Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California

MICHAEL SCHEUERER, MATTHEW B. SWITANEK, AND ROCHELLE P. WORSNOP

Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, and NOAA/Physical Sciences Laboratory, Boulder, Colorado

THOMAS M. HAMILL

NOAA/Physical Sciences Laboratory, Boulder, Colorado

(Manuscript received 30 March 2020, in final form 6 June 2020)

ABSTRACT

Forecast skill of numerical weather prediction (NWP) models for precipitation accumulations over California is rather limited at subseasonal time scales, and the low signal-to-noise ratio makes it challenging to extract information that provides reliable probabilistic forecasts. A statistical postprocessing framework is proposed that uses an artificial neural network (ANN) to establish relationships between NWP ensemble forecast and gridded observed 7-day precipitation accumulations, and to model the increase or decrease of the probabilities for different precipitation categories relative to their climatological frequencies. Adding predictors with geographic information and location-specific normalization of forecast information permits the use of a single ANN for the entire forecast domain and thus reduces the risk of overfitting. In addition, a convolutional neural network (CNN) framework is proposed that extends the basic ANN and takes images of large-scale predictors as inputs that inform local increase or decrease of precipitation probabilities relative to climatology. Both methods are demonstrated with ECMWF ensemble reforecasts over California for lead times up to 4 weeks. They compare favorably with a state-of-the-art postprocessing technique developed for medium-range ensemble precipitation forecasts, and their forecast skill relative to climatology is positive everywhere within the domain. The magnitude of skill, however, is low for week-3 and week-4, and suggests that additional sources of predictability need to be explored.

1. Introduction

A number of recent publications have investigated the levels of subseasonal forecast skill of operational ensemble weather prediction systems for precipitation accumulations over the contiguous United States (DelSole and Trenary 2017; Wang and Robertson 2019; Vigaud et al. 2020). California has experienced prolonged droughts from 2011 to 2016 that placed great stress on the state's water resources. Productive cool-season atmospheric events can alleviate current and/or building drought conditions, but they also have the potential to yield extreme rainfall and even floods. Improving the skillfulness of subseasonal precipitation forecasts would be extremely valuable for reservoir management, as many decisions are made 1–4 weeks ahead. Unfortunately, forecast skill for precipitation accumulations over California drops off significantly after week 2, and the low signal-to-noise ratio makes it challenging for statistical algorithms to extract information from the ensembles and generate probabilistic forecasts (e.g., for "below-normal" or "above-normal" precipitation amounts) that are reliable and skillful (Vigaud et al. 2020, their Figs. 6 and 7).

In the context of short- to medium-range forecasting (i.e., up to 7 days in advance), a variety of parametric (Sloughter et al. 2007; Wilks 2009; Scheuerer and Hamill 2015), nonparametric (Hamill and Whitaker 2006; Gagne et al. 2014; Henzi et al. 2019), and semiparametric (Taillardat et al. 2019; Schlosser et al. 2019) statistical postprocessing approaches have been proposed to obtain reliable probabilistic guidance from

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-20-0096.s1.

Corresponding author: Michael Scheuerer, michael.scheuerer@noaa.gov

DOI: 10.1175/MWR-D-20-0096.1

ensemble precipitation forecasts. These methods typically require a sufficiently large ''reforecast'' (i.e., retrospective forecasts from the same model) dataset that can be compared to past observations to identify systematic model errors and establish statistical relationships between ensemble forecasts and observations. Even larger training samples are needed in a subseasonal forecasting context where the signal-to-noise ratio is low and hence the risk of overfitting a statistical model increases. To reduce the overall number of model parameters, Stauffer et al. (2017) standardize both forecast and observation data to remove seasonal and location-specific climatological characteristics and fit a single statistical model that relates the standardized forecasts and observations. We follow the same general strategy, but our proposed data normalization is more tailored to the mixed discrete-continuous nature of the distribution of precipitation accumulations. Observed precipitation amounts are discretized into a fixed number of categories defined by climatological quantiles, an artificial neural network (ANN) model is used to relate normalized predictors to the precipitation categories, and the ANN-based, real-time probability forecasts are interpolated to full predictive cumulative distribution functions (CDFs). The ANN framework eliminates the need for parametric assumptions about the predictive distribution and the predictor–predictand relationships. Our working hypothesis is that by providing geographic information as additional predictors, the ANN's ability to model nonlinear predictor interactions permits a spatially adaptive adjustment of the raw ensemble forecasts, while training data are used more efficiently than by postprocessing models that are fitted separately for each location. We also propose an extension of the basic ANN model to a convolutional neural network (CNN) model that can take images of large-scale predictors as inputs and use them to predict local precipitation accumulations.

In section 2 we provide more detail about the ensemble forecasts and gridded observation data used in this study. Section 3 describes the preprocessing of predictors and predictands, the proposed ANN/CNN-based postprocessing models, and the interpolation of class probabilities to full CDFs. Verification results and a discussion of possible extensions of the proposed methodology are provided in sections 4 and 5, respectively.

All computations were performed in Python (Python Software Foundation 2018), visualizations are based on the recommendations by Stauffer et al. (2015) using the colorspace package (https://github.com/retostauffer/python-colorspace). Our neural network routines are implemented using the Python libraries Keras (Chollet and others 2015) and TensorFlow (Abadi et al. 2016). Python code for reproducing the results presented in this paper is available online (https://github.com/mscheuerer/NeuralNetworkS2S).

2. Data used in this study

The neural network methods developed here are used to predict week-2, week-3, and week-4 precipitation accumulations over California during the 20 cold seasons (October–April) in the period from 1997/98 to 2016/17. The predictions are based on retrospective 11-member ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) Cycle 43r3, which were produced every Monday and Thursday between 13 July 2017 and 4 June 2018 with 0000 UTC initial conditions (ECMWF 2017). With 20 years of reforecasts and 61 initialization dates falling within the October–April period, our dataset comprises 1220 seven-day forecast periods for each lead time. Reforecasts for total precipitation, total column water (TCW), and 500-hPa geopotential height (Z500) were retrieved from the ECMWF MARS archive system at a horizontal resolution of 0.25° over the area between 134° and 113°W longitude and between 29° and 46°N latitude, for lead times up to 28 days. Since the precipitation dataset used for verification (see below) is only available in the form of 24-h accumulations starting at 1200 UTC, we define week-2 precipitation forecasts to be the predicted accumulations between 156- and 324-h lead time, week-3 precipitation forecasts the predicted accumulations between 324- and 492-h lead time, and week-4 precipitation forecasts the predicted accumulations between 492- and 660-h lead time.

The ECMWF reforecasts are calibrated and verified against the daily accumulated PRISM precipitation dataset (PRISM 2019), which was obtained for the years 1981–2017, upscaled to the 0.25° resolution of the forecasts using arithmetic averaging, and clipped to the bounds of California. This results in 663 grid points for which forecast are generated. Compared to other gridded precipitation datasets, PRISM uses observations from a relatively dense station dataset, which results in improved representation of precipitation, especially over mountainous and coastal areas of the western United States (Daly et al. 2008). Analyzed TCW and Z500 data required for the training of our CNN model was obtained from the ERA5 dataset (Hersbach et al. 2019) provided by the Copernicus Climate Change Service (C3S) (2017).

For verifying and comparing the performance of the different methods discussed in this article, we need to set

aside independent test data that is not used for training. To ensure that we have a large enough test dataset from which we can calculate verification statistics with minimal sampling variability, we use leave-one-season-out cross validation. That is, we set aside one of the 20 cool seasons for testing, train the models on the remaining 19 cool seasons, and cycle the 20 cool seasons in the dataset so that eventually every cool season is evaluated, while the corresponding models are trained on independent data. For selecting the hyper-parameters defining the specific neural network architecture, a validation dataset has to be split off the training dataset. Details about this procedure are provided at the beginning of section 4a.

3. Statistical postprocessing methodology

Before we describe the ANN framework used here for subseasonal probabilistic quantitative precipitation forecasting, we briefly review the parametric approach that will be used as a benchmark. This method has been proposed by Scheuerer and Hamill (2015) in the context of statistical postprocessing of medium-range ensemble precipitation forecasts and provides probabilistic forecasts in the form of censored, shifted gamma distributions (CSGDs). Comparisons performed by Scheuerer and Hamill (2015), Baran and Nemoda (2016), and Zhang et al. (2017) suggest that the CSGD method and variants of it compare favorably with other state of the art postprocessing approaches for precipitation amounts; therefore, we select it as a reference method in the present study.

a. *Nonhomogeneous regression based on censored shifted gamma distributions*

The version of the method proposed by Scheuerer and Hamill (2015) used here proceeds in three steps. First, a climatological CSGD is fitted to the observed precipitation amounts at each grid point and for each month, using data from a 61-day time window centered around the 15th day of this month. The parameters μ_{cl}, σ_{cl}, and δ_{cl} defining the mean, standard deviation, and shift of the monthly climatological CSGDs are then linearly interpolated to day-specific parameters and are used later in the regression equations for the predictive CSGD parameters.

Second, the raw ensemble precipitation forecasts are spatially smoothed using the same smoothing kernel as Scheuerer and Hamill (2015) with a neighborhood radius $r = 300\,\mathrm{km}$. The rationale behind smoothing is that medium-range precipitation forecasts are often subject to substantial displacement errors, and spreading out a forecast signal related to increased precipitation amounts over a larger area can mitigate the "double penalty" issue and improve forecast skill (Ben Bouallègue et al. 2013; Scheuerer 2014; Scheuerer and Hamill 2015). To keep the algorithm simple, we omit the quantile-mapping step performed by Scheuerer and Hamill (2015) before the smoothing (see online supplement A for a justification of that omission). The mean $\overline{f}$ of the smoothed ensemble forecasts is then calculated and divided by its average over all reforecast dates in the training dataset within a 61-day time window centered around the 15th of the month. This multiplicative standardization yields a dimensionless predictor $\overline{f}_{\mathrm{ano}}$ that can now be linked to the parameters μ and σ of the predictive CSGD.

