Using Artificial Neural Networks to Improve CFS Week-3–4 Precipitation and 2-m Air Temperature Forecasts

YUN FAN,[a] VLADIMIR KRASNOPOLSKY,[b] HUUG VAN DEN DOOL,[a] CHUNG-YU WU,[a] AND JON GOTTSCHALCK[a]

[a] *Climate Prediction Center, NOAA/Center for Weather and Climate Prediction, College Park, Maryland*
[b] *Environmental Modeling Center, NOAA/Center for Weather and Climate Prediction, College Park, Maryland*

(Manuscript received 24 January 2020, in final form 16 November 2020, accepted 12 January 2021)

ABSTRACT: Forecast skill from dynamical forecast models decreases quickly with projection time due to various errors. Therefore, postprocessing methods, from simple bias correction methods to more complicated multiple linear regression–based model output statistics, are used to improve raw model forecasts. Usually, these methods show clear forecast improvement over the raw model forecasts, especially for short-range weather forecasts. However, linear approaches have limitations because the relationship between predictands and predictors may be nonlinear. This is even truer for extended range forecasts, such as week-3–4 forecasts. In this study, neural network techniques are used to seek or model the relationships between a set of predictors and predictands, and eventually to improve week-3–4 precipitation and 2-m temperature forecasts made by the NOAA/NCEP Climate Forecast System. Benefitting from advances in machine learning techniques in recent years, more flexible and capable machine learning algorithms and availability of big datasets enable us not only to explore nonlinear features or relationships within a given large dataset, but also to extract more sophisticated pattern relationships and covariabilities hidden within the multidimensional predictors and predictands. Then these more sophisticated relationships and high-level statistical information are used to correct the model week-3–4 precipitation and 2-m temperature forecasts. The results show that to some extent neural network techniques can significantly improve the week-3–4 forecast accuracy and greatly increase the efficiency over the traditional multiple linear regression methods.

KEYWORDS: Climate prediction; Numerical weather prediction/forecasting; Statistical forecasting; Intraseasonal variability; Machine learning; Neural networks

1. Introduction

The public demand for Subseasonal to Seasonal Prediction project (S2S) forecasts has been steadily increasing in recent years, primarily driven by certain industries, such as water management, agriculture, financial markets, energy, transportation, commerce, tourism, and insurance, etc., to prepare for and reduce risk from damaging meteorological events well in advance. In 2016, the National Oceanic and Atmospheric Administration (NOAA) initiated efforts to improve its capability for weeks 3 and 4 (i.e., 15–28 days ahead) extended range forecasts. Covering the week-3–4 extended-range lead time will enable NOAA to provide seamless S2S forecasts to the public for protecting life and property.

Numerical forecasts on the week-3–4 time scale are relatively new and are some of the most challenging and difficult to make. Past forecast efforts have focused on the short-term weather forecasts out to at most 7–10 days, operational short-term climate outlooks from 6 to 10 days and 8–14 days, and months-long integrations out to several seasons. There is a clear forecast gap around week 3 and 4. This is because current numerical weather models perform well up to about seven days in advance, and climate outlooks become more reliable as the time horizon extends from months to seasons. Subseasonal (e.g., week 3–4) forecasts are a middle ground, where the memory of the initial conditions that impact short-term weather is diminished after 7–10 days, while the impact of monthly and seasonal factors such as the state of El Niño, soil moisture, snow, and sea ice, along with others, is not yet well established for subseasonal forecasts. Sharma et al. (2017) and Pan et al. (2019) studied precipitation forecasts in the eastern United States and the West Coast from short to extended range and found the current state-of-the-art models provide little useful forecast skill beyond week 1–2. Numerical forecast of the atmospheric rivers, atmospheric blocking, and tropical cyclones showed similar results (Wick et al. 2013; Nayak et al. 2014; Nardi et al. 2018; Zhong et al. 2018). Modulation of some low-frequency modes, such as the Madden–Julian oscillation (MJO), quasi-biennial oscillation (QBO), and sea surface temperature (SST) suggests potential predictability for subseasonal forecasts (Johnson et al. 2014; DelSole et al. 2017; Vigaud et al. 2018; Baggett et al. 2018; Mundhenk et al. 2018; Wang and Robertson 2018; Jenney et al. 2019).

Corresponding author: Yun Fan, Yun.Fan@noaa.gov

DOI: 10.1175/WAF-D-20-0014.1

The Subseasonal Experiment (SubX), a research-to-operations project launched recently, provides a comprehensive research infrastructure for developing better S2S forecasts (Pegion et al. 2019).

Numerical weather and climate forecast models have been improving continuously during the last several decades (Warner 2011; Bauer et al. 2015). However, forecasts from direct dynamical model outputs still suffer from large forecast errors with lead time increasing due to the deficiency of model physics, errors in initial and boundary conditions, and other reasons. Therefore, various dynamical model postprocessing strategies are developed to remove forecast biases and errors, and to nudge model predictions toward observations, before forecasts are issued to the public.

Linear statistical postprocessing methods show some success in improving direct model prediction skill. One of those techniques is the model output statistics (MOS), which relates observed weather elements (predictands) to appropriate model forecast variables (predictors) via a statistical approach [e.g., multiple linear regression (MLR)]. MOS provides a tool for forecasters to objectively interpret numerical model output, quantifying uncertainties, remove biases, derive forecast variables not directly available from numerical forecast models, and provide improved weather forecast guidance. It is used routinely in different operational centers worldwide (Glahn and Lowry 1972; Klein and Glahn 1974; Wilson and Vallee 2002, 2003; Glahn et al. 2009; Gneiting 2014). However, the linear approach has some limitations, such as the huge number (millions) of MOS forecast equations trained pointwise for different variables over different sites, projection times, and weather regimes. Moreover, with increasing lead time, the relationship between predictands and predictors may be more nonlinear. This is even truer for the extended range forecasts, such as the week-3–4 forecasts.

In recent years, the great advances in machine learning (ML) in different fields have received much attention, due to the invention of more flexible and sophisticated ML methodologies and also the availability of larger datasets (i.e., "big data") for exploring challenging issues (Schmidhuber 2015; LeCun et al. 2015). ML technology has been developed to work with big data across a variety of disciplines and impacts almost every aspect of modern society from automation, classification, analysis, to detection. Modern ML (e.g., deep learning) techniques allow computational models to learn representations of large datasets with multiple levels of abstraction. Using a training algorithm, ML methods allow for identifying and modeling of more complicated relationships between variables that are not limited by linearity with a given optimization procedures.

Different ML techniques have been used to extract useful information and insights, and find the "known unknowns" from big data to solve the more challenging issues and make more accurate weather and climate forecasts. McGovern et al. (2017) showed that using artificial intelligence (AI) (e.g., decision-tree-based methods) can improve high-impact weather forecasting. Totz et al. (2017) used a cluster analysis for winter season precipitation anomaly outlooks, which outperforms both dynamical forecast models and a canonical correlation analysis based method. Cohen et al. (2019) showed ML techniques are far more powerful at mining data and recognizing patterns, and may be appropriate for subseasonal to seasonal (S2S) predictions. Neural networks (NN) are one of the most useful methods used in ML technologies. Modern NNs are able to learn high-level representations of a broad class of patterns from large datasets and are very good at discovering intricate structures hidden within high-dimensional big data. Krasnopolsky and Lin (2012; Krasnopolsky 2013) showed that neural networks can be used to improve daily (lead time of 24 h) precipitation forecast and in many other applications in the Earth system. Liu et al. (2016) used deep convolution neural networks to detect extreme weather (e.g., tropical cyclones and atmospheric rivers) in climate data. Rasp and Lerch (2018) demonstrated that neural network approaches can significantly outperform traditional state-of-the-art postprocessing methods for 2-m temperature forecasts at lead time of 48 h while being computationally affordable. NN techniques have a number of advantages. Their flexible and user-friendly algorithms can be used to simulate arbitrary nonlinear relationships. NN techniques can also more easily handle a large number of predictors/predictands and may help to discover complex nonlinear interconnections between predictors and predictands from large datasets.

So far, the daily week-3–4 forecast skill from direct dynamical forecast models is much lower than that of the short-range forecasts, such as 1–7 days and the week-1–2 forecasts. In this paper some NN architectures that are more beneficial for using model-derived fields are proposed. These NNs will be used to explore and evaluate their capability to improve the week-3–4 precipitation and 2-m air temperature forecasts. The rest of this paper is organized as follows: the dataset used for the NN training/testing and detailed NN methodology used in this study is highlighted in section 2. The NN check, optimal hidden neurons, data representation, and analysis of the week-3–4 model forecast errors are described in section 3. The NN forecast analysis and evaluation are presented in section 4, and conclusions and discussions are given in section 5.

2. Data and methodology

a. Data for NN training and validation

The datasets used for the NN training and testing consist of daily paired predictor and predictand variables. The dataset for the predictors used here includes the daily bias-corrected week-3–4 lead time forecast for total precipitation (P), mean 2-m air temperature (T2m), and 500-hPa height (Z500), and some others, which are obtained from the NOAA Climate Forecast System (CFS) (Saha et al. 2006, 2014) for the period 1 January 1999–31 December 2018. Since bias correction (by removing differences between model climatology and observed climatology) is one of the easiest and most effective ways to improve the raw model forecasts, one of the goals of this study is to see if the method introduced here can further improve the bias-corrected CFSv2 forecasts. The data domain used here covers the conterminous United States (CONUS) only. The data has been regridded to $1° \times 1°$ spatial resolution, nine selected vertical levels (pressure: 1000, 850, 700, 500, 300, 200, 100, 50, and 10), and is on a daily temporal resolution initialized at four different times (0000, 0600, 1200, and

1800 UTC) per day. Other predictors are also used, including daily P, T2m, and Z500 climatologies, latitudes, longitudes, elevations, station ID, and $\sin(\tau)$ and $\cos(\tau)$ where $\tau = (2\pi/365)t$ and t is the day of the year, all on the same spatial–temporal resolutions. These auxiliary predictors are also commonly used in the MOS and other NN systems.

The datasets used for corresponding target variables (predictands) include the daily observed P from the gauge-based daily CPC Unified Precipitation Analysis, the observed daily T2m from the Global Telecommunications System (GTS) based daily 2-m temperature analysis (Chen et al. 2008; W. Shei 2009, personal communication; Fan and van den Dool 2008). Both daily observed P and T2m are converted to two-weekly total and two-weekly means, and regridded to the same spatial–temporal resolutions as the above predictors.

The above 20 years of daily paired (predictors and predictands) datasets have 7305 daily records and can be split into two parts, the first part (about 6575 daily records, from 1 January 1999 to 31 December 2016) was used for training and the remaining part (about 730 daily records, from 1 January 2017 to 31 December 2018) was used for validation (independent forecast test). Three different k-fold cross-validation tests are also performed to verify the NN generalization in different periods.

b. Methodology

1) Formulation of the problem

Usually, the statistical post processing of model output is based on the reasonable assumption that there is a relationship between target variables/predictands (e.g., observed weather and climate elements) and input variables/predictors (e.g., the corresponding forecast variables of numerical prediction model). In a generic symbolic way, this relationship can be represented as

$$\mathbf{Z} = M(\mathbf{X});\ \mathbf{X} \in \Re^n,\ \ \mathbf{Z} \in \Re^m, \tag{1}$$

where $\mathbf{X}$ is an input vector composed of model forecast variables or predictors, $\mathbf{Z}$ is an output vector composed of observed meteorological elements or predictands, n is the dimensionality of the vector $\mathbf{X}$ (or input space), and m is the dimensionality of the vector $\mathbf{Z}$ (or output space). The term M denotes the mapping (relationship between the two vectors) that relates vectors $\mathbf{X}$ and $\mathbf{Z}$. In a particular case when a single predictand is considered, the mapping Eq. (1) turns into a single valued function of multiple variables. This function/mapping is expected to be a complex nonlinear function.

