DeHaan, R. Weihs, J. Helly, D. Reynolds, and S. Iacobellis, 2018: Evaluation of atmospheric river predictions by the WRF Model using aircraft and regional mesonet observations of orographic precipitation and its forcing. *J. Hydrometeor.*, **19**, 1097–1113, https://doi.org/10.1175/JHM-D-17-0098.1.

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.

McCarty, W., L. Coy, R. Gelaro, A. Huang, D. Merkova, E. B. Smith, M. Sienkiewicz, and K. Wargan, 2016: MERRA-2 input observations: Summary and assessment. NASA Tech. Memo. NASA/TM-2016-104606/Vol. 46, 51 pp.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

——, R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Meech, S., S. Alessandrini, W. Chapman, and L. Delle Monache, 2021: Post-processing rainfall in a high-resolution simulation of the 1994 Piedmont flood. *Bull. Atmos. Sci. Technol.*, **1**, 373–385, https://doi.org/10.1007/s42865-020-00028-z.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor. Climatol.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Nair, V., and G. E. Hinton, 2010: Rectified linear units improve restricted Boltzmann machines. *Proc. 27th Int. Conf. on Machine Learning*, Haifa, Israel, ICML, 807–814.

Nardi, K. M., E. A. Barnes, and F. M. Ralph, 2018: Assessment of numerical weather prediction model reforecasts of the occurrence, intensity, and location of atmospheric rivers along the West Coast of North America. *Mon. Wea. Rev.*, **146**, 3343–3362, https://doi.org/10.1175/MWR-D-18-0060.1.

Nayak, M. A., G. Villarini, and D. A. Lavers, 2014: On the skill of numerical weather prediction models to forecast atmospheric rivers over the central United States. *Geophys. Res. Lett.*, **41**, 4354–4362, https://doi.org/10.1002/2014GL060299.

Nielsen, M., 2015: *Neural Networks and Deep Learning*. Determination Press, http://neuralnetworksanddeeplearning.com/.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Ralph, F. M., J. J. Rutz, J. M. Cordeira, M. Dettinger, M. Anderson, D. Reynolds, L. J. Schick, and C. Smallcomb, 2018: A scale to characterize the strength and impacts of atmospheric rivers. *Bull. Amer. Meteor. Soc.*, **100**, 269–289, https://doi.org/10.1175/BAMS-D-18-0023.1.

——, and Coauthors, 2020a: West Coast forecast challenges and development of atmospheric river reconnaissance. *Bull. Amer. Meteor. Soc.*, **101**, E1357–E1377, https://doi.org/10.1175/BAMS-D-19-0183.1.

——, M. D. Dettinger, L. J. Schick, and M. L. Anderson, 2020b: Introduction to atmospheric rivers. *Atmospheric Rivers*, F. M. Ralph et al., Eds., Springer, 1–13.

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.

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

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, N. Navab et al., Eds., Springer, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

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.

——, 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.

Schulz, B., and S. Lerch, 2021: Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. Preprint, 44 pp., https://arxiv.org/abs/2106.09512.

Schulzweida, U., L. Kornblueh, and R. Quast, 2006: CDO user guide: Version 1.9.5, Vol. 1. Climate Data Operators Doc., 215 pp.

Siegert, S., J. Bröcker, and H. Kantz, 2012: Rank histograms of stratified Monte Carlo ensembles. *Mon. Wea. Rev.*, **140**, 1558–1571, https://doi.org/10.1175/MWR-D-11-00302.1.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014: Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Stone, R. E., C. A. Reynolds, J. D. Doyle, R. H. Langland, N. L. Baker, D. A. Lavers, and F. M. Ralph, 2020: Atmospheric river reconnaissance observation impact in the Navy Global Forecast System. *Mon. Wea. Rev.*, **148**, 763–782, https://doi.org/10.1175/MWR-D-19-0101.1.

Theocharides, S., G. Makrides, A. Livera, M. Theristis, P. Kaimakis, and G. E. Georghiou, 2020: Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. *Appl. Energy*, **268**, 115023, https://doi.org/10.1016/j.apenergy.2020.115023.

Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173A**, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x.

——, and M. S. Johnson, 2012: Probabilistic wind gust forecasting using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.*, **140**, 889–897, https://doi.org/10.1175/MWR-D-11-00075.1.

Torrey, L., and J. Shavlik, 2010: Transfer learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. E. Soria Olivas et al., Eds., IGI Global, 242–264.

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

van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.*, **117**, 2230–2247, https://doi.org/10.1175/1520-0493(1989)117<2230:ANLAWF>2.0.CO;2.

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

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, https://doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2.

Wick, G. A., P. J. Neiman, F. M. Ralph, and T. M. Hamill, 2013: Evaluation of forecasts of the water vapor signature of atmospheric rivers in operational numerical weather prediction models. *Wea. Forecasting*, **28**, 1337–1352, https://doi.org/10.1175/WAF-D-13-00025.1.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, https://doi.org/10.1002/met.134.

——, 2010: Sampling distributions of the Brier score and Brier skill score under serial dependence. *Quart. J. Roy. Meteor. Soc.*, **136**, 2109–2118, https://doi.org/10.1002/qj.709.

——, 2016: “The stippling shows statistically significant grid points”: How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, **97**, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1.

Wilson, A. M., and Coauthors, 2020: Training the next generation of researchers in the science and application of atmospheric rivers. *Bull. Amer. Meteor. Soc.*, **101**, E738–E743, https://doi.org/10.1175/BAMS-D-19-0311.1.

Wirth, V., M. Riemer, E. K. M. Chang, and O. Martius, 2018: Rossby wave packets on the midlatitude waveguide—A review. *Mon. Wea. Rev.*, **146**, 1965–2001, https://doi.org/10.1175/MWR-D-16-0483.1.

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

Zhang, A., Z. C. Lipton, M. Li, and A. J. Smola, 2021: *Dive into Deep Learning*. Release 0.17.0, self published, 823 pp., https://arxiv.org/abs/2106.11342.