This link is established in a third step where we define

$$\mu = \frac{\mu_{\mathrm{cl}}}{a_1}\log\{1 + [(\exp(a_1) - 1)(a_2 + a_3\overline{f}_{\mathrm{ano}})]\},$$
$$\sigma = a_4\sigma_{\mathrm{cl}}\sqrt{\frac{\mu}{\mu_{\mathrm{cl}}}}, \tag{1}$$

while the shift parameter $\delta = \delta_{\mathrm{cl}}$ is kept fixed. The regression parameters are chosen such that the continuous ranked probability score (CRPS) (Matheson and Winkler 1976) obtained by applying these regression equations to a training dataset is minimized. The regression equations in (1) are simpler than those used by Scheuerer and Hamill (2015): they involve only four parameters and a single predictor $\overline{f}_{\mathrm{ano}}$ summarizing information in the raw ensemble forecasts. Simplification is required to avoid overfitting the model in the present situation of poor signal-to-noise ratio of subseasonal forecasts and the smaller sample size of weekly (vs daily) accumulations. We felt that the POP_f (ensemble probability of precipitation) predictor could be omitted, as it is not as useful for the 7-day accumulations considered here than it was found to be for the 12-h accumulations studied by Scheuerer and Hamill (2015). A predictor that measures the ensemble spread could be useful, but it is particularly prone to overfitting and was therefore also omitted (see online supplement A for a justification of this omission). Finally, while separate parameters a_1, a_2, a_3, and a_4 are fitted for each forecast lead time and each grid point, the same parameters are used across the entire cool season considered here in order to increase the training sample size. This implies an assumption that the NWP forecast skill does not vary too much during the cool season, while including the parameters μ_{cl}, σ_{cl}, and δ_{cl} in the regression equations ensures that at least seasonal variations in the precipitation climatology are accounted for.

Hamill and Scheuerer (2018) employed modified regression equations and included additional predictors

that account for spatial variations to allow sharing regression parameters across different grid points, but they found that the CSGD method did not retain the good reliability and performance reported in other publications. Here, we instead use the simplified implementation of the CSGD method described above which otherwise follows the usual approach of fitting separate parameters at each grid point to avoid local biases. The desire to share parameters across different grid points, however, is one of our motivations to explore the use of ANNs which are more flexible than a nonhomogeneous regression framework like the CSGD method, and allow one to include predictors that can interact in a nonlinear way and thus account for spatial variations in the model parameters.

b. *Neural network–based framework for probabilistic quantitative precipitation forecasting*

The development of increasingly sophisticated ANNs is driven by applications such as speech recognition, visual object recognition and object detection (LeCun et al. 2015), but ANNs are also increasingly used in the context of weather prediction. Examples include prediction of tornadoes (Marzban and Stumpf 1996; Lakshmanan et al. 2005), damaging winds (Marzban and Stumpf 1998), hail (Marzban and Witt 2001; Gagne et al. 2019), snowfall (Ware et al. 2006; Roebber et al. 2007), synoptic-scale fronts (Lagerquist et al. 2019), and hurricane intensity (Cloud et al. 2019). Two recent publications propose ANN-based statistical postprocessing methods for continuous surface weather variables and obtain full predictive distributions in two different ways: 1) Rasp and Lerch (2018) still assume a parametric distribution family but use an ANN to establish (possibly nonlinear) predictor–predictand relationships. 2) Bremnes (2020) characterizes predictive distributions in the form of conditional quantile functions represented as a linear combination of Bernstein polynomials. The coefficients of these basis functions are linked to the predictors via an ANN, and this approach permits flexibility with regard to both shape of the predictive distribution and predictor–predictand relationships.

Statistical postprocessing of precipitation forecasts always comes with additional challenges. First, the point mass at zero makes it difficult to find suitable parametric distribution families. The CSGD introduced above has been demonstrated to be capable of representing uncertainty about precipitation amounts in both wet and dry weather situations, but important aspects of the distribution such as the associated probability of precipitation are controlled by a complex interaction of the three distribution parameters, which makes it difficult to share model parameters across grid points with different climatologies. A quantile-based modeling approach, on the contrary, would have to deal with the fact that a variable number of these quantiles may be zero. Here, we explore a different approach to approximating a full predictive distribution by partitioning the range of possible outcomes into $m + 1$ categories, predicting the probabilities for observed precipitation amounts to fall into each category, and interpolating the resulting probability vector to a full predictive distribution.

In contrast to Li et al. (2019) who propose a similar approach to probabilistic forecasting in a more general context, we use a partitioning scheme that takes the climatology at each day of the year (doy) and location into account. This way it is possible to share ANN model parameters across time and space while avoiding highly unequal frequencies of occurrence of each category across different seasons and different grid points. Denote by $B_i = [c_{i-1}, c_i]$ the bin defining the ith category for $i \in \{0, \ldots, m\}$. The first ("no precipitation") bin is defined by $c_{-1} = 0$ mm and $c_0 = 0.254$ mm (i.e., any 7-day precipitation accumulation smaller or equal to 0.01 in. is considered negligible). Denoting the climatological probability of B_0 by $p_{\mathrm{cl},0}$, we set

$$\alpha_{\mathrm{cl},i} := p_{\mathrm{cl},0} + (1 - p_{\mathrm{cl},0})\frac{i}{m}, \quad i = 0, \ldots, m,$$

and define c_i as the climatological quantile of level $\alpha_{\mathrm{cl},i}$ with $c_m := \infty$. With this definition, each category $i \in \{1, \ldots, m\}$ occurs with climatological probability $p_{\mathrm{cl},i} = (1 - p_{\mathrm{cl},0})/m$. A climatology is constructed by composing a sample of observed precipitation amounts for each doy and each grid point, using data from a 61-day time window centered around this doy and all years for which precipitation analyses are available (here: 1981–2017). The "no precipitation" probability $p_{\mathrm{cl},0}$ and the quantiles $c_1, \ldots, c_{m-1}$ are then calculated from this sample. In our case, the sample size is 2257, which permits a sufficiently robust calculation of these quantiles. An example of the partition boundaries at the grid points along a 39.375°N latitude transect for doy = 15 (15 January) is depicted in Fig. 1. Presumably due to rounding effects, our climatological samples contain a number of duplicate values, and as a result it occasionally happens that $c_i = c_{i+1}$ for at least one i. Since the boundaries $c_0, \ldots, c_{m-1}$ are included in both of the bins they are separating, all bins are still well defined, but the category assignment is ambiguous. If the categorical precipitation data are represented by a binary indicator matrix with one column for each category and a value of 1 if the precipitation accumulation is in that category and 0 otherwise, the ambiguous assignment translates into multiple cases of 1 in each row.

To normalize the meteorological predictors (here: precipitation forecasts from K different ensemble members) we calculate the extreme forecast index (EFI) in the revised

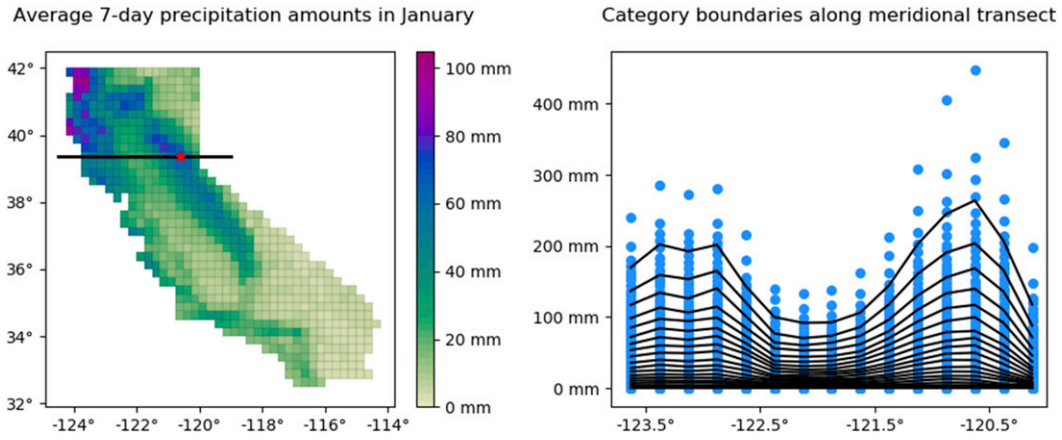

FIG. 1. (left) Climatological mean of 7-day precipitation amounts in January and meridional transect at 39.375°N latitude with the grid point in the Tahoe National Forest studied later marked by a red asterisk. (right) Boundaries for a partition into 20 categories, based on the respective climatological distributions for all grid points along that transect, with percentiles depicted as blue dots.

form proposed by Zsoter (2006). First, we spatially smooth each ensemble member in the same way as described in section 3a to reduce the impact of displacement errors in the forecast fields. Then, we compose a sample of these smoothed precipitation forecasts (pooling across all ensemble members) for each doy and each grid point, using data from a 61-day time window centered around this doy and all years except the year left out for verification. This sample represents the climatological reference distribution for the calculation of the EFI (see section A for details), which quantifies the departure of the ensemble forecast distribution from the model climatology. The EFI takes values in $[-1, 1]$, and its interpretation is independent from the climatology at the respective doy and grid point, which makes it a perfect statistic to summarize ensemble information for use as a meteorological predictor in our ANN model. Longitude and latitude of each grid point of the observation dataset, normalized to $[-1, 1]$, are provided as additional geographic predictors in order to allow the postprocessing model to adapt to possible variations of skill across the domain.

The ANN architecture used to link these predictors with observed precipitation categories is rather basic, except for one important modification discussed below. The input layer is fully connected to a (single) hidden layer with 10 nodes, and this hidden layer is fully connected to a preliminary output layer. For both layers we use exponential linear units (ELUs, Clevert et al. 2015) with $\alpha = 1$ as activation functions and L1 regularization with regularization parameter determined as described in section 4a. ELUs were preferred over rectified linear units (ReLUs) because they have similar properties but lead to smoother partial dependence curves. Experiments with L2 and dropout regularization yielded very similar results in this basic setup (not shown). Possible benefits of a larger number of nodes or more hidden layers are studied in section 4a.