Since both model forecast variables (predictors) and observations (predictands) contain errors in their data representations due to model deficiency, noise, uncertainty in initial and boundary conditions, and limited spatial and temporal resolutions, etc., a statistical approximation of the mapping Eq. (1) can be written as

$$\mathbf{Y} = M_s(\mathbf{X}). \tag{2}$$

Here the vector $\mathbf{Y}$ can be considered as an estimated predictand vector based on model variables $\mathbf{X}$, while M_s is a statistical approximation for the mapping M in Eq. (1). In the majority of modern MOS systems a single valued and pointwise MLR is used as the method of statistical approximation. In this case, the mapping Eq. (2) can be represented by a system of m independent linear regression equations:

$$y_q = a_{q0} + \sum_{j=1}^{n} a_{qj} \cdot x_j;\quad q = 1, \ldots, m. \tag{3}$$

The coefficients a_{qj} of various equations of the system (3) are different and usually calculated for each equation (for each corrected model variable y_q) individually and independently.

The linear regression approach Eq. (3) has three major disadvantages. First, the essentially nonlinear relationship/mapping Eq. (2) is approximated by linear dependencies in Eq. (3), which loses nonlinear components of the relationship between input vector and output vector. Second, the linear approach, as designed in most MOS procedures, does not consider the covariability between output variables (e.g., the observed 2-week total P and mean T2m here), whereas the nonlinear relationship/mapping Eq. (2) can take into account the relationships between different observed weather elements (components of vector $\mathbf{Y}$). Third, the approximation Eq. (3) splits the vector $\mathbf{Y}$ (e.g., P, T2m, wind, and other variables) into single components y_q that are usually treated not only individually and independently, but also location by location (i.e., point by point), thus losing the spatial dependency (or pattern relationship). Therefore, the approach Eq. (3), by definition, does not completely use relationships and correlations (or consistency constraints) offered by the observed data. It also does not use the pattern relationships (or space dependency) offered by the big datasets.

In the following sections, it will be shown that the NN approach allows users not only to address the aforementioned important problems and to improve the approximation, but also greatly reduces the number of approximation equations which improves training efficiency at the same time.

2) NN emulation for the linear mapping

The NN techniques are generic, accurate, flexible, and convenient mathematical/statistical models that can enable users to emulate/approximate different complicated nonlinear input/output relationships, by using statistical ML algorithms (Krasnopolsky 2013). NN can be applied to any problem that can be formulated as a mapping (input vector vs output vector relationship). The simplest NN approximations use a family of analytical functions such as

$$y_q = \mathrm{NN}(\mathbf{X}, \mathbf{a}, \mathbf{b}) = a_{q0} + \sum_{j=1}^{k} a_{qj} \cdot f_j;\quad q = 1, 2, \ldots, m, \tag{4}$$

where

$$f_j = F\left(b_{j0} + \sum_{i=1}^{n} b_{ji} \cdot x_i\right) = \tanh\left(b_{j0} + \sum_{i=1}^{n} b_{ji} \cdot x_i\right). \tag{5}$$

Here, x_i and y_q are components of the input and output vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively; vector $\mathbf{a}$ and vector $\mathbf{b}$ are the NN weights; n and m are the number of inputs and outputs, respectively; and k is the number of nonlinear basis activation

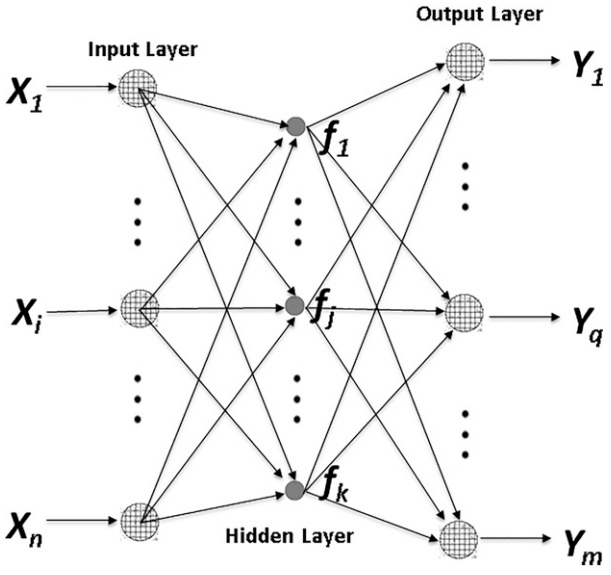

FIG. 1. The simplest NN with one hidden layer and linear neurons in the output layer. The hidden layer derives nonlinear transformations of the inputs and then linear combinations of these nonlinear transformations are used to model the outputs.

functions f_j (or hidden neurons). The hyperbolic tangent is used as an activation function (Hornik 1991, 1993). Other activation functions can be used depending on the problem at hand (Liu et al. 2016; McGovern et al. 2017; Rasp and Lerch 2018). Equation (4) is a mapping that can approximate any continuous or almost continuous (with finite discontinuities) mapping (Krasnopolsky 2013). A pictographic representation of the entire NN is shown in Fig. 1 and the connections (arrows) correspond to the NN weights. The NN complexity can be quickly increased by adding variables in the input layer and/or output layer, and neurons in the hidden layer.

To find the coefficients a_{ij} and b_{ij} in NN Eqs. (4) and (5), an error function E is created:

$$E = \frac{1}{N}\sum_{t=1}^{N}[\mathbf{Z}_t - \mathrm{NN}(\mathbf{X}_t)]^2, \tag{6}$$

where vector $\mathbf{Z}_t$ is composed of observed weather and climate elements, vector $\mathbf{X}_t$ is composed of all predictors, and N is the total number of paired records included in the training dataset. Then, the error function (or cost function) (6) is minimized to obtain an optimal set of all coefficients a_{ij} and b_{ij} via a simplified version of the procedure known as the back propagation training algorithm. The back propagation algorithm searches for the minimum of the error function in weight space through a simplified version of the steepest (gradient) descent method. It partitions the final total cost to each of the single neurons in the network and repeatedly adjusts the weights of neurons whose cost is high, and back propagates the error through the entire network from the output to its inputs.

It is noteworthy, that all NN outputs y_q are included in the same error function (6) and are trained simultaneously using all observed weather variables included in the output vector $\mathbf{Z}_t$. Therefore, during the training, in addition to diminishing the difference between the model variables and corresponding observations, the NN also learns the statistical patterns

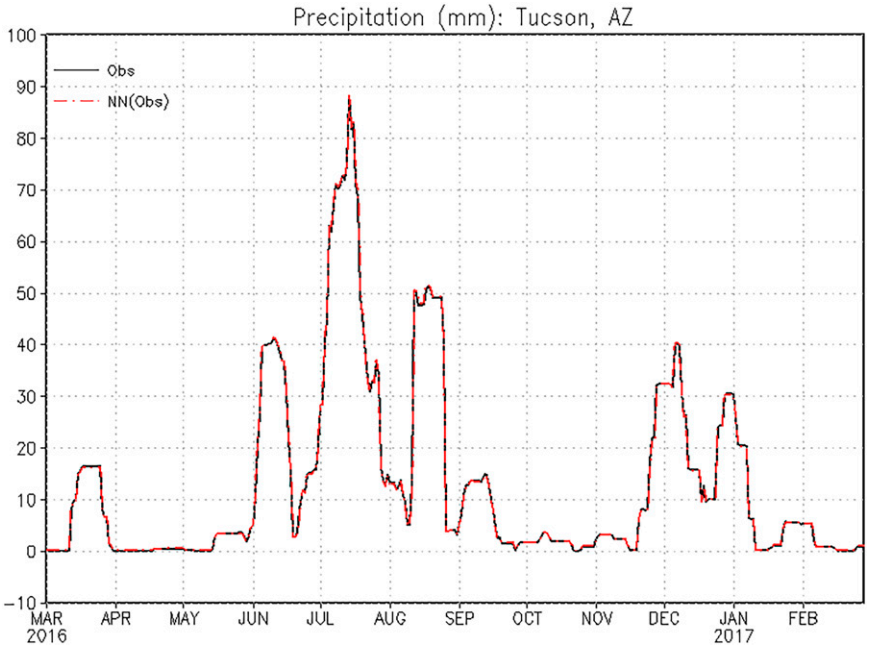

FIG. 2. Time series of the observed 2-week total precipitation from Tucson, AZ (black line), and the independent forecast week-3–4 total precipitation (red line) (with observed precipitation input) from NN-1 at same location and same period.

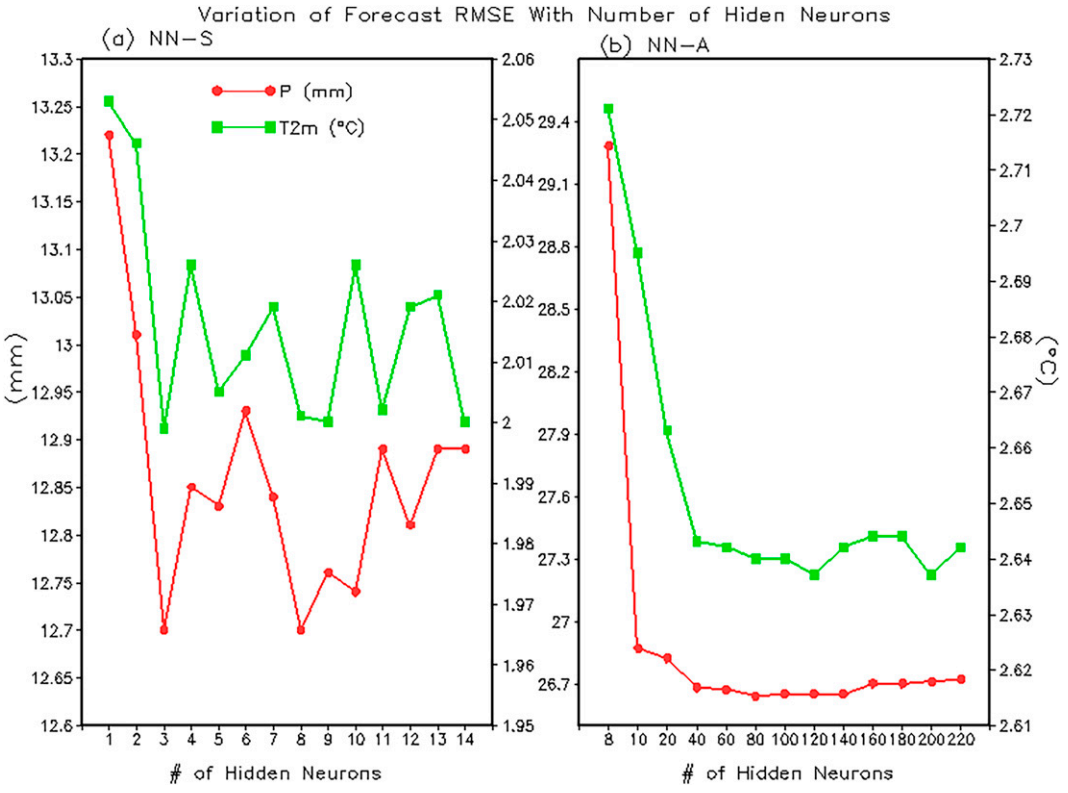

FIG. 3. The mean week-3–4 forecast RMSE (left scale in mm for P; right scale in °C for T2m) as a function of hidden neurons (k) on the independent forecasts from (a) the NN-S on randomly selected nine neighbor points and (b) the NN-A on the CONUS domain. Training period: 1 Jan 1999–December 2016. Validation period: 1 Jan 2017–31 Dec 2018.

and relationships between the model and observed meteorological variables, as well as nonlinear dependencies between variables included in the input vector **X** and in the output vector **Z** in the training dataset. These learned patterns and relationships are used by NN to correct the output of the numerical prediction model.

When the trained NN is applied to new data, all Eq. (4) are evaluated simultaneously using the same coefficients a_{ij} and b_{ij} for all outputs. Thus, from an algorithmic point of view, all m Eq. (4) are one object—a mapping; whereas the MLR approach (3) is a set of independent functions. At first sight, Eqs. (3) and (4) look similar; however, three important differences should be emphasized. In Eq. (4):

(i) The relationship between x_i and y_q is nonlinear when the activation function is nonlinear, such as tanh; therefore, the NN approximation (4) is capable of approximating both the linear and nonlinear components of the mapping (4) (Krasnopolsky 2013).
(ii) The NN approximation (4) can approximate not only pattern relationships and correlations between input variables and output variables, but also the relationship (or covariability) between different observed variables offered by the observed data included in the NN output vector **Z**.
(iii) By including multiple variables in the NN output vector at multiple locations, the NN approach (4) also allows the algorithm to significantly reduce the maintenance burden on the NN equations by generating all weights in one training cycle and storing them in one array file. In contrast, the MLR (e.g., MOS) approach in Eq. (3) usually consists of several thousand to several million individual and independent equations.

3) DESIGN NN ARCHITECTURES

Effective training the NN system requires not only designing the NN architecture with faithful representation of training data, but also careful tuning of the parameters, such as the number of neurons, learning rate, regularization, and adding appropriate auxiliary variables in order to achieve more optimal results, avoid overfitting, and achieve better generalization (Krasnopolsky 2007, 2013; Rasp and Lerch 2018; Fan et al. 2019). In this study, three different NN architectures are designed or configured as follows:

(i) NN-1, which can produce one corrected CFS variable (e.g., P or T2m) at one location (grid point) like the MLR. This pointwise NN setting has an architecture n:K:1 (n inputs at one location: K hidden neurons: one output at one location).
(ii) NN-S, which can produce one and/or several corrected CFS variables (e.g., P and/or T2m) at one or several

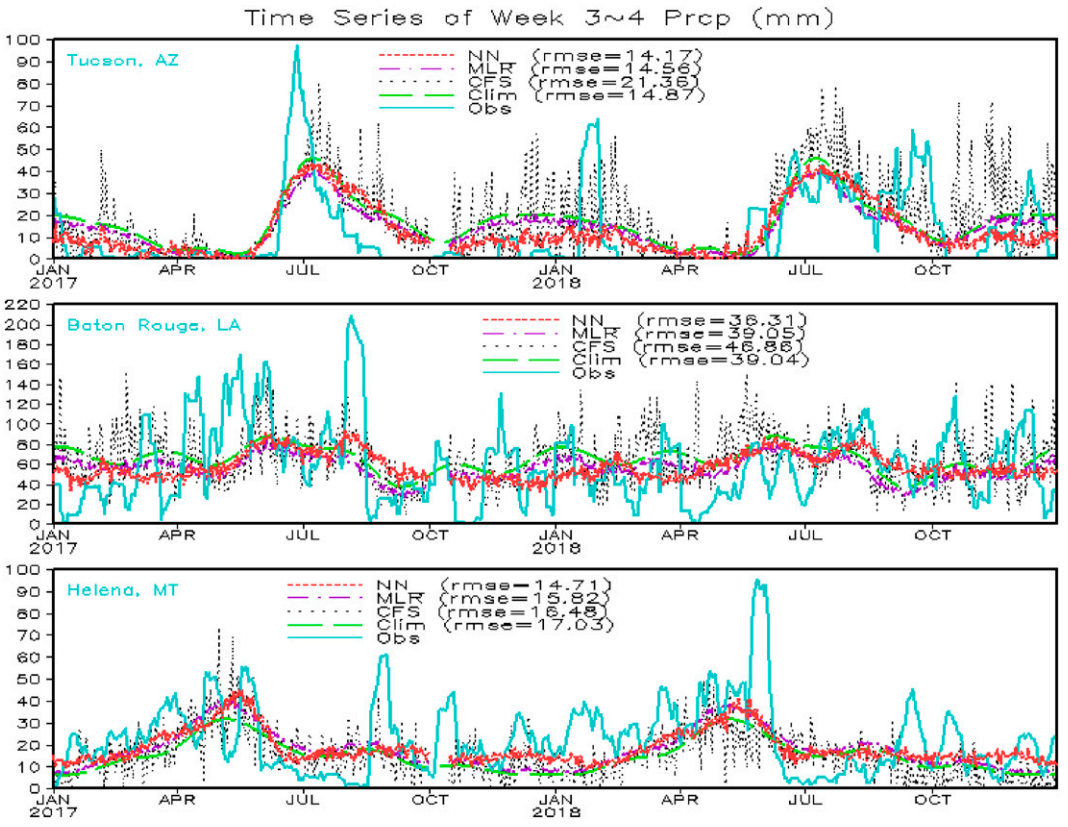

FIG. 4. Time series of daily week-3–4 total P on independent data by NN-A forecasts (red dash), MLR forecasts (purple dot–dash), bias-corrected CFS forecasts (black dot), observed climatology (green long dash), and observations (light blue solid) at three randomly selected locations. The values of RMSE are the averages over the two years.

locations (grid points) simultaneously. One NN-S training can replace two or more MLR equations needed to reach the same goal in the traditional MLR approach. This NN setting has an architecture n:K:m (n inputs from one or several locations: K hidden neurons: m outputs at one or several locations). NN-S can be treated as a small regionalized architecture by setting n:K:m (n inputs from a small region: K hidden neurons: m outputs in a small region that is not necessarily the same as the input region).

(iii) NN-A, which can produce one and/or several corrected CFS variables (e.g., P and/or T2m or more variables) for the entire forecast domain simultaneously. In this case, both $\mathbf{X}_i$ and $\mathbf{Y}_q$ in Fig. 1 are vector variables. This NN setting has an architecture L:K:M (L inputs from all available predictors over all input locations: K hidden neurons: M outputs from all available predictands over the all forecast domain). Here L and M are not necessarily in the same domain. In principle, one NN-A training could replace several thousand MLR equations needed to reach the same goal in the traditional MLR approach. NN-A not only benefits from the flexible NN algorithms, but also takes full advantage of the available big data.

It should be emphasized that the NN-A architecture allows the algorithm to account for both nonlinear relationships among input and output variables, and for the spatial dependency and the covariability among the predictors and predictands by training different predictor and predictand variables over the entire forecast domain simultaneously. During the NN-A training, the NN algorithm tries to minimize the differences between all predictors and predictands at all input and output locations simultaneously to obtain an optimal set of the NN weighting coefficients for all locations. The statistical patterns and relationships learned during the NN training processes are then used by the NN to make the corrected forecasts for each location. Doing it all at once in an NN method does not mean regional differences are neglected.

It should also be noted that the complexity of the NN approximation is partly controlled by the number of hidden neurons K. The more complicated the mapping, the more hidden neurons K are required. However, there is always a trade-off between the desired mapping accuracy and complexity of the NN emulation. The number K should be carefully controlled and kept to a minimum in order to avoid overfitting and to allow a smooth and accurate mapping. The weight initialization method (Nguyen and Widrow 1990) is used for reducing the effects of overfitting and achieving better generalization. The NN weights can be updated inexpensively on a daily basis in real time, through a sequential training approach that works with the training data arriving in real time (record by record).

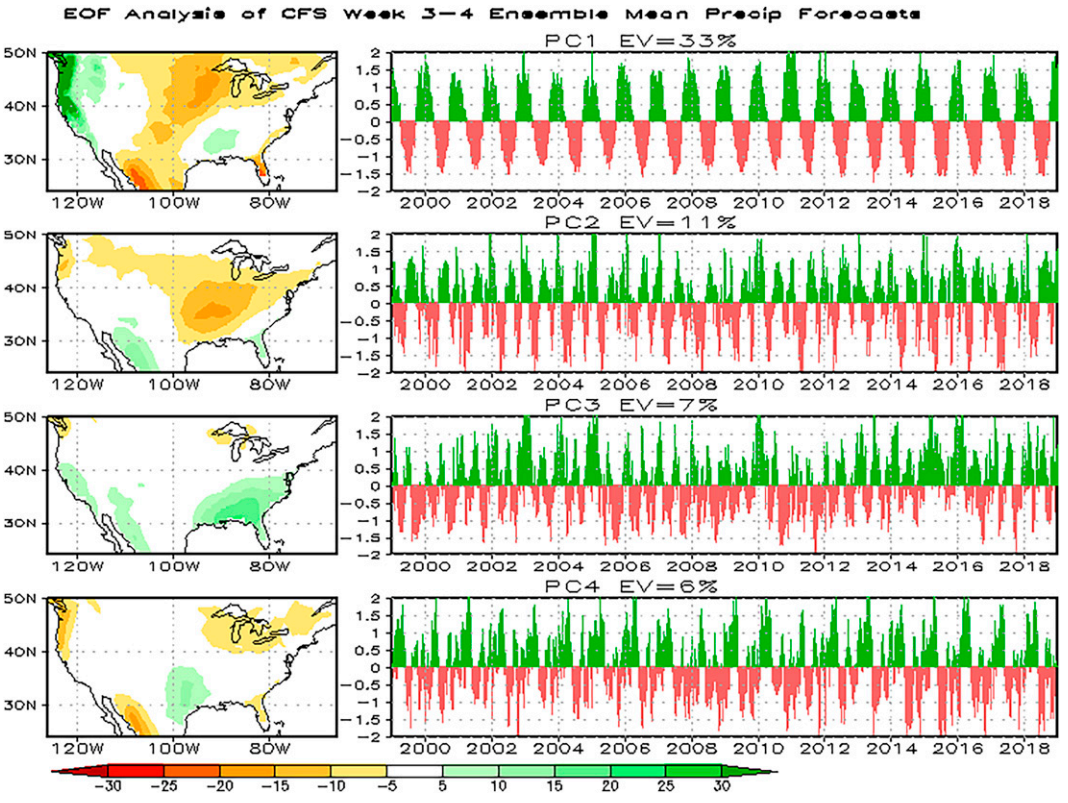

FIG. 5. The first four leading EOF patterns (scaled by the RMS value of the associated PCs; mm) and their correspondent time series (normalized to unit variance) from bias-corrected CFS ensemble mean (average of 0000, 0600, 1200, and 1800 UTC) week-3–4 forecast total precipitation.