Let $x_0, \ldots, x_m$ be the preliminary output from the network described above. A softmax activation function applied to these values directly would result in forecast probabilities:

$$p_i = \frac{\exp(x_i)}{\sum_{j=0}^{m} \exp(x_j)}, \quad i = 0, \ldots, m.$$

With increasing forecast lead time the predictors become less and less informative, and the best a postprocessing model can do in this case is revert to the climatological probabilities $p_{\mathrm{cl},0}, \ldots, p_{\mathrm{cl},m}$ at each doy and each grid point. We can help it in doing so by providing the logarithms of these probabilities as additional predictors which bypass the hidden layer(s) (and any regularization), and are added to the preliminary output values $x_0, \ldots, x_m$. That is, we apply the final softmax activation function to the values:

$$z_i = x_i + \log(p_{\mathrm{cl},i}), \quad i = 0, \ldots, m \tag{2}$$

to obtain

$$p_i = \frac{\exp(x_i) p_{\mathrm{cl},i}}{\sum_{j=0}^{m} \exp(x_j) p_{\mathrm{cl},j}}, \quad i = 0, \ldots, m.$$

The last equation shows that the values $x_0, \ldots, x_m$ can be understood as (logarithms of) multiplicative anomalies from the climatological probabilities. It also shows that

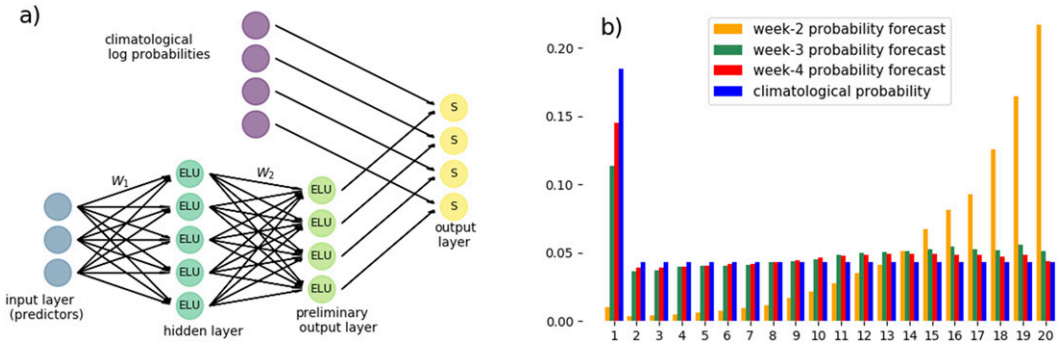

FIG. 2. (a) A schematic of the ANN architecture used in this study. For improved readability the ANN is depicted with 5 (instead of 10) nodes in the hidden layer, and 4 (instead of 20) categories for which forecast probabilities are calculated. (b) The categorical forecast probabilities for 8–14 Jan 2017 precipitation accumulations at a grid point in the Tahoe National Forest.

in the case of no skill, the ANN only needs to revert to preliminary outputs equal to zero in order for the final output vector to be equal to $(p_{\mathrm{cl},0},\ldots,p_{\mathrm{cl},m})$. Figure 2 illustrates the ANN architecture and the use of the additional climatology predictors. It also shows how in a case study of an atmospheric river event (accumulation period: 8–14 January 2017) the forecast probabilities at a grid point in the Tahoe National Forest converge to the climatological probabilities with increasing forecast lead time. Results of additional experiments that demonstrate the benefit of this use of climatological information are reported in the online supplement B.

A special kind of loss function is required to deal with the possibly ambiguous category assignments that can occur with the gridded precipitation observations as mentioned above. In appendix B, we propose a modified categorical cross-entropy score (MCCES) that can handle this type of ''censored'' observation data in those (few) cases where the category assignment is indeed ambiguous, while it reduces to the standard categorical cross-entropy loss in the cases where the observed precipitation amount falls into one definite category. To avoid introducing additional stochasticity into the optimization (keeping the low signal-to-noise ratio in our setup in mind), we decided to estimate the ANN parameters via batch gradient descent (i.e., the model is only updated after all training examples have been evaluated) using the Adam optimization method (Kingma and Ba 2014) with a learning rate of 0.05. This parameter was found to ensure stable convergence, and optimization was stopped after 100 epochs since additional iterations did not improve the training scores.

We finally describe how a forecast probability vector $\mathbf{p} = (p_0,\ldots,p_m)$ can be interpolated to a full predictive cumulative distribution function F. The value of F at each category boundary c_i can be calculated as the sum of the first i components of $\mathbf{p}$. For the forecast example from Fig. 2b these cumulative probabilities are depicted in Fig. 3a. Now consider the *cumulative hazard function:*

$$H(x) = -\log[1 - F(x)].$$

For the 7-day precipitation accumulations studied here and larger indices i, we find that the points $[c_i, H(c_i)]$ tend to be aligned along a straight line (see Fig. 3b). A perfectly linear function H would imply that the

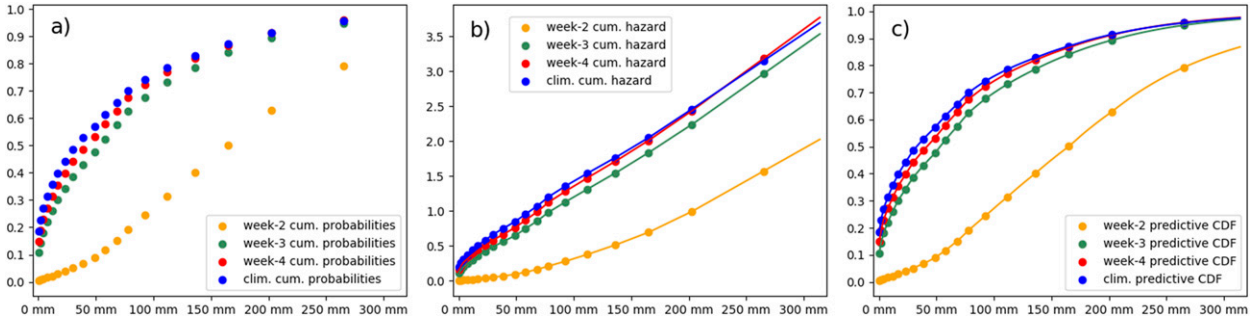

FIG. 3. (a) Cumulative probabilities for the forecast example from Fig. 2b), (b) corresponding cumulative hazards, and (c) predictive CDFs obtained by back-transforming the linearly interpolated cumulative hazards.

predictive distribution is exponential. Approximating H by a piecewise linear function that interpolates the points $[c_0, H(c_0)], \ldots, [c_{m-1}, H(c_{m-1})]$ and extrapolates linearly beyond c_{m-1} is equivalent to approximating the predictive CDF by a piecewise exponential function, and seems to provide a good reconstruction of F as shown in Fig. 3c. From this predictive CDF any probabilistic forecast of interest (e.g., exceedance probabilities for fixed or arbitrary climatology-based thresholds) can be derived.

c. *Neural network–based prediction using NWP forecasts of large-scale predictors*

For longer lead times, we can expect timing and displacement errors to become larger and larger, and the different ensemble members to diverge more and more. At some point it may be more useful to consider large-scale predictors that provide information about the general atmospheric state over the area of interest instead of forecasts of specific precipitation amounts. Here, we consider forecasts of geopotential heights at 500 hPa (Z500) and total column water (TCW) over an area between 134° and 113°W longitude and between 29° and 46°N latitude, averaged over the 7-day forecast period and upscaled to a 1° resolution. From these forecasts, we hope to obtain information about the direction and strength of the atmospheric flow, and about the available amount of precipitable water, respectively. The two large-scale predictors are normalized in different ways. For TCW, we are mostly interested in the anomalies from climatology (''moist'' or ''dry''), and we therefore normalize separately for each forecast grid point. We first calculate the 10th and the 90th percentiles across all years for which we have data, and then approximate the climatology at this grid point for each doy by fitting a simple, harmonic regression curve (a constant, a sine, and a cosine term) to the respective percentile values during the doys in the cool season period considered here. The TCW data are then normalized such that the 10th percentile is mapped to -1 and the 90th percentile is mapped to 1. For the Z500 predictor we expect that gradient information may be more important than the actual values, and therefore a local normalization is less appropriate as it removes the climatological (typically north–south) gradient from the fields. Instead, we normalize with respect to the 1st and 99th climatological percentile, smoothed over the cool season as for TCW, but calculated across all forecast grid points rather than for each grid point individually.

In contrast to section 3b, the predictor is now an image with 22×18 pixels and two channels, and this predictor is no longer specific to each observation grid point but meant to provide forecast information for the entire domain $\mathscr{D}$. A number of changes to the ANN framework proposed in section 3b are made to account for these differences. First, we note that *convolutional neural networks* (CNNs) are the perfect tool to deal with a predictor of that type (i.e., an image with multiple channels). Therefore, the input layer in Fig. 2a) is replaced by a sequence of 2D convolutional layers and max pooling layers. Specifically, the two-channel image is run through

- a 3×3 convolutional layer with four filters, no padding, and ELU activation.
- a 2×2 max pooling layer.
- another 3×3 convolutional layer with eight filters, no padding, and ELU activation.
- another 2×2 max pooling layer.

The output of the second max pooling layer is then flattened and used as input to the hidden layer in Fig. 2a).

The preliminary output layer in Fig. 2a) also needs to be modified. In the ANN framework described in section 3b, each grid point of the observation dataset is a separate case, and the preliminary output $x_0, \ldots, x_m$ determines the probability forecast for each of the $m + 1$ categories at that grid point. Now, the predictor image needs to inform the probability forecasts at all observation grid points simultaneously. This can be done by defining local, spatially smooth basis functions $\phi_1, \ldots, \phi_l$, and changing the preliminary output layer such that it returns a tensor $(\tilde{x}_{j,i})_{j=1,\ldots,l;i=0,\ldots,m}$. The dot product with the tensor defined by evaluating each basis function at all locations $s \in \mathscr{D}$ yields a new tensor

$$x_{s,i} = \sum_{j=1}^{l} \tilde{x}_{j,i} \phi_j(s), \tag{3}$$

which represents the anomalies from the climatological probabilities at each location $s \in \mathscr{D}$ and for each precipitation category $i \in \{0, \ldots, m\}$. These anomalies are then used in the same way as in (2) (i.e., they are added to the logarithm of the climatological probabilities at each grid point and turned into probability forecasts by using a softmax activation function). Figure 4 depicts the $l = 5$ basis functions used in the present study, details about their construction are provided in appendix C. A schematic of the complete CNN model is shown in Fig. 5a).

The computational details for this CNN-based postprocessing model are similar to those in section 3b. The MCCES (see appendix B) is again used as a loss function to be minimized by the Adam optimization method which is now run with a learning rate of 0.01 and stopped after 150 epochs. Unlike for the basic ANN in section 3b,

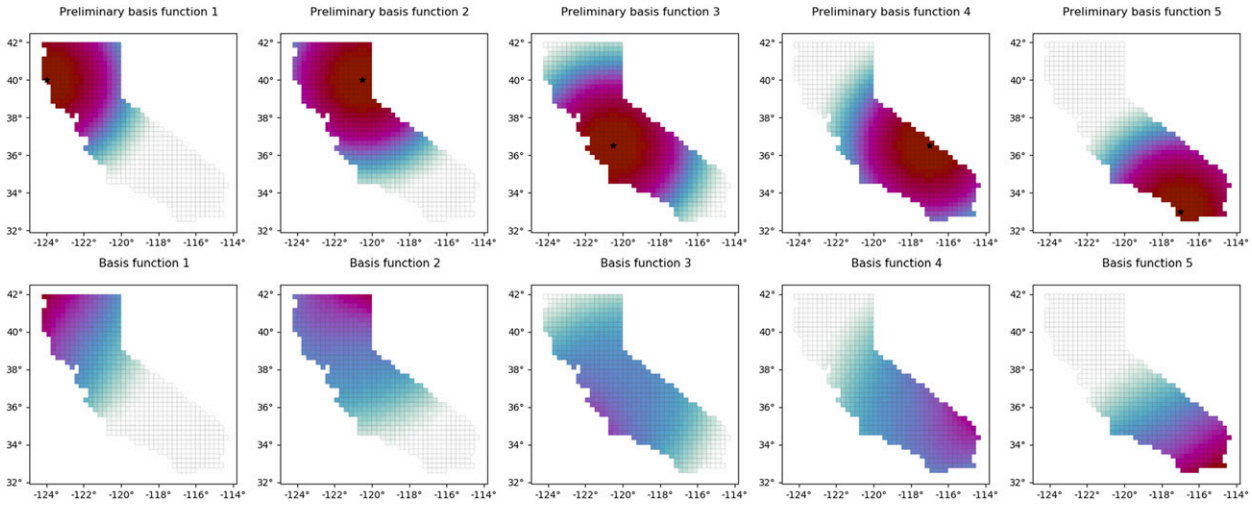

FIG. 4. (bottom) Basis functions used to transform the output from the hidden layer in our CNN into multiplicative probability anomalies over California and (top) preliminary radial basis functions from which the final basis functions were obtained through normalization. For details see appendix C.

dropout regularization (Srivastava et al. 2014) was found to be superior to L1 and L2 regularization for fitting the CNN model proposed here. Dropout is applied before the input of each hidden layer with a dropout rate determined as described in section 4a.

A major change is made in the application of the CNN framework compared to the basic ANN framework. Since the learning task—identifying features in a two-channel image of Z500 and TCW that can be linked to precipitation amounts over California—is far more sophisticated than that of linking forecast to observed precipitation amounts, the low signal-to-noise ratio at longer lead times may make it very hard to learn useful filters in the convolutional layers. While we can hope that some of the forecast errors for Z500 and TCW average out when working with the ensemble mean, averaging over ensemble members may dilute interactions between those two predictor variables, and gradients of the different Z500 forecast fields may partly cancel out. We therefore take a different approach and apply the above CNN framework in two steps. In the first step, the CNN is used to establish a link between observed precipitation amounts and 7-day averages of *analyzed* Z500 and TCW, for which we can expect the link with precipitation accumulations to be much more clear-cut than for extended-range forecasts of these quantities. The week-2/-3/-4 ensemble forecasts are then normalized in the same way as the analyses (but with respect to the IFS model climatology), and each member k is run separately through the CNN fitted to the analyzed Z500 and TCW data. The preliminary output values $x_{s,i}^{(k)}$, $k = 1, \ldots, 11$, for category i and location $s \in \mathscr{D}$ obtained in this way can be viewed as an ensemble of multiplicative *perfect prog* probability anomalies. They can be averaged to

$$\overline{x}_{s,i} = \frac{1}{11}\sum_{k=1}^{11} x_{s,i}^{(k)}, \quad s \in \mathscr{D}, \quad i = 0, \ldots, m,$$

added to the climatological log-probabilities as in (2), and turned into probability forecasts via softmax activation functions. By normalizing forecasts and analyses with respect to their respective climatologies, we implicitly remove systematic biases that the forecasts may have. However, this perfect prog approach does not account for insufficient ensemble spread, and the resulting probabilities may be overconfident. In the second step, we therefore allow a relaxation of the ensemble mean perfect prog probability anomalies $\overline{x}_{s,i}$ toward climatology via a relaxation factor η that extends (2) to

$$z_{s,i}(\eta) = \eta\overline{x}_{s,i} + \log(p_{\mathrm{cl},i}), \quad i = 0, \ldots, m. \tag{4}$$

A reasonable parameter range for η is [0, 1] where any value $\eta < 1$ reduces sharpness (and thus potential overconfidence) of the probability forecasts and $\eta = 0$ entails a complete relaxation to climatology (which is called for in the case where the ensemble forecasts have no skill). A schematic illustrating the complete procedure for generating probability forecasts for each category based on IFS ensemble forecasts of Z500 and TCW is depicted in Fig. 5b). The optimal η can be found by minimizing the same MCCES used to fit the neural network parameters. The training datasets used for this optimization are composed of the same dates that were

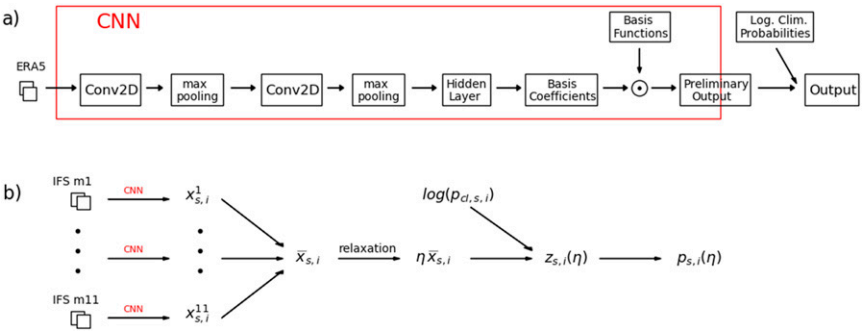

FIG. 5. (a) Schematic of the CNN model trained on ERA5 data. (b) Illustration of how, in a second step, the fitted CNN model is applied to IFS ensemble forecasts and how the preliminary output obtained separately for each member is aggregated to a single probability forecast vector.

used to train the CNN model. We assume η to be constant across the domain but specific to each forecast lead time. In our study, typical values were $\eta \approx 0.63$ for week 2, $\eta \approx 0.29$ for week 3, and $\eta \approx 0.23$ for week 4.

4. Verification of the probabilistic forecasts

a. Choice of hyperparameters

We first discuss the choice of several tuning parameters of the ANN/CNN frameworks proposed in section 3, which have either not been specified yet or need further justification in the form of a sensitivity analysis. If the goal is to find the best possible neural network configuration and hyperparameters, one could use one of the automated procedures that have been proposed in the literature (e.g., Hutter et al. 2011; Bergstra and Bengio 2012). Here, we take a different approach, starting with the relatively simple neural network architectures proposed in section 3, and providing a more detailed analysis of the potential benefits of more complex architectures by comparing with two specific, more complex alternatives. This may not lead to the overall best possible configuration, but we believe this approach yields more insights into the effects of different modeling choices. To make these choices, a three-way split of the dataset into a training, a validation, and a test dataset is required. For the calculations made within this subsection we therefore perform a 5-fold cross-validation inside the leave-one-season-out cross validation that already set aside one year for testing. The remaining data are split up into five disjoint, equally large time periods, and neural network models with different choices of tuning parameters are trained on 4 of these 5 folds and evaluated on the remaining fold. The validation fold is cycled so that we end up with five different validation scores for each year of the outer cross-validation loop. Depicting these five values separately gives us some idea about the variability that is due to 1) the stochasticity in the neural network model fitting and 2) possibly different predictability during the 5 cross-validation folds. We can compare that variability with the differences between, for example, different neural network architectures and check if these systematic differences are large enough compared to the random differences between the 5 folds to make a case for a more complex model.

This type of approach is first applied to compare the architecture of the ANN model from section 3b (one hidden layer with 10 nodes) with two alternative, more complex models:

- one hidden layer with 20 nodes
- two hidden layers with 10 nodes each

It is also used to check whether the number $m + 1$ of categories into which the observed precipitation amounts are discretized has an impact on the forecast quality. We make $m = 19$ our default choice and compare that against a coarser ($m = 9$) and a finer ($m = 29$) discretization. The advantage of the former is that fewer output nodes (and thus fewer parameters) are required, while the latter results in less information loss due to the discretization and permits more degrees of freedom in the predictive CDF that is reconstructed from the predicted category probabilities. The MCCES is not suitable for this comparison because larger values of m systematically entail a larger loss, so the comparison would be biased in favor of a smaller number of categories. We therefore use the CRPS as a loss function for this particular comparison. The CRPS can be calculated from the predictive CDFs via numerical integration and does not favor a particular choice of m.

Another important hyperparameter in the basic ANN framework from section 3b is the parameter λ that

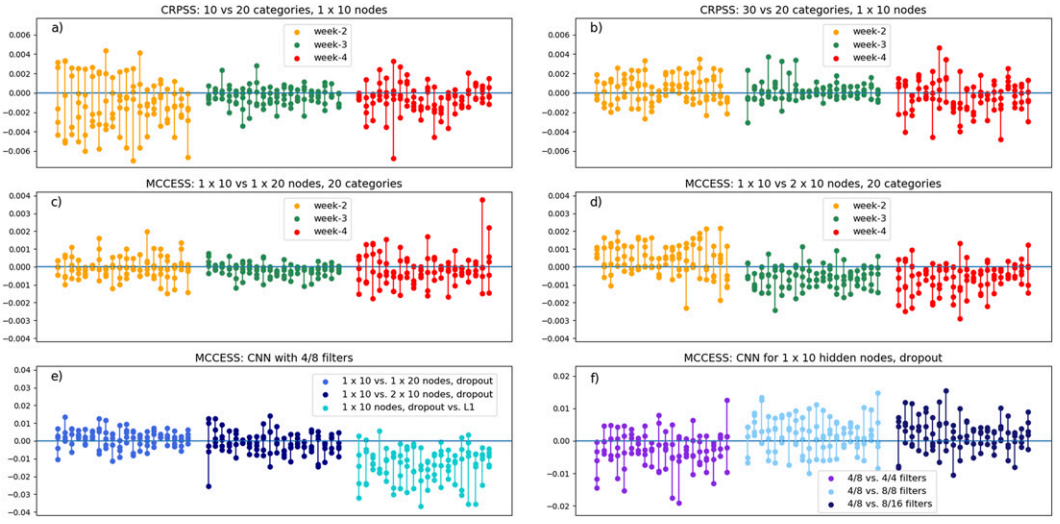

FIG. 6. (top) CRPSS for 10 or 30 categories relative to the default choice of 20 categories, (middle) MCCESS of more complex ANN architectures relative to the default architecture, and (bottom) MCCESS of different variations of the default CNN architecture. Each of the vertical lines corresponds to one cross-validation test year. The dots along this line are the validation scores of the 5 cross-validation folds associated with a single test year.

controls the strength of the L1 regularization. We test the choices $\lambda \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ and select the value that yields the lowest MCCES averaged over the 5 validation folds, separately for each of the neural network architectures and each choice of m. The most common choice for the week-2 forecasts was $\lambda = 10^{-6}$, while for week 3 and week 4 larger values were more common, especially for the more complex neural network architectures.