3. NN check, optimization, data representation, and predictability analysis

a. *NN sanity check*

To evaluate the accuracy of the NN approximation and also the applicability of NN software used, the NN-1 was trained at several randomly selected locations to approximate the identical mapping:

$$\mathbf{X} = M(\mathbf{X}), \tag{7}$$

where **X** could be any predictor and predictand variable. If the NN is working properly, a mapping performed between a variable **X** and itself should return the variable **X**. Figure 2 shows the independent week-3–4 precipitation mapping from the NN-1 approximation and the observed 2-week total precipitation in the same period from one of several randomly selected locations (Tucson, Arizona). The NN-1 training period is from 1 January 1999 to 31 December 2015. The experiment indicates that the NN algorithm can almost perfectly reproduce the observed precipitation for the independent forecast period from 1 March 2016 to 28 February 2017. The difference between the above NN-1 and the observation varies between −0.2 and 0.4 mm. Similar mapping also has been done on the CFS week-3–4 forecast precipitation with similar results. The NN-1 also can reproduce the noisier CFS model forecasts very well with slightly higher mapping errors for reasons noted below in section 3c.

b. *Optimal number of hidden neurons*

The complexity of the NN mapping can be controlled by varying the number of the NN hidden neurons. To evaluate the optimal size (k) of the hidden neurons in Eq. (4) for the NN week-3–4 P and T2m forecasts, some criteria, such as the root-mean-square error (RMSE), bias, correlation, scatter, skewness, and others are used together to select the optimal number of hidden neurons. A set of 14 NN-S (9 neighbor points used here) are trained with varying k from 1 to 14. The results based on the widely used RMSE are shown in Fig. 3a with the independent NN forecasts. For the precipitation forecast on a pointwise basis, $k = 3$ is the optimal number of NN hidden neurons. Under the chosen NN-S setting, using more neurons ($k > 3$) does not reduce the forecast error, probably because the NN-S starts to fit more noise from the data. The results from NN-1 are very similar to those from the above NN-S settings. When compared with the benchmark MLR method with the same predictors, both NN-1 and NN-S do a better job at predicting the observed precipitation. However, in terms of optimal hidden neurons, the mapping from both NN-1 and NN-S is not strongly nonlinear (i.e., only a small number k can be used beneficially).

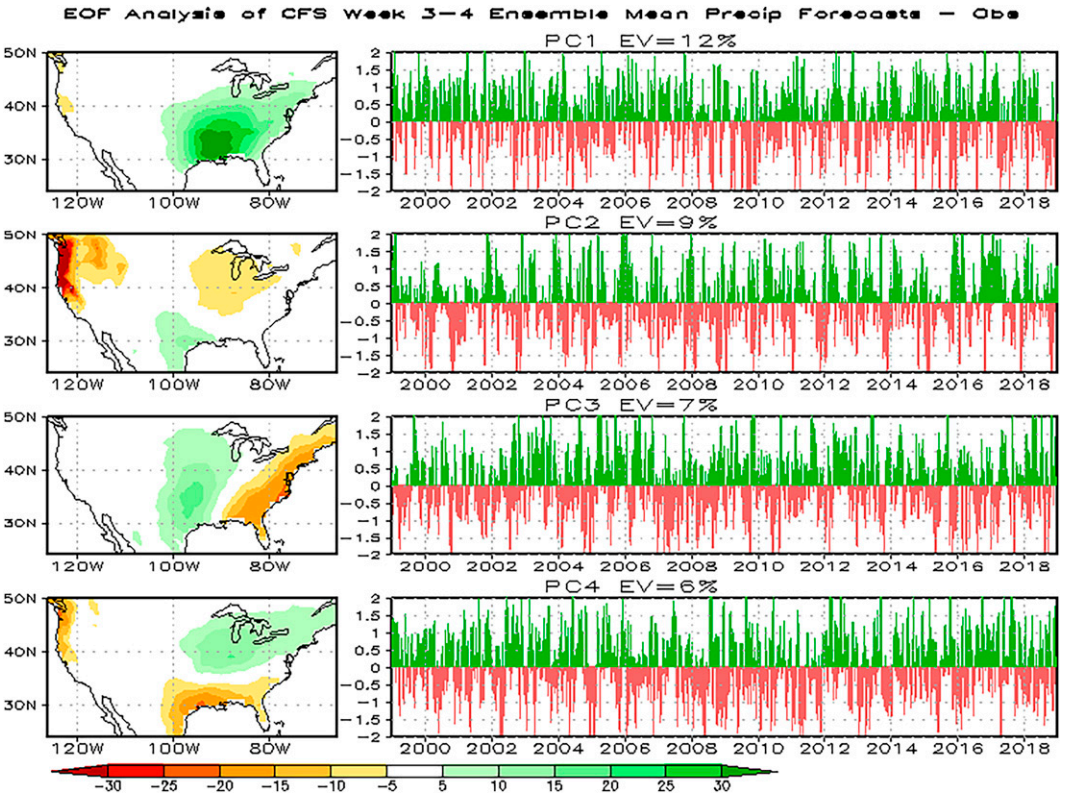

FIG. 6. The first four leading EOF patterns (scaled by the RMS value of the associated PCs; mm) and their related time series (normalized to unit variance) from forecast errors (bias-corrected CFS week-3–4 ensemble total P minus observation).

To evaluate the optimal number of hidden neurons in Eq. (4) for the NN-A, another set of 15 NN-A tests k varying from 10 to 220 has been conducted. The mean forecast RMSE derived from the week-3–4 forecast P and T2m using independent testing dataset is shown in Fig. 3b. The results indicate that if separately forecasting P or T2m, $k = 120$ is near the optimal number of hidden neurons. In contrast, forecasting P and T2m together requires $k = 200$ hidden neurons for optimal results. This indicates that the NN-A architecture with more than 100 hidden neurons is significantly more nonlinear than NN-1 and NN-S architectures with a far lower set of hidden neurons (~2–3). In other words, with NN-A the nonlinear and pattern-wise corrections for the week-3–4 forecasts of both P and T2m over the entire forecast domain (CONUS) is much more ambitious and potentially beneficial than the point-wise correction for just a single location or several neighboring sites. Therefore, the NN-A mapping, which is designed to take advantage of the flexible NN algorithm and big datasets and to do more sophisticated patternwise corrections, presents much more nonlinear features, as expressed in terms of the optimal number of hidden neurons. In general, the computational cost increases linearly with hidden neurons used.

c. *Data representation*

It is important to understand the characteristics of the data being analyzed because it will inform choice in the NN architectures. When looking at the time series of the daily CFS week-3–4 forecast total precipitation and its corresponding observed total precipitation, two significant differences emerge:

1) The observed total precipitation (e.g., light blue solid curves in Fig. 4) is smoother than its corresponding CFS week-3–4 forecast total precipitation (black dot curves in Fig. 4). This is because each of the daily observed 2-week total precipitation has a 13-day overlap of data on its adjacent date. However, for each of the daily CFS week-3–4 forecast total precipitation, the model forecasts do have such 13-day overlap in terms of dates, but they come from different initializations. Due to forecast error growth, the CFS data are noisier compared to observations.
2) The trajectories of the daily CFS week-3–4 forecast total precipitation at the same location, but initialized at four different initial times (0000, 0600, 1200, or 1800 UTC on each day), can be very different after 4 weeks of model integration. However, how to address the above issues in training datasets properly is crucial for improving the NN training.

To minimize impacts related to the above two issues, the empirical orthogonal function (EOF) analysis was used to explore the spatial–temporal variations of the bias-corrected CFS week-3–4 forecast P and T2m in the period covering

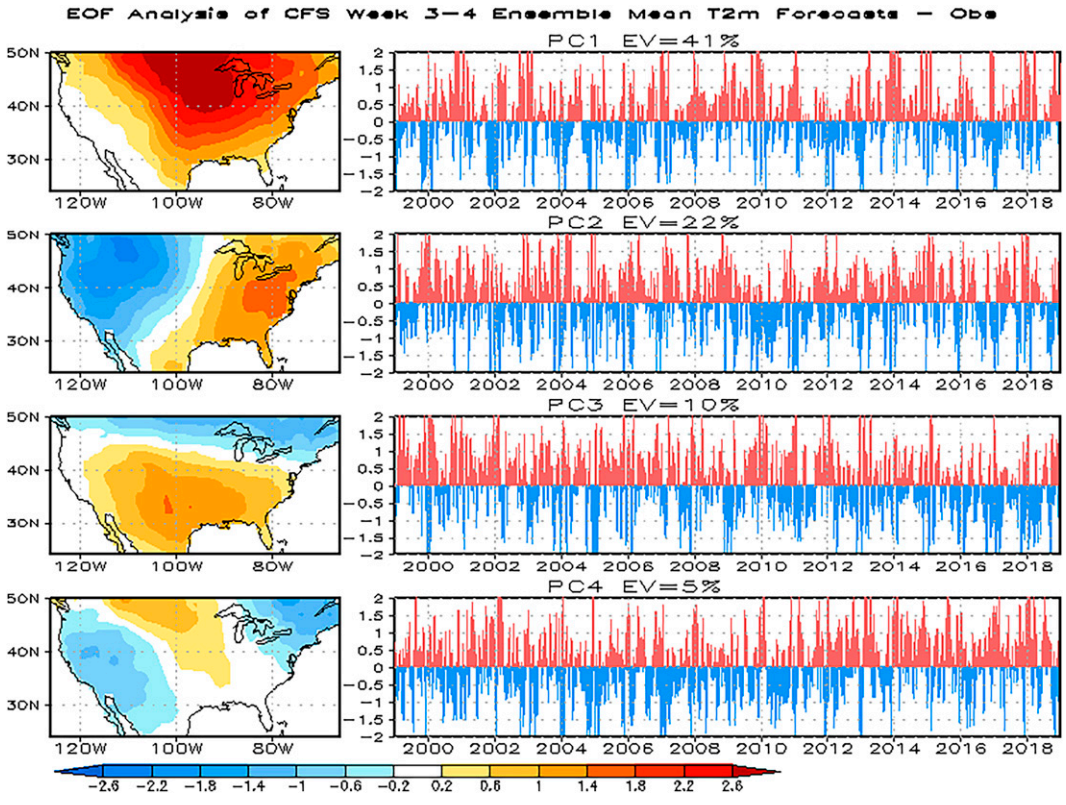

FIG. 7. The first four leading EOF patterns (scaled by the RMS value of the associated PCs; °C) and their related time series (normalized to unit variance) from forecast errors (bias-corrected CFS week-3–4 ensemble mean T2m minus observation).

1 January 1999–31 December 2018 from four different initial times. The encouraging results indicate that the leading EOF patterns and the variations of their corresponding time series are quite similar from the four different initial times (0000, 0600, 1200, and 1800 UTC). Figure 5 depicts the first four leading EOF patterns and their corresponding time series from the CFS week-3–4 ensemble mean total precipitation (averaged from 0000, 0600, 1200, and 1800 UTC), which are similar to results from the individual CFS week-3–4 total precipitation forecasts initialized at 0000, 0600, 1200, and 1800 UTC. The spatial patterns of the leading EOF modes are relatively large-scale and the temporal variations are dominated by annual and semiannual cycles. The first four modes account for about 57% of the total variance from the ensemble mean forecasts, but only about 44% of the total variance from individual members.

The EOF analysis was applied to the corresponding observed 2-week total precipitation. The first four leading EOF modes account for about 42% of the total variance from the observed 2-week total precipitation. It shows that at large scales (the first four leading EOF patterns) the CFS week-3–4 forecast total precipitation bears many similarities with observational data. However, the corresponding time series from the observational data are noisier, except for the first leading EOF, the variation of its time series is also dominated by a very strong annual cycle.

The same EOF analysis was also applied to the CFS week-3–4 ensemble mean forecast T2m and its corresponding observed 2-week mean T2m. The results (not shown) reveal that the leading EOF spatial patterns from the CFS week-3–4 forecast T2m are dominated by large-scale patterns and are remarkably similar to those from the observational data. However, the amplitudes and timing are main issues for the CFS week-3–4 forecasts. The first four leading EOF modes account for 84% of the total variance for the CFS week-3–4 ensemble mean T2m forecasts and 78% of the total variance for the observational data. This suggests that the structures of the T2m are simpler than the *P*.