Figures 6a,b show continuous ranked probability skill scores (CRPSSs) for 10 and 30 categories (using the respective optimal λ) relative to the default choice of 20 categories. For almost every test year, the five different validation CRPSSs fall both below and above zero with no clear tendency toward a significant improvement or deterioration of skill. We conclude that in our setting the forecast results are not sensitive to the number of categories into which the gridded precipitation observations are discretized. A similar conclusion can be reached with regard to an increase of the number of nodes in the hidden layer of our default ANN architecture, which does not show systematic improvement of the modified categorical cross-entropy skill scores (MCCESSs) shown in Fig. 6c). Figure 6d) shows some tendency for improvement of week-2 forecast skill if a more complex architecture with two hidden layers is used, while that same architecture entails a deterioration of skill during week 3 and week 4. To keep things simple, we continue with our default model for all forecast lead times, but we note that if this ANN postprocessing framework were to be applied in the context of medium-range forecasting for which we usually see a better signal-to-noise ratio, the issue of optimal network architecture would have to be reassessed.

For the CNN framework proposed in section 3c the primary hyperparameter that needs to be selected is the dropout rate. Moreover, since the learning task is more complex and dropout reduces the effective number of nodes, we check again if more complex architectures yield better results than the proposed basic architecture with one hidden layer with 10 nodes. We also include the comparison that led us to the conclusion that dropout regularization is preferable to L1 regularization for the CNN framework, and a comparison of the forecast performance obtained with different configurations of filters in the two convolutional layers. Results of additional experiments with different configurations of the convolutional part of the CNN are reported in online supplement C. Since the CNN is trained with ERA5 data, tuning parameters and modeling choices are independent of forecast lead time. Dropout rates are selected separately for each architecture and each test year from the candidate set {0.1, 0.2, 0.3, 0.4, 0.5} as the minimizers of the average MCCES over the 5 validation folds. The optimal value was typically between 0.3 and 0.5 for the two architectures with a single hidden layer and between 0.1 and 0.3 for the architectures with two hidden layers (with dropout being applied before each of them). Figure 6e) suggests that the CNN architecture with two hidden layers does not yield systematically better probabilistic forecasts, and shows only a slight tendency toward improved forecasts with 20 nodes in

TABLE 1. RPSSs of probability forecasts at each analysis grid point, aggregated over the forecast domain and the 20 seasons. The "ERA5" column shows the RPSS of probability forecasts obtained by applying the fitted CNN models to (cross-validated) ERA5 data instead of IFS ensemble forecasts.

	ERA5	IFS ENS week 2	IFS ENS week 3	IFS ENS week 4
Raw ensemble	—	0.133	−0.088	−0.087
Bias-corrected ensemble	—	0.157	−0.054	−0.066
CSGD	—	0.190	0.025	0.003
ANN	—	**0.207**	**0.036**	0.010
ANN, EFI only	—	0.200	0.034	0.010
CNN	0.419	0.180	0.034	**0.016**
CNN, Z500 only	0.347	0.155	0.025	0.012
CNN, TCW only	0.307	0.147	0.030	0.012

the single hidden layer. Since the signal-to-noise ratio is less favorable when IFS ensemble forecasts of Z500 and TCW are used, we decided to continue with the simpler architecture with 10 nodes. Regarding regularization, however, we see much clearer evidence that dropout yields better results than L1 regularization, so we make this our default choice in the CNN framework. Figure 6d) finally shows results for different numbers of filters in the convolutional layers. There is a slight tendency toward better results if more filters are used, but again we feel that the case for a more complex model is not strong enough given the added challenges when using the fitted CNN model with IFS ensemble forecasts. Therefore, we continue with our default choice of four filters in the first and eight filters in the second convolutional layer.

b. *Ranked probability skill scores*

We present verification results obtained by evaluating the probabilistic forecasts by the basic ANN method and the CNN-based approach over the cool season left out for testing and then aggregating over all 20 cross-validated seasons. In the subseasonal context evaluation typically focuses on "below normal" and "above normal" probabilities, but since our National Weather Service partner, NOAA's Climate Prediction Center (CPC), provides probabilities of exceeding the 85th climatological percentile in their hazard assessment tool (www.cpc.ncep.noaa.gov/products/predictions/threats/extremesTool.php) we include this threshold in our performance metric and study ranked probability scores (Epstein 1969; Murphy 1971) of predicted probabilities for exceeding the 33rd, 67th, and 85th climatological percentiles at each grid point of our observation dataset. Ranked probability skill scores (RPSSs) are obtained using climatological frequencies (calculated from observed precipitation amounts across all 20 seasons and the 15 forecast valid dates closest to the forecast date of interest) of threshold exceedance as a reference.

Table 1 shows RPSSs for the two neural network approaches, the variant of the CSGD method described in section 3a, raw ensemble probability forecasts, and bias-corrected ensemble probability forecasts obtained by calculating the relative frequencies of members exceeding percentiles of the IFS model climatology rather than observation-based percentiles. This simple bias correction improves the ensemble forecast skill somewhat, but more improvement is realized with the more sophisticated postprocessing methods. The most skillful week-2 forecasts are obtained with the basic ANN approach, which suggests that IFS week-2 precipitation forecasts provide valuable forecast information which machine learning methods can extract successfully. In week 4, on the contrary, the CNN approach fares best, which supports our initial hypothesis that IFS information about larger-scale weather patterns is more useful than its precipitation forecasts at this lead time. Unfortunately, while forecast skill is positive across the entire domain (see Fig. 7), its magnitude is rather small, so without additional improvements to the forecast system or successful use of additional predictors the value of these probabilistic week-4 forecasts is still limited. Applying the CNN approach to only Z500 fields or only TCW fields both decreases forecast performance at all lead times compared to the full model that uses both predictors. Whether, conversely, providing additional large-scale predictors can further improve forecast performance is a question that will be addressed in future research.

Let us now take a closer look at the differences in performance of the CSGD, the basic ANN, and the CNN approach. Figure 7 shows RPSS maps for each of these methods and the three different forecast lead times considered here. One-sided, paired t tests were performed, separately for each grid point, to test whether the improvements of ANN/CNN-based forecast skill over CSGD-based forecast skill are statistically significant. Due to the overlapping forecast periods, the samples of RPS differences are not independent, but

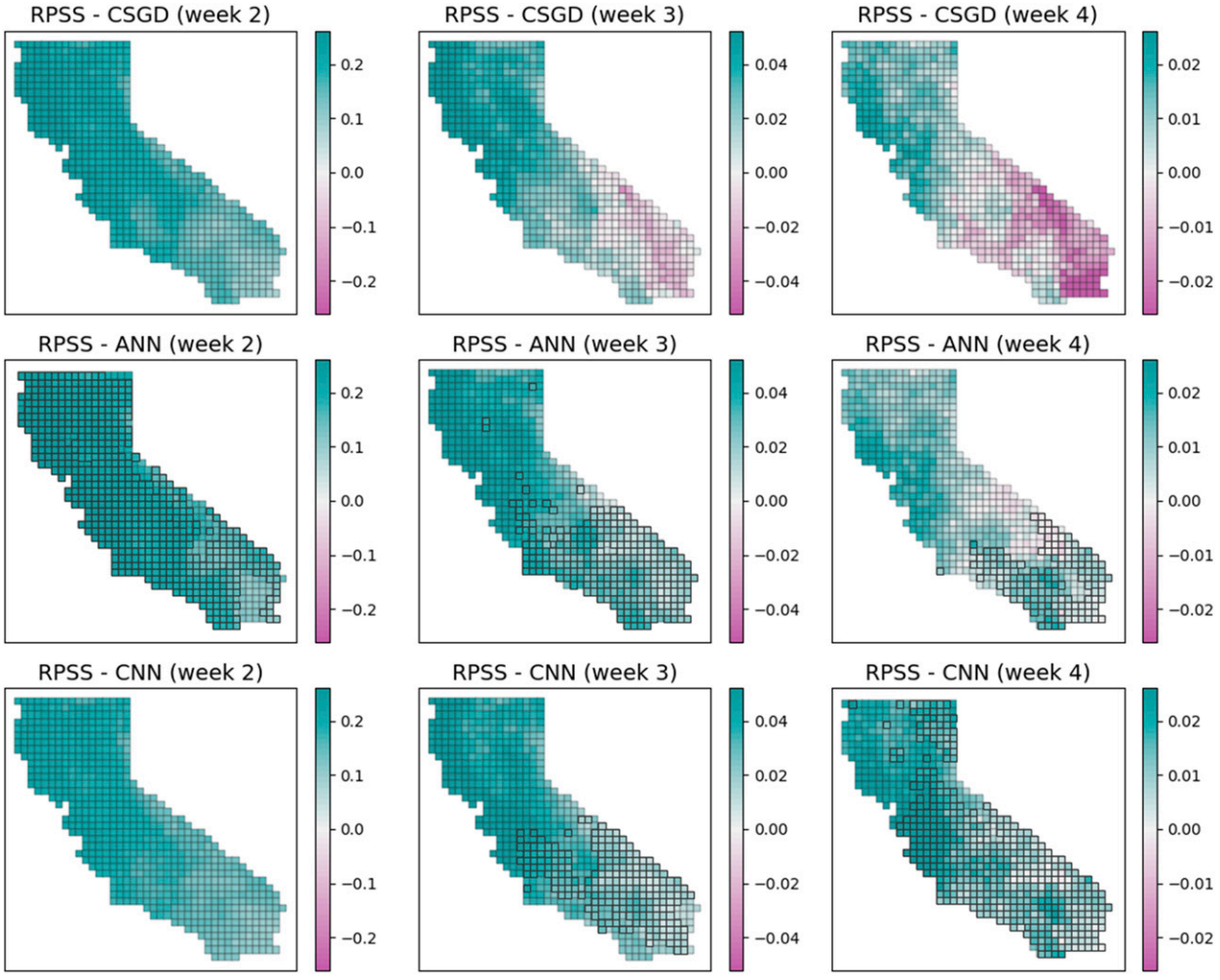

FIG. 7. RPSSs of probability forecasts for exceeding the 33rd, 67th, and 85th climatological percentiles at each analysis grid point, aggregated over the 20 seasons. For (middle) ANN and (bottom) CNN, grid boxes where the improvement over the (top) CSGD forecasts is statistically significant when controlling the false discovery rate at the level $\alpha_{\mathrm{FDR}} = 0.1$ are emphasized with thicker border lines.