The above results suggest that the CFS is comparatively better at predicting large-scale patterns and low-frequency variations in the observed *P* and T2m than at capturing fine-scale variations of those highly parameterized and unresolved physical processes. These results indicate important suggestions in the NN training processes:

(i) Using more reliable and robust large-scale pattern information in the NN predictors (e.g., the CFS forecast *P*, T2m, and Z500) may prove to be more beneficial for the NN forecasts.
(ii) Using ensemble means (average from 0000, 0600, 1200, and 1800 UTC) may further improve data representation, because ensemble mean not only smooths spatial–temporal

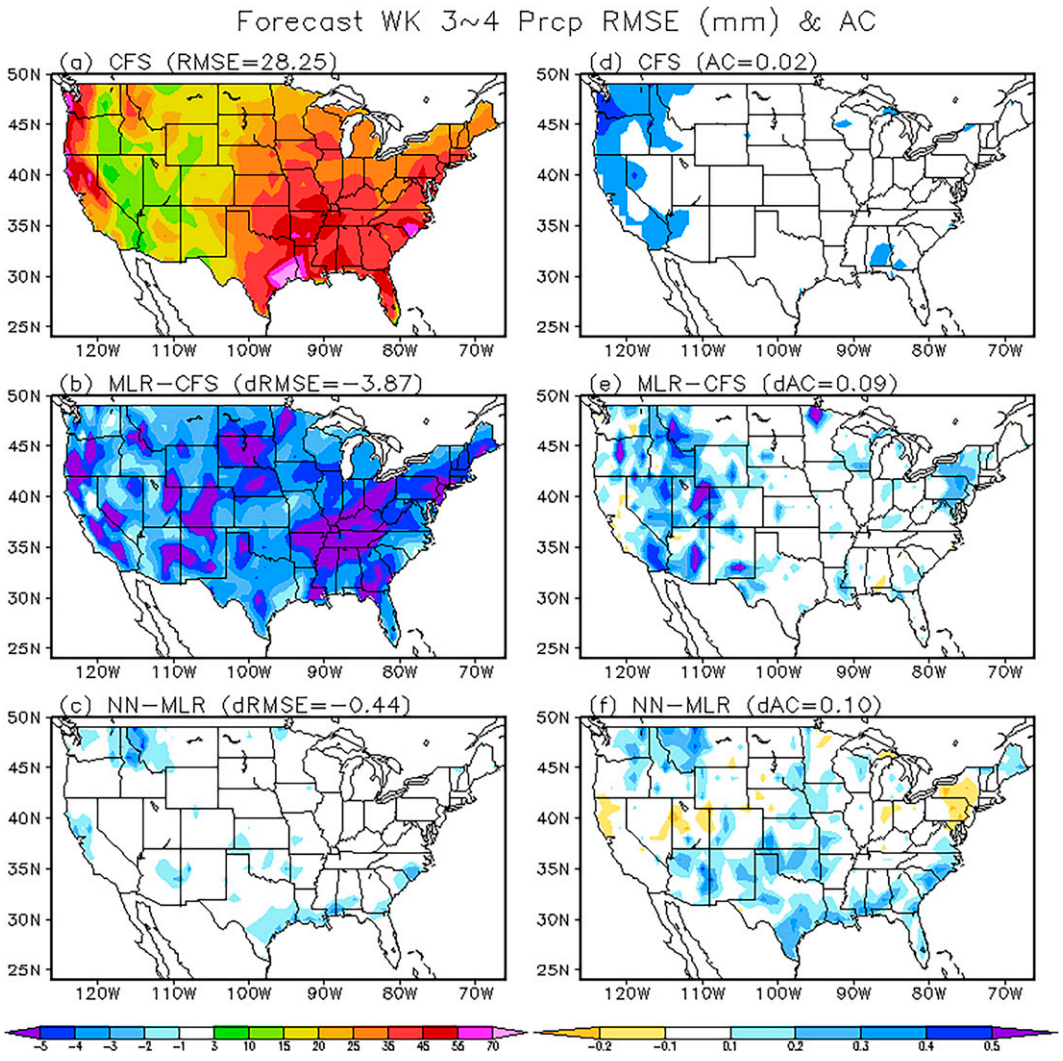

FIG. 8. (left) The RMSE and (right) AC of daily week-3–4 *P* by (a),(d) bias-corrected CFS forecasts against observation; (b),(e) RMSE differences (dRMSE) and AC differences (dAC) between CFS and MLR forecasts; and (c),(f) as in (b) and (e), but between MLR and NN forecasts. Training period: 1 Jan 1999–31 Dec 2016. Testing period: 1 Jan 2017–31 Dec 2018. The values in panel titles are the averages over the CONUS domain. For the AC and dAC, the shaded regions exceed the 99% confidence level.

noise in the input data, but also increases the percentage representation (explanation) of the total variance of the data.

d. Analysis of the CFS week-3–4 P and T2m forecast errors

The similar EOF analysis was also applied to the bias-corrected CFS week-3–4 ensemble mean *P* and T2m forecast errors (i.e., forecast minus observation). Moreover, such an EOF analysis can also provide insight into limits of CFS week-3–4 *P* and T2m forecasts (in other words, what do the CFS week-3–4 forecast errors look like and to what extent can the errors be removed?). Ideally, if the forecast errors are either constant or vary regularly, then nearly all errors can be removed easily. If the forecast errors are characterized by large-scale spatial pattern and low-frequency temporal variations, then at least part of the errors can be corrected in most cases. The worst scenario is if forecast errors are white noise–like. In that scenario, there is no way the forecast errors can be corrected or removed, no matter what methods are used.

The results (Fig. 6) reveal that in general these leading EOF patterns from the bias-corrected CFS week-3–4 ensemble mean *P* forecast errors are relatively large-scale patterns and feature some low-frequency variations (e.g., annual cycle), but are much noisier compared with the time series in Fig. 5. These forecast errors are caused by the model deficiency, errors in initial and boundary conditions, definition differences of the model forecast versus observed variables, and the nature of predictands. They may be partly removable by some post-processing techniques. For example, the first and second leading forecast error modes feature relatively large-scale patterns and

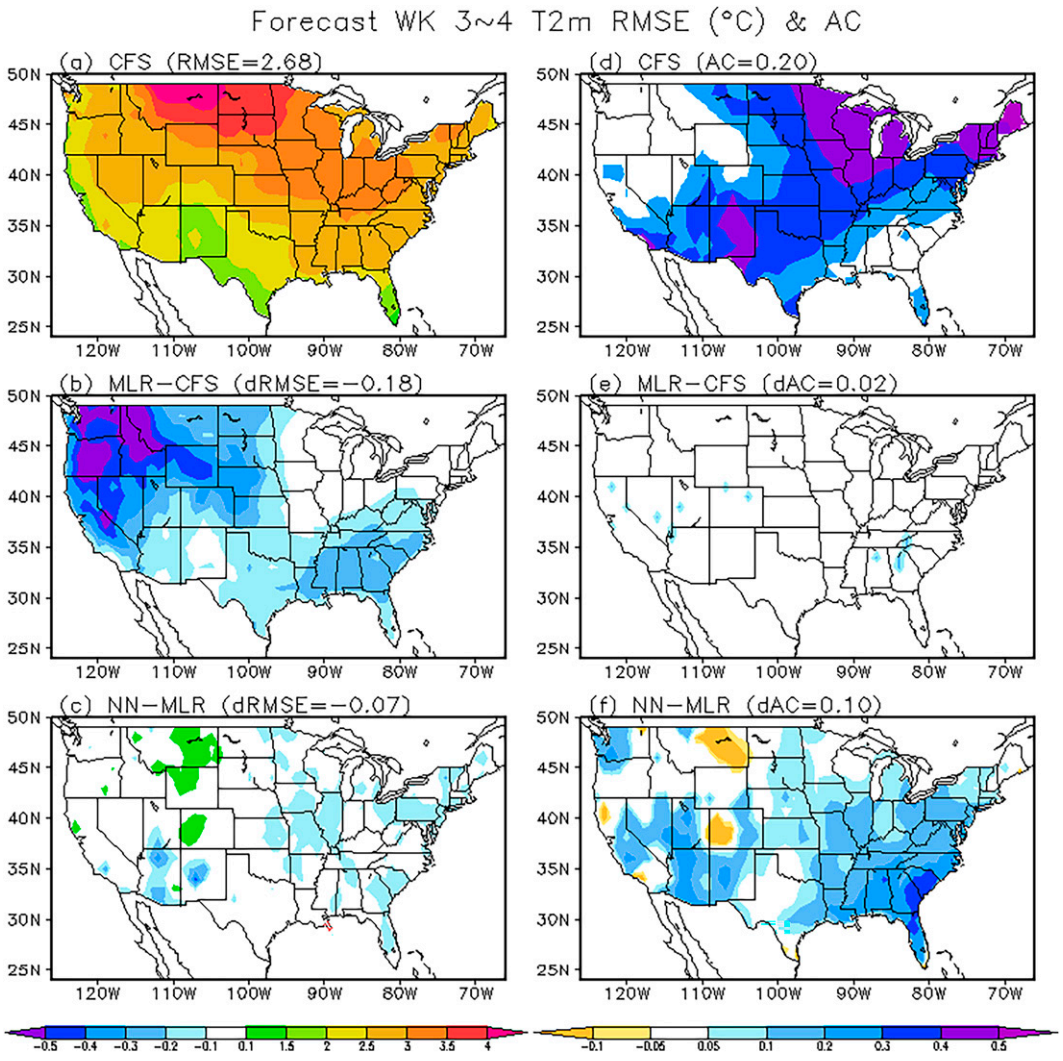

FIG. 9. (left) The RMSE and (right) AC of daily week-3–4 T2m by (a),(d) bias-corrected CFS forecasts against observation; (b),(e) RMSE differences (dRMSE) and AC differences (dAC) between CFS and MLR forecasts; and (c),(f) as in (b) and (e), but between MLR and NN forecasts. Training period: 1 Jan 1999–31 Dec 2016. Testing period: 1 Jan 2017–31 Dec 2018. The values in panel titles are the averages over the CONUS domain. For the AC and dAC, the shaded regions exceed the 99% confidence level.

are dominated by the annual cycle. This means that the CFS does not produce a satisfactory forecast for the observed annual cycle in precipitation over CONUS in terms of the amplitudes and phases. The good news is that usually part of these climate-like forecast errors (or climate biases) can be easily removed by some bias correction methods (Fan and van den Dool 2011).

It should be mentioned that the above forecast errors from the first four leading EOF modes only account for about 34% of the total variance from the week-3–4 ensemble mean *P* forecast errors, meaning limited opportunity for forecast improvement over the CFS week-3–4 *P* forecasts. Not all of these forecast errors are correctable (or removable). In general, the higher the EOF leading mode, the smaller the scale in spatial pattern and the noisier in temporal variation. Usually these small-scale and high-frequency forecast errors are even more difficult to remove. To some extent, they may reflect the prediction limits for the CFS week-3–4 precipitation forecasts.

For the bias-corrected CFS week-3–4 ensemble mean T2m forecast errors, the most dominant (i.e., the first 4 leading EOF) forecast error patterns (Fig. 7) show large-scale spatial patterns very similar to the forecasts and observations individually. The corresponding time series also feature some low-frequency (e.g., annual cycle) variations. The first four leading EOF modes account for 78% of the total variance for the forecast errors, much higher than the forecast *P* and therefore potentially more predictable than the *P*. This means that a large part of the T2m forecast errors can be represented by just a few leading EOF modes. These climate bias–like

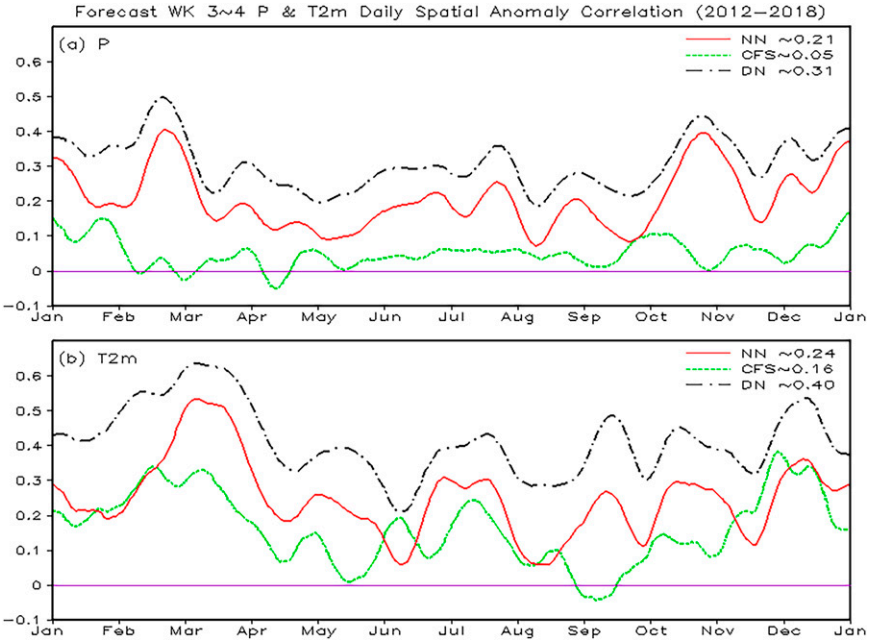

FIG. 10. Mean time series of the daily week-3–4 (a) *P* and (b) T2m spatial anomaly correlations over the CONUS (5-day running mean applied) among 1) NN (red): independent NN forecasts (Test 3) and observations, 2) CFS (green): bias-corrected CFS forecasts and observations, and 3) DN (black): dependent NN (all data January 1999–31 Dec 2018 used for NN training) and observations. Validation period: 1 Jan 2012–31 Dec 2018.

forecast errors indicate that the CFS is very good at forecasting the week-3–4 T2m spatial patterns but has problems forecasting their amplitudes and timing. Compared with the forecast *P* errors, these T2m forecast errors should be comparatively easier to remove in general. However, because the time series of the first four leading EOF modes are as noisy as is the case for forecast *P* errors, it may still be difficult to remove these T2m forecast errors. Some features of the above forecast errors are also true for the short-range weather forecasts from day 1 up to week 2 in some forecast systems, such as the NCEP Global Forecast System (GFS) (Fan and van den Dool 2011; Fan et al. 2015).

4. NN week-3–4 P and T2m forecasts

a. Forecasts from different NN architectures

There is considerable need for skillful week-3–4 forecasts. However, forecasting for this time scale is one of toughest areas and prediction skills are very low in general. One open question to be explored here is if the ML (e.g., the nonlinear NN systems used here) techniques with the bias-corrected CFS predictors as input can outperform the bias-corrected CFS *P* and T2m forecasts in week-3–4 time scale and the benchmark MLR tools with the same inputs as the NN systems.

The daily time series of the week-3–4 *P* from the observational data, the bias-corrected CFS forecast, the MLR forecast, and the NN forecast for three randomly selected locations are also shown in Fig. 4 above. Overall, the results from the NN forecasts are slightly better than the results from the MLR method. In general, both methods beat the climatology forecasts, but not by much. The forecast skill (in terms of the RMSE) for the week-3–4 precipitation is still quite low. The results also indicate that the resulting week-3–4 NN precipitation forecasts by using the ensemble mean from four initial times (0000, 0600, 1200, and 1800 UTC) are in general better than the NN forecasts by using the CFS forecasts from an individual ensemble member. Similarly, the week-3–4 forecasts by using NN-A generally outperform those from the NN-1 or NN-S, since NN-1 and NN-S settings do not fully take the benefits offered by the NN algorithms and big data by only working on very small portions of data at a given time.

As mentioned earlier, the NN-A setting can take advantage of the flexible NN algorithm that accounts for complicated linear and nonlinear relationships, spatial dependency, and covariability among predictors and predictands. The NN-A setting was explored with a variety of predictors and predictands. The results show that using observed daily *P* and T2m climatologies as predictors outperforms other auxiliary predictors, such as $\sin(\tau)$, $\cos(\tau)$, latitude, longitude, elevation, station ID, etc., because all these effects are already well represented by the climatology variables. It also shows that using the same group of predictors to forecast the week-3–4 *P* and T2m together (covariability between observed *P* and T2m counted) is better than forecasting the same *P* and T2m separately.

In the following part of this paper, the focus will turn to the more beneficial NN-A setting by forecasting *P* and T2m together. The five predictors used in NN training include the

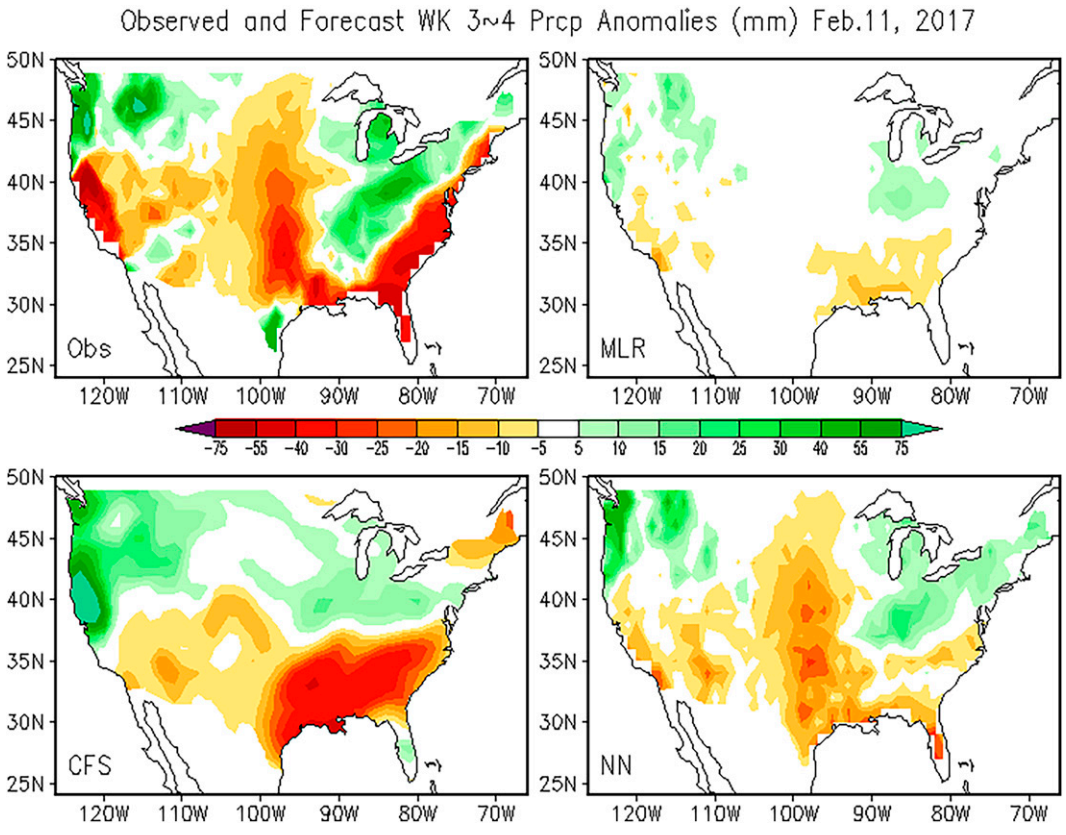

FIG. 11. The observed (Obs), CFS, MLR, and NN forecast week-3–4 *P* anomalies on 11 Feb 2017.

CFS bias-corrected ensemble mean week-3–4 total *P*, anomaly T2m and Z500, and the observed *P* and T2m climatologies. The two predictands are observed total *P* and anomaly T2m.

b. *Verification of the daily NN week-3–4 P and T2m forecasts*

In this subsection, the spatial-temporal distribution of the week-3–4 *P* and T2m forecast skill will be explored. Figure 8 shows that the root-mean-square error (RMSE) and anomaly correlation coefficients (AC) of the bias-corrected ensemble mean CFS precipitation forecasts when adjusted by the NN-A are overall better than the adjusted forecast obtained from the benchmark pointwise MLR method. Here the NN-A and MLR training period is from 1 January 1999 to 31 December 2016. The period of 1 January 2017–31 December 2018 is used as an independent verification period. The above results indicate both the NN-A and MLR methods improve the bias-corrected CFS week-3–4 precipitation forecasts, and especially the forecast skill in various parts of the western CONUS is encouraging (AC over 0.4 or 0.5). However, some degradation is also seen in limited areas for the NN-A when compared with the MLR, such as near the northeastern United States. In general, the NN-A forecasts show better forecast skills than the MLR forecasts over most locations in term of both the RMSE and AC, with the AC improvement more robust. This may also indicate that accounting for the nonlinear relationship between the predictors and predictands, as well as making use of colinearity plays an important role in precipitation forecasting.

As Fig. 8, Fig. 9 shows the week-3–4 T2m forecast skills from the bias-corrected CFS, the MLR, and the NN-A methods. Both the NN-A and the MLR are able to reduce the CFS forecast errors in terms of the RMSE, although not as much. The performance of the NN-A is slightly better than the MLR method, in terms of the RMSE forecast skill. However, in terms of the AC forecast skill it is encouraging that the NN-A method is significantly better than the MLR method in most places except some degradation in limited areas. Again, this may indicate that the nonlinear relationship plays an important role between the predictors and predictands at improving the week-3–4 T2m forecasting.

c. *Three k-fold cross validations*

In this subsection, three multiyear daily NN week-3–4 *P* and T2m (independent) forecast experiments were conducted to further explore whether the week-3–4 *P* and T2m forecast improvement from the ML (e.g., the nonlinear NN systems used here) technologies are robust, reliable, and meaningful when compared with the bias-corrected CFS week-3–4 forecasts. Three *k*-fold cross-validation tests were performed as follows:

Test 1: Remove 3 years of daily paired data from a 20-yr period (1999–2018) of daily pooled data sequentially and yearly,