we found that they can be approximated by first-order autoregressive processes. The variance of the sampling distribution, required in the denominator of the test statistic, can then be estimated as described in Eq. (2.15) of Jones (1975) to reflect the influence of the serial correlation of RPS differences. We account for test multiplicity (each grid point is tested individually) by controlling the false discovery rate (FDR) (Benjamini and Hochberg 1995) at the level $\alpha_{\mathrm{FDR}} = 0.1$ as described by Wilks (2016). The skill increase of ANN week-2 forecasts relative to CSGD forecasts is statistically significant at 86% of all grid points within California. By weeks 3 and 4, this number drops down to 42% and 22%, respectively. At these longer lead times, the improvement mainly occurs in areas where the CSGD forecasts have negative skill, which we hypothesize is due to overfitting. These areas happen to be the dryer ones where a larger fraction of the observed precipitation amounts is equal to zero, and thus fewer data pairs that are informative for model fitting are available. By removing climatological characteristics and fitting a single model that works for all grid points simultaneously, the ANN method avoids overfitting and skill never drops much below zero. The CNN approach performs worse than the CSGD method in week 2, comparable to the basic ANN method in week 3, and better than both of the other methods in week 4, where the skill improvement over the CSGD method is statistically significant at 68% of all grid points. Like the basic ANN method, the CNN method provides forecasts which at the very least do not fare worse than climatological forecasts, and in week 4 it can extract

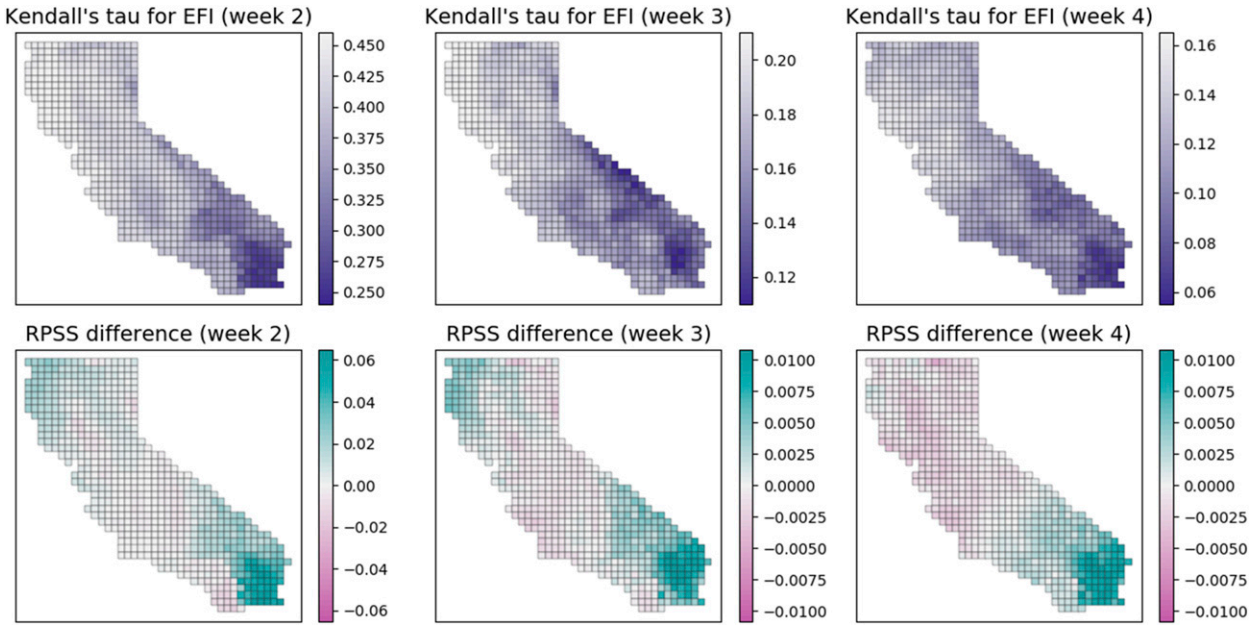

FIG. 8. (top) Kendall rank correlation coefficients between the EFI and the categorized precipitation observations at each grid point and (bottom) differences between the RPSS obtained with the full ANN model and the ANN model without the longitude and latitude predictors.

somewhat more forecast information from the IFS ensemble and attain better skill in Central and Northern California.

c. *Spatial adaptivity of the basic ANN and the CNN model*

Apart from the common idea of discretizing the predictand into climatology-dependent categories and providing the corresponding climatological frequencies as ''auxiliary predictors'' to the network, both neural network approaches use different strategies to permit a spatially varying response to the inputs.

The basic ANN model removes climatological characteristics from the IFS ensemble precipitation forecasts and provides geographic coordinates as additional predictors which allow the ANN model to account for spatial variations in forecast skill. Are these predictors useful? Table 1 includes results obtained with a variation of the basic ANN model in which the EFI of the IFS ensemble precipitation forecasts is used as the only predictor (i.e., the ANN can no longer use geographic information to modulate the influence of the EFI predictor on the local probability anomalies). As a result, the forecast skill decreases compared to the proposed implementation of the ANN model. To illustrate that this decrease in skill is indeed related to spatial variations of IFS skill, Fig. 8 depicts the Kendall rank correlation coefficients between the EFI and the categorized precipitation observations at each grid point. It also depicts the differences between the RPSS obtained with the full ANN model and the variant without the longitude and latitude predictors. For all forecast lead times, IFS skill is markedly lower over Southern California, and this region happens to be the one that benefits most from the use of geographic coordinates as additional predictor. Apparently, the ability to model nonlinear predictor interactions permits the ANN model to use this information to modulate the local response to a given EFI value. This allows us to use a single postprocessing model to generate spatially adaptive, probabilistic forecasts over the entire domain.

The CNN approach proposed in section 3c uses identical predictors (here: two-channel images) for each grid point of the observation dataset and models the influence of these predictor in different geographic areas through the use of local basis functions (see Fig. 4). To demonstrate that this indeed permits different local responses, Fig. 9 depicts the normalized ERA5-based Z500 and TCW fields associated with the lowest and highest predicted probabilities of exceeding the 85th climatological percentile at Eureka (Northern California) and San Diego (Southern California). The large-scale scenarios that minimize (maximize) the probabilities are different for the two locations with respect to both direction of atmospheric flow and area of the driest (wettest) TCW anomalies. Even though we made no attempt to optimize their number, positions, or radius

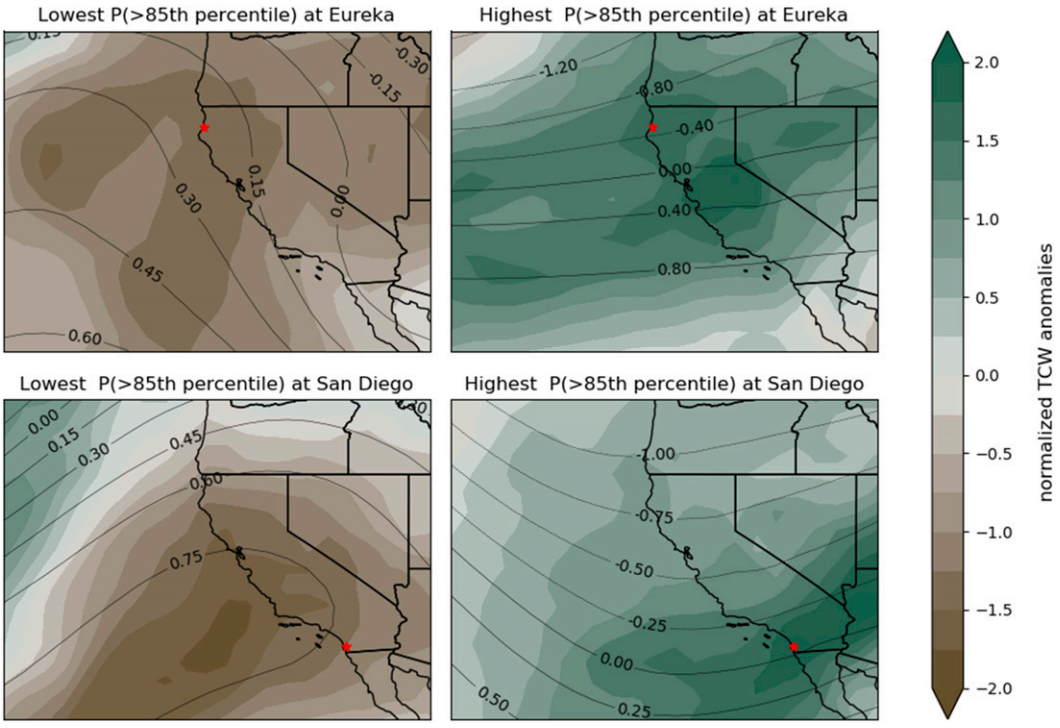

FIG. 9. Normalized Z500 (contours) and TCW fields associated with the (left) lowest (right) highest predicted probabilities of exceeding the 85th climatological percentile at (top) Eureka and (bottom) San Diego.

of influence, the general concept of local basis functions seems to be effective for modeling the spatially varying influence of large-scale predictors that come in the form of multichannel images.

5. Discussion

We have proposed a new ANN-based approach for generating probabilistic subseasonal precipitation forecasts based on the output of an ensemble weather prediction system. The low signal-to-noise ratio at subseasonal lead times requires highly efficient use of training data, and we achieve this by removing climatological characteristics from both forecast and observation data. This way, a single ANN model can generate probabilistic forecasts at every grid point within the forecast domain. Providing climatological information to the ANN helps ensure that the resulting probabilistic forecasts never perform worse than climatological forecasts. We also found that adding predictors with geographic information can account for spatially varying skill of the underlying NWP forecasts.

An extension of the basic ANN to a CNN has also been proposed and allows one to use a different type of predictor: instead of using NWP model output that is specific to each grid point, the CNN model can extract information from multichannel images, and is thus capable of making local predictions based on forecast large-scale atmospheric conditions. Both ANN and CNN frameworks were demonstrated over California with subseasonal forecasts from the IFS ensemble. The basic ANN outperformed a state-of-the-art postprocessing method developed for medium-range forecasting at all lead times, the CNN yielded the most skillful week-4 probabilistic forecasts.

Despite the improvements obtained with the proposed ANN/CNN methodology, forecast skill beyond week 2 is rather limited. The focus of this paper is to propose new postprocessing methodology that addresses the particular challenges with subseasonal probabilistic forecasting, but further efforts are required to explore additional sources of subseasonal predictability. More systematic investigations into selecting an optimal set of predictors could be performed. The ANN-based approach presented here uses only IFS precipitation forecasts and the CNN-based approach uses only forecasts of Z500 and TCW, but best results across all forecast lead times are likely achieved by using a combination of surface weather variables and large-scale variables as predictors. Moreover, there may well be sources of predictability that are not optimally used by NWP models. In the context of seasonal precipitation forecasting over

the western United States, Switanek et al. (2020) found that rather basic statistical models can extract forecast information from past sea surface temperature analyses that allows them to outperform state-of-the-art NWP models, and recent research by M. Switanek (2020, personal communication) suggests that some of the skill improvement carries over to subseasonal precipitation forecasts. It has also been established that the Madden–Julian oscillation (MJO) (Guan et al. 2012; Zhang 2013), especially in combination with the quasi-biennial oscillation (QBO) (Mundhenk et al. 2018; Zhang and Zhang 2018) influences winter precipitation over California, and that poor representation of the Arctic Oscillation in NWP models can limit their forecast skill (Singh et al. 2018). These indices would all be candidates that could be used as additional predictors in our ANN or CNN framework, hoping that their interaction with the other predictor variables improves forecast skill.

Another path forward is to look for ''forecasts of opportunity'' (i.e., to find ways for a priori identification of skillful subseasonal forecasts) (e.g., Albers and Newman 2019). This information could be provided to the ANN/CNN model as an additional predictor and allow it to deviate more strongly from climatology in situations of enhanced expected predictability. Somewhat better skill can also be obtained by scaling back expectations on temporal and spatial resolution. Combining week-3 and week-4 lead times to week-3–4 outlooks, for example, might make it easier to capture the time scales of rainfall relationships to shifts in the jet stream and storm tracks (Vigaud et al. 2020).

We finally note that the proposed ANN and CNN methodology, while developed with subseasonal precipitation forecasts in mind, may well be suitable for other weather variables and/or shorter forecast lead times. This may require adjustments to the interpolation scheme that reconstructs full predictive distributions based on the predicted category probabilities from the neural networks, and reassessment of the optimal neural network complexity.

Acknowledgments. We are grateful to Steve Penny, Sebastian Lerch, and two anonymous reviewers for valuable comments that helped improve this paper. Our research was supported by a grant from the California Department of Water Resources through federal Grant 4BM9NCA-P00.

APPENDIX A

Efficient Calculation of the Extreme Forecast Index (EFI)

Denote by F_{cl} the CDF of the climatological distribution at a fixed grid point and fixed day of the year. The revised EFI proposed by Zsoter (2006) is defined as the weighted integral over departures of the proportion of EPS members lying below the α quantile of F_{cl} from that level α:

$$\mathrm{EFI}(F_{\mathrm{cl}}, \mathbf{x}) = \frac{2}{\pi}\int_0^1 \frac{\alpha - \frac{1}{K}\sum_{k=1}^{K}\mathbf{1}\{x_k \le F_{\mathrm{cl}}^{-1}(\alpha)\}}{\sqrt{\alpha(1-\alpha)}}\,d\alpha, \tag{A1}$$

where $\mathbf{x} = (x_1, \ldots, x_K)$ denotes the ensemble and $\mathbf{1}$ denotes the indicator function that is equal to 1 if the condition in the bracket is fulfilled and 0 otherwise. Using the definition and properties of the beta function $B(\cdot, \cdot)$ and integration by substitution we can rewrite (A1) as

$$\begin{aligned}\mathrm{EFI}(F_{\mathrm{cl}}, \mathbf{x}) &= \frac{2}{\pi}\int_0^1 \frac{\sqrt{\alpha}}{\sqrt{1-\alpha}}\,d\alpha - \frac{2}{\pi K}\sum_{k=1}^{K}\int_0^1 \frac{\mathbf{1}\{F_{\mathrm{cl}}(x_k) \le \alpha\}}{\sqrt{\alpha(1-\alpha)}}\,d\alpha = \frac{2}{\pi}B\left(\frac{3}{2},\frac{1}{2}\right) - \frac{2}{\pi K}\sum_{k=1}^{K}\int_{F_{\mathrm{cl}}(x_k)}^1 \frac{1}{\sqrt{\alpha(1-\alpha)}}\,d\alpha \\ &= 1 - \frac{4}{\pi K}\sum_{k=1}^{K}\int_{\sqrt{F_{\mathrm{cl}}(x_k)}}^1 \frac{1}{\sqrt{1-\alpha^2}}\,d\alpha = -1 + \frac{4}{\pi K}\sum_{k=1}^{K}\arcsin\left[\sqrt{F_{\mathrm{cl}}(x_k)}\right] \\ &= -1 + \frac{2}{\pi K}\sum_{k=1}^{K}\arccos[1 - 2F_{\mathrm{cl}}(x_k)],\end{aligned}$$

so the calculation of $\mathrm{EFI}(F_{\mathrm{cl}}, \mathbf{x})$ reduces to a sum that involves probability integral transforms $F_{\mathrm{cl}}(x_1), \ldots, F_{\mathrm{cl}}(x_K)$ of the ensemble forecasts with respect to the climatological CDF. If the climatology is represented by a sample $\xi = (\xi_1, \ldots, \xi_L)$, F_{cl} can be approximated by an empirical CDF $\hat{F}_{\mathrm{cl}}$, and the probability integral transform of x_k is $\hat{F}_{\mathrm{cl}}(x_k) = (1/L)\sum_{l=1}^{L}\mathbf{1}\{x_k \le \xi_l\}$.

Here, we take a different approach and make the same assumption as for the forecast distributions in section 3b that F_{cl} can be approximated by a piecewise exponential distribution. We calculate the 19 quantiles with levels $\{0.05, 0.1, \ldots, 0.95\}$ from the sample ξ and approximate F_{cl} by linear interpolation of the associated cumulative hazard function as described in section 3b.

APPENDIX B

Modified Categorical Cross-Entropy Score (MCCES)

Denote by $\mathbf{p} = (p_0, \ldots, p_m)$ a probability vector representing the forecast probabilities for each of the $m + 1$ categories, and let $\mathbf{y} = (y_0, \ldots, y_m)$ be the corresponding binary observation vector. In the usual situation where the assignment to a category is unambiguous, there is exactly one component ι such that $y_\iota = 1$ and $y_i = 0$ for $i \neq \iota$. In this case the categorical cross-entropy loss function is defined as

$$\mathscr{L}(\mathbf{p}, \mathbf{y}) = -\sum_{i=0}^{m} y_i \log(p_i), \tag{B1}$$

and it follows from Gibb's inequality that $\mathscr{L}$ is a strictly proper scoring rule for multinomial forecast distributions.

Assume now that there is a censoring mechanism that prevents us from knowing the exact outcome and only reveals that for a given partition $A_0, \ldots, A_r$ of $\{0, \ldots, m\}$ and some index j, $\iota \in A_j$. We denote the partition by $\mathscr{A}$ and define a modified categorical cross-entropy loss function via

$$\mathscr{L}_{A_0,\ldots,A_r}(\mathbf{p}, \mathbf{y}) = -\sum_{j=0}^{r} y_{A_j} \log(p_{A_j}), \tag{B2}$$

where $y_{A_j} := \mathbf{1}\{\iota \in A_j, y_\iota = 1\}$ and $p_{A_j} := \sum_{i \in A_j} p_i$. We demonstrate that this is a proper scoring rule for outcomes distributed according to a multinomial distribution with probability vector $\mathbf{p}$. For the expected loss for a probability forecast vector $\mathbf{q}$ we obtain

$$\begin{aligned}\mathbf{E}_{\mathbf{p}}\left[\mathscr{L}_{A_1,\ldots,A_r}(\mathbf{q}, \mathbf{y})\right] &= -\sum_{\iota=0}^{m} p_\iota \sum_{j=0}^{r} \mathbf{1}\{\iota \in A_j\} \log(q_{A_j}) \\ &= -\sum_{i=0}^{r} \underbrace{\sum_{\iota \in A_i} p_\iota}_{=p_{A_i}} \sum_{j=0}^{r} \underbrace{\mathbf{1}\{\iota \in A_j\}}_{=0 \text{ for } j \neq i} \log(q_{A_j}) = -\sum_{i=0}^{r} p_{A_i} \log(q_{A_i}) \geq -\sum_{i=0}^{r} p_{A_i} \log(p_{A_i}),\end{aligned}$$

where the last inequality is Gibb's inequality applied to the probability vectors $(p_{A_0}, \ldots, p_{A_r})$ and $(q_{A_0}, \ldots, q_{A_r})$. The last term is equal to $\mathbf{E}_{\mathbf{p}}[\mathscr{L}_{A_0,\ldots,A_r}(\mathbf{p}, \mathbf{y})]$ and thus $\mathscr{L}_{A_0,\ldots,A_r}$ is proper. Is is strictly proper only in the trivial case where (after index permutation) $A_j = \{j\}, j = 0, \ldots, m$.

We note that $\mathscr{L}_{A_1,\ldots,A_r}$ can be written in a more convenient form if the incompletely known outcome is encoded in such a way that for some index j, $y_i = 1$ for all $i \in A_j$ and $y_i = 0$ for all $i \notin A_j$. This is how we encoded the possibly ambiguous category assignments in section 3b. The definition of the modified categorical cross-entropy loss function in (B2) is then equivalent to

$$\mathscr{L}_{A_0,\ldots,A_r}(\mathbf{p}, \mathbf{y}) = -\log\left(\sum_{i=0}^{m} y_i p_i\right). \tag{B3}$$

This form is easy to implement and computationally efficient, and in the situation where $y_\iota = 1$ for one single component ι it is equivalent to the standard categorical cross-entropy loss in (B1).