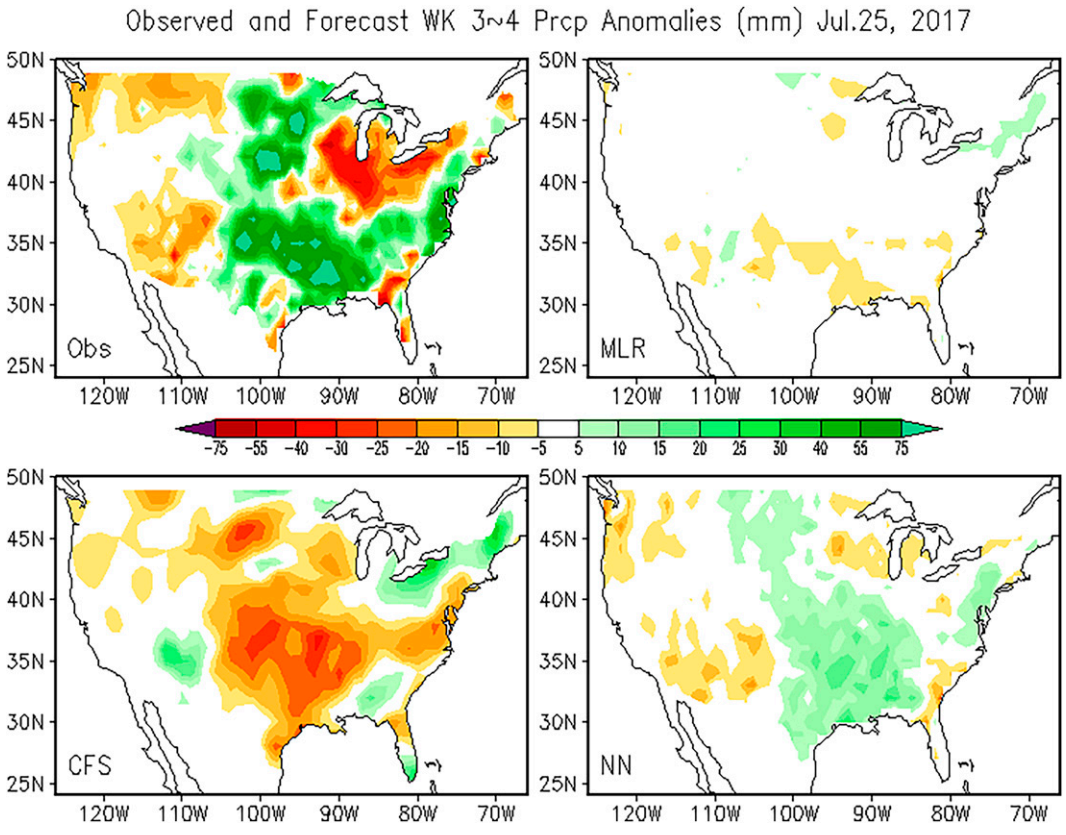

FIG. 12. As in Fig. 11, but for 25 Jul 2017.

then use the middle year only as the independent forecast (testing) dataset, with the remaining 17 years daily paired data employed as training dataset. For the year 1999 and 2018, only 2 years of daily data are removed, with the far side year (i.e., 1999 or 2018) used only as an independent forecast (testing) dataset, and the remaining 18 years daily data used as the primary training dataset. The above procedure was repeated yearly for 20 times so that the independent NN-A experiments were performed every year from 1999 to 2018.

Test 2: Remove one year of daily paired data from a 20-yr period (1999–2018) of daily pooled data sequentially and yearly, taking these as the independent forecast (testing) dataset and the remaining 19 years daily paired data as the training dataset. The above procedure was repeated yearly for 20 times. Therefore, another 20 yearly independent NN-A experiments were performed from 1999 to 2018.

Test 3: Remove 60 days of daily paired data (each in 2012–18) from a 20-yr period (1999–2018) of daily pooled data sequentially as the independent forecast (testing) dataset, using the remaining 19 plus years of daily paired data as the training dataset. A total of 42 NN-A 60-day independent experiments cover the period from 2012 to 2018.

Figure 10 shows the time series of the daily week-3–4 forecast *P* and T2m spatial anomaly corrections (AC) averaged over 2012–18 from (i) the NN (Test 3) independent forecasts, (ii) the NN dependent forecasts (training data covering 1999–2018, can be viewed as the upper limit of NN forecasts) and (iii) the bias-corrected CFS forecasts. The results indicate that the NN techniques indeed can make a robust improvement for the week-3–4 *P* and T2m forecasts over the bias-corrected CFS forecasts. Both of the independent NN week-3–4 *P* and T2m forecasts are improved over the bias-corrected CFS *P* and T2m forecasts, with the NN week-3–4 *P* forecast improvement (mean AC from 0.05 to 0.21) being a more robust improvement across all times of the year, while the NN week-3–4 T2m forecast improvement (mean AC from 0.16 to 0.24) is less robust than the *P* forecasts. The results also show that the independent NN (Test 3) week-3–4 *P* and T2m forecasts have very similar tendencies as the dependent NN week-3–4 *P* and T2m forecasts. This indicates that sometimes the dependent NN forecast systems are more predictable than other times and the independent NN forecast systems follow the same ups and downs.

For Test 2, the mean time series of the daily NN week-3–4 forecast *P* and T2m spatial anomaly correlations closely follow the results from Test 3, with forecast skill degraded slightly (mean AC from 0.21 to 0.20 for *P* and from 0.24 to 0.22 for T2m), due to the training sample data being farther away from the dependent training sample data. For Test 1, its mean time series of the daily NN week-3–4 forecast *P* and T2m spatial anomaly correlations also follow the results from Test 3 quite well with forecast skill further degraded (mean

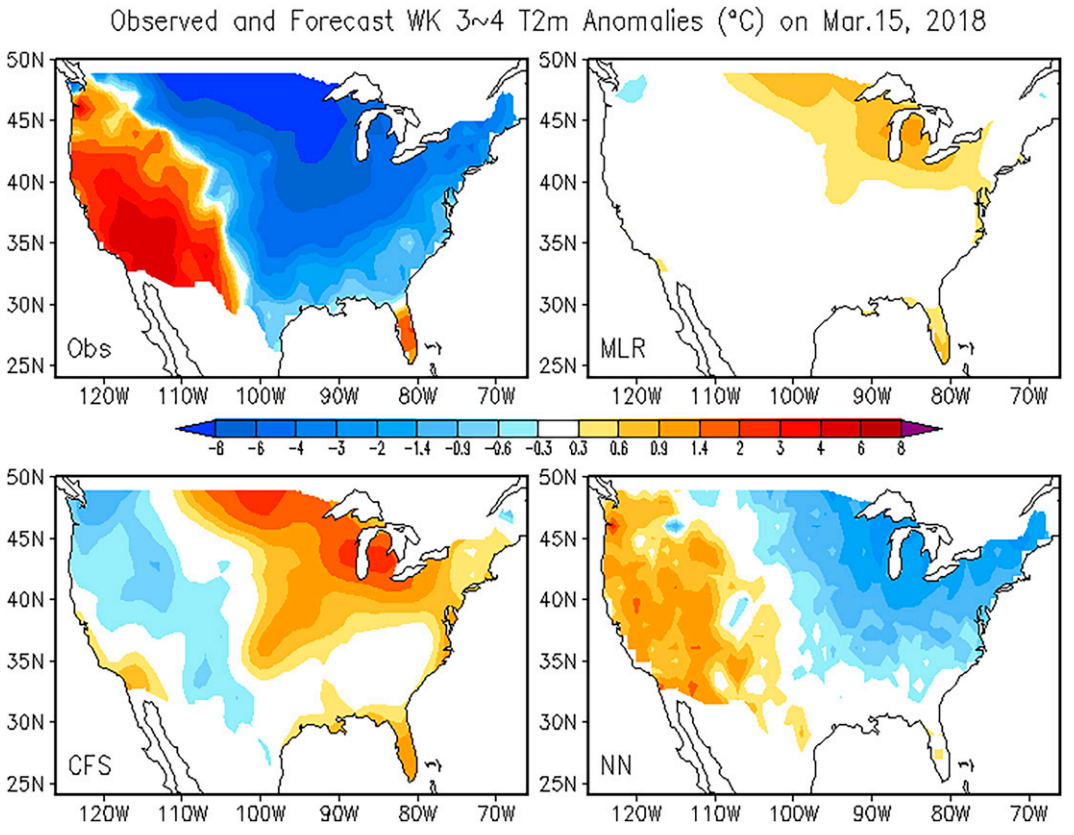

FIG. 13. The observed (Obs), CFS, MLR, and NN forecast week-3–4 T2m anomalies on 15 Mar 2018.

AC from 0.21 to 0.19 for *P* and from 0.24 to 0.20 for T2m), due to the training sample data being even farther away from the dependent training data. Therefore, in principle, if we can nudge the training sample (e.g., Test 3, but withhold 30 days data as independent test data) closer to the dependent training sample, the forecast skill should be further improved when compared with Test 3.

d. Comparison of different forecast methods

Finally, when checking the overall week-3–4 forecast performance of three (CFS, MLR, and NN) forecasts over the multiyear verification period, both the MLR and the NN consistently beat the bias-corrected CFS. Of the MLR and NN forecasts, the NN forecasts significantly outperformed the pointwise MLR forecasts in many respects. Figures 11–14 depict examples of the observed week-3–4 *P* and T2m anomalies, together with the corresponding week-3–4 CFS, MLR, and NN forecast *P* and T2m anomalies. In these cases, the NN techniques show very encouraging and impressive ability to turn around or reverse the incorrect *P* and T2m forecast patterns seen in the bias-corrected CFS forecasts. Usually the above "turn around" events can persist for several days and can happen in any season. One possible explanation for this is that model forecast spatial patterns are systematically and frequently offset in certain time frames and locations with certain *P*, T2m, and Z500 patterns, and the NN architecture used here has the ability to allow the NN algorithm to remember what happened. Then, the NN system can determine what is (are) the best and most important forecast input(s), where these (group) points are located, and how to minimize the forecast errors in multiple dimensions for best mapping the target (predictand) points, an accomplishment that cannot be done with the traditional pointwise and spatially independent MLR method.

5. Conclusions and discussion

In this study, NN techniques are used to improve the NCEP CFS week-3–4 *P* and T2m forecasts, and to explore the predictability of the CFS week-3–4 *P* and T2m forecasts. Benefiting from the great advances in ML in recent years, NN techniques show some advantages over traditional statistical methods such as MLR: its flexible algorithms can account for complicated linear and nonlinear relationships, spatial dependency, and covariability in predictors and predictands, and at the same time, it is able to handle big data easily and efficiently.

Knowing the datasets well and using a better data representation are very important before applying NN training. The EOF analysis indicates that the CFS is very good at predicting large-scale patterns and low-frequency variations in observed *P* and T2m, but less so at capturing highly parameterized and unresolved processes in *P* and T2m. Better data representation

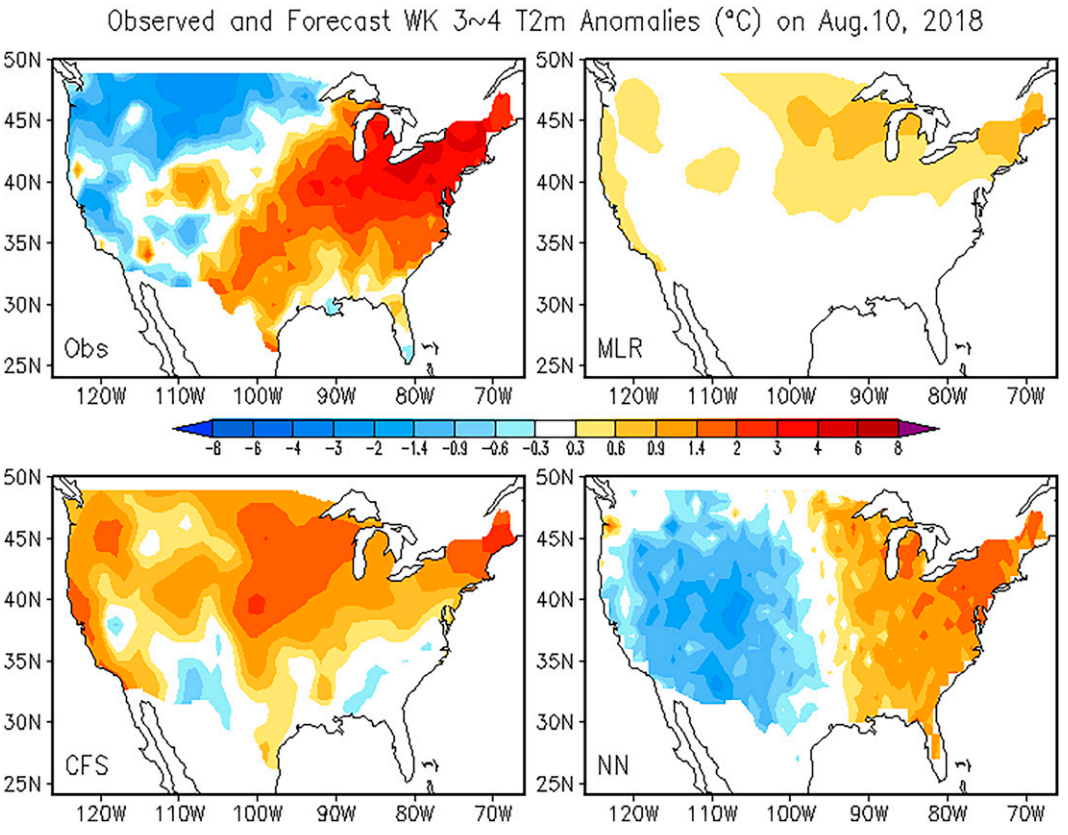

FIG. 14. As in Fig. 13, but for 10 Aug 2018.

can also be achieved by using ensemble means to increase the explained percentage of total variance and to reduce noise in the data.

The EOF analysis of the CFS week-3–4 *P* and T2m forecast errors provides some insight on the extent that forecast errors are correctable. The results reveal that the spatial-temporal structures of the most dominant CFS week-3–4 forecast errors have relatively large-scale spatial patterns with low-frequency variations, such as the annual cycle, namely, climate biases. This is also true for some short-range weather forecast systems from day 1 up to 2 weeks. In general, at least part of these large-scale and low-frequency forecast errors are removable.

Different NN configurations are used to compare to the benchmark MLR postprocessing method. By designing more beneficial NN setups, the NN-A architecture, is able to account for not only nonlinear features or relationship within a given large dataset, but also spatial dependency (e.g., pattern relationships) by training different predictors and predictands from the entire forecast domain simultaneously. Moreover, the NN-A architecture can also account for the covariability among the predictands by training different predictands simultaneously. Together, these learned statistical patterns and relationships from the NN training processes are then used to correct the CFS week-3–4 forecasts. The NN-A has the ability to extract more complicated and high-level information hidden behind big data. Thus, the NN-A can perform more sophisticated forecast corrections, such as reversing incorrect forecast patterns, which is impossible for the traditional method like pointwise MLR.

Although the improvement for the week-3–4 *P* and T2m is very encouraging, the overall forecast skill (in terms of both RMSE and AC skills) for the week-3–4 *P* and T2m predictions is still quite low, when compared to the week-2 outlooks. Since the NN forecasts here critically depend on the quality of the CFS forecast inputs, improving the CFS itself remains critically important to improve the week-3–4 forecasts. Another potential way to improve the CFS week-3–4 *P* and T2m forecasts is to do more detailed dynamic analyses and to consider including more related predictors. Using more advanced NN architectures (e.g., deep NNs) and more advanced ML techniques could also help to improve the forecasts. Further studies can advance this capability.

Acknowledgments. We thank Drs. Peitao Peng and Kyle MacRitchie for their constructive suggestions and insightful comments. Thanks also due to the three anonymous reviewers for their excellent comments and suggestions. This work was partly supported by the NOAA OSTI program.

Data availability statement. CFSv2 can be downloaded from the NOAA/NCEI (former NCDC) website, CPC daily T2m analysis is available at ftp://ftp.cpc.ncep.noaa.gov/precip/PEOPLE/wd52ws/global_temp/, and CPC daily *P* analysis is

available at ftp://ftp.cpc.ncep.noaa.gov/precip/CPC_UNI_PRCP/GAUGE_GLB/.

REFERENCES

Baggett, C., K. Nardi, S. Childs, S. Zito, E. Barnes, and E. Maloney, 2018: Skillful subseasonal forecasts of weekly tornado and hail activity using the Madden–Julian Oscillation. *J. Geophys. Res. Atmos.*, **123**, 12 661–12 675, https://doi.org/10.1029/2018JD029059.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Chen, M., W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. Wayne Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.*, **113**, D04110, https://doi.org/10.1029/2007JD009132.

Cohen, J., D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, 2019: S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdiscip. Rev.: Climate Change*, **10**, e00567, https://doi.org/10.1002/wcc.567.

DelSole, T., L. Trenary, M. K. Tippett, and K. Pegion, 2017: Predictability of week-3–4 average temperature and precipitation over the contiguous United States. *J. Climate*, **30**, 3499–3512, https://doi.org/10.1175/JCLI-D-16-0567.1.

Fan, Y., and H. van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, **113**, D01103, https://doi.org/10.1029/2007JD008470.

——, and ——, 2011: Bias correction and forecast skill of NCEP GFS ensemble week-1 and week-2 precipitation, 2-m surface air temperature and soil moisture forecasts. *Wea. Forecasting*, **26**, 355–370, https://doi.org/10.1175/WAF-D-10-05028.1.

——, K. Gilbert, D. Rudack, W. Yan, S. Scallion, and P. Shafer, 2015: The characteristics of GFS MOS temperature forecast guidance errors for the past decade. *Special Symp. on Model Postprocessing and Downscaling*, Phoenix, AZ, Amer. Meteor. Soc., 4.2, https://ams.confex.com/ams/95Annual/webprogram/Paper265326.html.

——, C.-Y. Wu, J. Gottschalck, and V. Krasnopolsky, 2019: Improve CFS Week 3-4 precipitation and 2 meter air temperature forecasts with neural network techniques. *43rd NOAA Annual Climate Diagnostics and Prediction Workshop*, Santa Barbara, CA, NOAA/National Weather Service, 59–63, https://www.nws.noaa.gov/ost/climate/STIP/43CDPW/43cdpw-YFan.pdf.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

——, K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009: The gridding of MOS. *Wea. Forecasting*, **24**, 520–529, https://doi.org/10.1175/2008WAF2007080.1.

Gneiting, T., 2014: Calibration of medium-range weather forecast. ECMWF Tech. Memo. 719, 30 pp., https://www.ecmwf.int/en/elibrary/74623-calibration-medium-range-weather-forecasts.

Hornik, K., 1991: Approximation capabilities of multilayer feedforward network. *Neural Networks*, **4**, 251–257, https://doi.org/10.1016/0893-6080(91)90009-T.

——, 1993: Some new results on neural network approximation. *Neural Networks*, **6**, 1069–1072, https://doi.org/10.1016/S0893-6080(09)80018-X.

Jenney, A., K. Nardi, E. Barnes, and D. Randall, 2019: The seasonality and regionality of MJO impacts on North American temperature. *Geophys. Res. Lett.*, **46**, 9193–9202, https://doi.org/10.1029/2019GL083950.

Johnson, N. C., D. C. Collins, S. B. Feldstein, M. L. L'Heureux, and E. E. Riddle, 2014: Skillful wintertime North American temperature forecasts out to 4 weeks based on the state of ENSO and the MJO. *Wea. Forecasting*, **29**, 23–38, https://doi.org/10.1175/WAF-D-13-00102.1.

Klein, W. H., and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.*, **55**, 1217–1227, https://doi.org/10.1175/1520-0477(1974)055<1217:FLWBMO>2.0.CO;2.

Krasnopolsky, V., 2007: Reducing uncertainties in neural network Jacobians and improving accuracy of neural network emulations with NN ensemble approaches. *Neural Networks*, **20**, 454–461, https://doi.org/10.1016/j.neunet.2007.04.008.

——, 2013: *The Application of Neural Networks in the Earth System Sciences: Neural Network Emulations for Complex Multidimensional Mappings*. Springer, 200 pp.

——, and Y. Lin, 2012: A neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental U.S. *Adv. Meteor.*, **2012**, 649450, https://doi.org/10.1155/2012/649450.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444, https://doi.org/10.1038/nature14539.

Liu, Y., E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, 2016: Application of deep convolutional neural networks for detecting extreme weather in climate datasets. arXiv, 1605.01156v1, https://arxiv.org/abs/1605.01156.

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.

Mundhenk, B., E. Barnes, E. Maloney, and C. Baggett, 2018: Skillful empirical subseasonal prediction of landfalling atmospheric river activity using the Madden–Julian Oscillation and quasi-biennial oscillation. *Nat. Climate Atmos. Sci.*, **1**, 20177, https://doi.org/10.1038/s41612-017-0008-2.

Nardi, K., E. Barnes, and F. 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., G. Villarini, and D. 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.

Nguyen, D., and B. Widrow, 1990: Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. *1990 IJCNN Int. Joint Conf. on Neural Networks*, San Diego, CA, Institute of Electrical and Electronics Engineers, 21–26, https://doi.org/10.1109/IJCNN.1990.137819.

Pan, B., K. Hsu, A. AghaKouchak, S. Sorooshian, and W. Higgins, 2019: Precipitation prediction skill for the west coast United States: From short to extended range. *J. Climate*, **32**, 161–182, https://doi.org/10.1175/JCLI-D-18-0355.1.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

Rasp, S., and S. Lerch, 2018: Neural network for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate*, **19**, 3483–3517, https://doi.org/10.1175/JCLI3812.1.

——, and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Schmidhuber, J., 2015: Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003.

Sharma, S., and Coauthors, 2017: Eastern U.S. verification of ensemble precipitation forecasts. *Wea. Forecasting*, **32**, 117–139, https://doi.org/10.1175/WAF-D-16-0094.1.

Totz, S., E. Tziperman, D. Coumou, K. Pfeiffer, and J. Cohen, 2017: Winter precipitation forecast in the European and Mediterranean regions using cluster analysis. *Geophys. Res. Lett.*, **44**, 12 418–12 426, https://doi.org/10.1002/2017GL075674.

Vigaud, N., A. Robertson, and M. Tippett, 2018: Predictability of recurrent weather regimes over North America during winter from submonthly reforecasts. *Mon. Wea. Rev.*, **146**, 2559–2577, https://doi.org/10.1175/MWR-D-18-0058.1.

Wang, L., and A. Robertson, 2018: Week 3–4 predictability over the United States assessed from two operational ensemble prediction systems. *Climate Dyn.*, **52**, 5861–5875, https://doi.org/10.1007/s00382-018-4484-9.

Warner, T. T., 2011: *Numerical Weather and Climate Prediction.* Cambridge University Press, 550 pp.

Wick, G., P. Neiman, F. Ralph, and T. 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.

Wilson, L. J., and M. Vallee, 2002: The Canadian Updateable Model Output Statistics (UMOS) system: Design and development tests. *Wea. Forecasting*, **17**, 206–222, https://doi.org/10.1175/1520-0434(2002)017<0206:TCUMOS>2.0.CO;2.

——, and ——, 2003: The Canadian Updateable Model Output Statistics (UMOS) system: Validation against perfect prog. *Wea. Forecasting*, **18**, 288–302, https://doi.org/10.1175/1520-0434(2003)018<0288:TCUMOS>2.0.CO;2.

Zhong, Q., J. Li, L. Zhang, R. Ding, and B. Li, 2018: Predictability of tropical cyclone intensity over the western North Pacific using the IBTrACS dataset. *Mon. Wea. Rev.*, **146**, 2741–2755, https://doi.org/10.1175/MWR-D-17-0301.1.