APPENDIX C

Construction of the Basis Functions in Fig. 4

The basis functions $\phi_1, \ldots, \phi_5$ depicted in Fig. 4 are constructed such as to induce a smooth function when linearly combined as in (3) while having a local support (i.e., they are zero beyond a certain radius ρ of influence), so that the associated coefficients $x_{1,i}, \ldots, x_{5,i}$ only affect the forecasts within a limited geographic area. We achieve that by first defining smooth, locally supported radial basis functions:

$$\vartheta_{\tilde{s}_j}(\cdot) := \left[1 - \left(\frac{\|\cdot - \tilde{s}_j\|}{\rho}\right)^3\right]_+^3, \tag{C1}$$

where $\|\cdot\|$ denotes the Euclidean norm (here applied to differences in geographic coordinates), $[\cdot]_+ = \max(\cdot, 0)$, and $\tilde{s}_1, \ldots, \tilde{s}_5$ are the center points of the radial basis functions. The center points are chosen as

$$\begin{aligned}&\tilde{s}_1 = (-124°, 40°), \quad \tilde{s}_2 = (-120.5°, 36.5°),\\ &\tilde{s}_3 = (-120.5°, 40°), \quad \tilde{s}_4 = (-117°, 33°),\\ &\tilde{s}_5 = (-117°, 36.5°),\end{aligned}$$

that is, they are 3.5° apart in both longitude and latitude direction and inside the forecast domain. The basis functions $\phi_1, \ldots, \phi_5$ are then obtained by normalizing these functions at each analysis grid point:

$$\phi_j(\cdot) := \frac{\vartheta_{\tilde{s}_j}(\cdot)}{\sum_{j'=1}^{5} \vartheta_{\tilde{s}_{j'}}(\cdot)}. \quad \text{(C2)}$$

This normalization ensures that a spatially constant function can be obtained in (3) if all coefficients $x_{1,i},\ldots,x_{5,i}$ are identical. Each function ϕ_j is equal to zero beyond a distance ρ from the associated center point. In this study we chose $\rho = 7°$.

REFERENCES

Abadi, M., and Coauthors, 2016: Tensorflow: A system for large-scale machine learning. *Proc. USENIX 12th Symp. on Operating Systems Design and Implementation*, Savannah, GA, Advanced Computing Systems Association, 265–283, https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Albers, J. R., and M. Newman, 2019: A priori identification of skillful extratropical subseasonal forecasts. *Geophys. Res. Lett.*, **46**, 12 527–12 536, https://doi.org/10.1029/2019GL085270.

Baran, S., and D. Nemoda, 2016: Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, **27**, 280–292, https://doi.org/10.1002/env.2391.

Ben Bouallègue, Z., S. E. Theis, and C. Gebhardt, 2013: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteor. Z.*, **22**, 49–59, https://doi.org/10.1127/0941-2948/2013/0374.

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

Bergstra, J., and Y. Bengio, 2012: Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, **13**, 281–305.

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.

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

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

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

Copernicus Climate Change Service (C3S), 2017: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), accessed 24 June 2019, https://cds.climate.copernicus.eu/cdsapp#!/home.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, https://doi.org/10.1002/joc.1688.

DelSole, T., and L. Trenary, 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.

ECMWF, 2017: Part V: Ensemble prediction system. ECMWF IFS Doc. 5, 23 pp., https://www.ecmwf.int/node/17737.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987, https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

——, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.*, **147**, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1.

Guan, B., D. E. Waliser, N. P. Molotch, E. J. Fetzer, and P. J. Neiman, 2012: Does the Madden–Julian Oscillation influence wintertime atmospheric rivers and snowpack in the Sierra Nevada? *Mon. Wea. Rev.*, **140**, 325–342, https://doi.org/10.1175/MWR-D-11-00087.1.

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

——, and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079–4098, https://doi.org/10.1175/MWR-D-18-0147.1.

Henzi, A., J. F. Ziegel, and T. Gneiting, 2019: Isotonic distributional regression. https://arxiv.org/abs/1909.03725.

Hersbach, H., and Coauthors, 2019: Global reanalysis: Goodbye ERA-Interim, hello ERA5. *ECMWF Newsletter*, No. 159, ECMWF, Reading, United Kingdom, 17–24, https://doi.org/10.21957/vf291hehd7.

Hutter, F., H. H. Hoos, and K. Leyton-Brown, 2011: Sequential model-based optimization for general algorithm configuration. *Proc. Fifth Int. Conf. on Learning and Intelligent Optimization*, Rome, Italy, LION'05, 507–523, https://doi.org/10.1007/978-3-642-25566-3_40.

Jones, R. H., 1975: Estimating the variance of time averages. *J. Appl. Meteor.*, **14**, 159–163, https://doi.org/10.1175/1520-0450(1975)014<0159:ETVOTA>2.0.CO;2.

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

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

Lakshmanan, V., G. Stumpf, and A. Witt, 2005: A neural network for detecting and diagnosing tornadic circulations using the mesocyclone detection and near storm environment algorithms. *21th Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, San Diego, CA, Amer. Meteor. Soc., J5.2, https://ams.confex.com/ams/Annual2005/techprogram/paper_82772.htm.

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

Li, R., H. D. Bondell, and B. J. Reich, 2019: Deep distribution regression. https://arxiv.org/abs/1903.06023.

Marzban, C., and G. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.*, **35**, 617–626, https://doi.org/10.1175/1520-0450(1996)035<0617:ANNFTP>2.0.CO;2.

——, and ——, 1998: A neural network for damaging wind prediction. *Wea. Forecasting*, **13**, 151–163, https://doi.org/10.1175/1520-0434(1998)013<0151:ANNFDW>2.0.CO;2.

——, and A. Witt, 2001: A Bayesian neural network for severe-hail size prediction. *Wea. Forecasting*, **16**, 600–610, https://doi.org/10.1175/1520-0434(2001)016<0600:ABNNFS>2.0.CO;2.

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.

Mundhenk, B. D., E. A. Barnes, E. Maloney, and C. F. Baggett, 2018: Skillful empirical subseasonal prediction of landfalling atmospheric river activity using the Madden–Julian Oscillation and quasi-biennial oscillation. *npj Climate Atmos. Sci.*, **1**, 20177, https://doi.org/10.1038/S41612-017-0008-2.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10**, 155–156, https://doi.org/10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2.

PRISM, 2019: Prism gridded climate data. Oregon State University, accessed 15 January 2019, http://prism.oregonstate.edu.

Python Software Foundation, 2018: Python language reference, version 3.6.6. Accessed 2018, http://www.python.org.

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

Roebber, P. J., M. R. Butt, S. J. Reinke, and T. J. Grafenauer, 2007: Real-time forecasting of snowfall using a neural network. *Wea. Forecasting*, **22**, 676–684, https://doi.org/10.1175/WAF1000.1.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, https://doi.org/10.1002/qj.2183.

——, 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.

Schlosser, L., T. Hothorn, R. Stauffer, and A. Zeileis, 2019: Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Ann. Appl. Stat.*, **13**, 1564–1589, https://doi.org/10.1214/19-AOAS1247.

Singh, D., M. Ting, A. A. Scaife, and N. Martin, 2018: California winter precipitation predictability: Insights from the anomalous 2015–2016 and 2016–2017 seasons. *Geophys. Res. Lett.*, **45**, 9972–9980, https://doi.org/10.1029/2018GL078844.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, https://doi.org/10.1175/MWR3441.1.

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.

Stauffer, R., G. Mayr, M. Dabernig, and A. Zeileis, 2015: Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. *Bull. Amer. Meteor. Soc.*, **96**, 203–216, https://doi.org/10.1175/BAMS-D-13-00155.1.

——, N. Umlauf, J. W. Messner, G. Mayr, and A. Zeileis, 2017: Ensemble postprocessing of daily precipitation sums over complex terrain using censored, high-resolution standardized anomalies. *Mon. Wea. Rev.*, **145**, 955–969, https://doi.org/10.1175/MWR-D-16-0260.1.

Switanek, M., J. J. Barsugli, M. Scheuerer, and T. M. Hamill, 2020: Present and past sea surface temperatures: A recipe for better seasonal climate forecasts. *Wea. Forecasting*, **35**, 1221–1234, https://doi.org/10.1175/WAF-D-19-0241.1.

Taillardat, M., A.-L. Fougères, P. Naveau, and O. Mestre, 2019: Forest-based methods and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Wea. Forecasting*, **34**, 617–634, https://doi.org/10.1175/WAF-D-18-0149.1.

Vigaud, N., M. K. Tippett, J. Yuan, A. W. Robertson, and N. Acharya, 2020: Spatial correction of multimodel ensemble subseasonal precipitation forecasts over North America using local Laplacian eigenfunctions. *Mon. Wea. Rev.*, **148**, 523–539, https://doi.org/10.1175/MWR-D-19-0134.1.

Wang, L., and A. W. Robertson, 2019: 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.

Ware, E. C., D. M. Schultz, H. E. Brooks, P. J. Roebber, and S. L. Bruening, 2006: Improving snowfall forecasting by accounting for the climatological variability of snow density. *Wea. Forecasting*, **21**, 94–103, https://doi.org/10.1175/WAF903.1.

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.

——, 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.

Zhang, C., 2013: Madden–Julian oscillation: Bridging weather and climate. *Bull. Amer. Meteor. Soc.*, **94**, 1849–1870, https://doi.org/10.1175/BAMS-D-12-00026.1.

——, and B. Zhang, 2018: QBO–MJO connection. *J. Geophys. Res. Atmos.*, **123**, 2957–2967, https://doi.org/10.1002/2017JD028171.

Zhang, Y., L. Wu, M. Scheuerer, J. Schaake, and C. Kongoli, 2017: Comparison of probabilistic quantitative precipitation forecasts from two postprocessing mechanisms. *J. Hydrometeor.*, **18**, 2873–2891, https://doi.org/10.1175/JHM-D-16-0293.1.

Zsoter, E., 2006: Recent developments in extreme weather forecasting. *ECMWF Newsletter*, No. 107, ECMWF, Reading, United Kingdom, 8–17, https://www.ecmwf.int/node/17958.