RUFCO: A Deep Learning Framework to Postprocess Subseasonal Precipitation Accumulation Forecasts

ROCHELLE P. WORSNOP,[a] MICHAEL SCHEUERER,[b] THOMAS M. HAMILL,[c] TIMOTHY A. SMITH,[a] AND JAKOB SCHLÖR[d]

[a] *NOAA/Physical Sciences Laboratory, Boulder, Colorado*
[b] *Norwegian Computing Center, Oslo, Norway*
[c] *The Weather Company, Boulder, Colorado*
[d] *Machine Learning in Climate Science, University of Tübingen, Tübingen, Germany*

(Manuscript received 12 March 2024, in final form 2 August 2024, accepted 3 October 2024)

ABSTRACT: Postprocessing is a critical step in attaining calibrated and reliable probabilistic forecast output from numerical weather prediction models. A novel deep learning framework is proposed to postprocess 20 years of 7- and 14-day precipitation accumulation reforecasts from the Global Ensemble Forecast System at subseasonal time scales (week 1, week 2, and combined weeks 3–4 forecasts) over the contiguous United States. The network builds upon previous studies and is a combination of three parallel-trained components suitable for subseasonal prediction. The first is a ResUnet architecture which learns nonlinear relationships between binned observed precipitation and input images of weather and geographical variables. The second conditions the network to the month-of-year via a feature-wise linear modulation (FiLM) layer. The third helps the network learn when to revert the forecast to that of climatology. The RUFCO (named for its components ResUnet, FiLM, and Climatological-Offramp) forecasts are compared against raw and climatological forecasts as well as those from a state-of-the-art distributional regression postprocessing model, "censored, shifted gamma distribution (CSGD)," and a simple bias-corrected model. At week 1, every method exhibited a competitive advantage over climatological forecasts. At week 2, RUFCO generated forecasts with statistically significant improvement over climatology at 82%–94% of the domain, beating CSGD's coverage of 76%–90% of the grid points. At week 3, RUFCO's skillful coverage was 65%–85%, while CSGDs dropped to only 12%–37%. At the longer lead times, RUFCO achieved the highest domain-averaged skill scores across seasons. However, the network tends to "smooth" forecast skill, making it less competitive with CSGD in limited areas with strongly spatially varying biases.

SIGNIFICANCE STATEMENT: Precipitation accumulation forecasts 1, 2, and 3–4 weeks in advance are increasingly in-demand for a variety of decision-making applications around hydrologic forecasting, flood and drought awareness, and wildfire preparedness. However, raw forecasts from numerical weather prediction systems have errors that hinder skill. Postprocessing methods remove those errors and provide more reliable and skillful forecasts. We show that a new neural network technique is an effective and competitive postprocessing tool compared to more traditional techniques.

KEYWORDS: Forecast verification/skill; Probabilistic Quantitative Precipitation Forecasting (PQPF); Probability forecasts/models/distribution; Postprocessing; Deep learning; Neural networks

1. Introduction

Forecasting at subseasonal time scales (typically between 2 weeks and 3 months ahead) is one of the latest prediction frontiers. Multiple projects and corresponding databases such as the subseasonal to seasonal (S2S) prediction project (Vitart et al. 2017) and subseasonal experiment (SubX; Pegion et al. 2019) were generated with the sole purpose of accelerating research to improve operational S2S forecasts. Regarding accumulated precipitation, which is the focus of this study, these forecasts are critical for assessing food security (Breeden et al. 2022), managing water resources (Yuan et al. 2014), evaluating the severity of ongoing and building droughts (Pendergrass et al. 2020; Hoell et al. 2020), and identifying locations and times with wildfire risk (White et al. 2017; Abatzoglou et al. 2023).

Numerical weather prediction (NWP) models generally produce skillful midlatitude forecasts ~7–10 days ahead (Zhang et al. 2019), albeit with more skill for variables like surface temperature (Hamill et al. 2004; DelSole et al. 2017) and for accumulations over longer durations and larger areas (Roberts 2008; Roberts and Lean 2008). At subseasonal time scales, the chaotic growth of model errors (Lorenz 1969; Slingo and Palmer 2011) limits the information provided by the initial conditions. Even more, the predictability associated with slowly varying boundary conditions (e.g., sea surface temperatures, soil moisture, and sea ice) does not provide substantive predictability enhancement until longer, seasonal time scales (Vitart 2004; Merryfield et al. 2020). Because subseasonal

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-24-0020.s1.

Corresponding author: Rochelle Worsnop, rochelle.worsnop@noaa.gov

DOI: 10.1175/AIES-D-24-0020.1 e240020

forecast applications often are desired for relatively short aggregations (i.e., daily, 1 week, or 2 weeks) and because they bridge this gap of predictability between weather and climate scales, they notoriously suffer from a lack of skill.

The aforementioned model errors may manifest as systematic biases and ensemble dispersion errors. These have many causes, including limited model resolution, biases in the initial and boundary conditions, and inaccurate representations of subgrid-scale processes associated with the use of parameterization schemes. Statistical postprocessing techniques can correct these errors within the raw subseasonal forecasts so that every forecast, at every location, is as skillful and reliable (i.e., probabilities will be more consistent with the observed relative frequencies) as possible. Machine learning (ML) methods are a natural extension of statistical postprocessing tools (Rasp and Lerch 2018; Haupt et al. 2021; Vannitsem et al. 2021, and references therein). The appeal is their ability to model highly nonlinear relationships among a suite of predictors and the predictand(s) and their flexibility to seamlessly incorporate a variety of data types. While some deep learning approaches do assume a predictive distribution a priori and use networks to learn their parameters (Rasp and Lerch 2018; Ghazvinian et al. 2021; Ghazvinian et al. 2022; Chapman et al. 2022; Hu et al. 2023), most require no assumption about the predictive distribution of the variable.

Currently, only a few studies have applied neural networks (NNs) for precipitation postprocessing at subseasonal time scales. Scheuerer et al. (2020), referred to as S20 hereafter, first demonstrated the use of two NN-based models to postprocess week 2-, 3-, and 4-forecasts of accumulated precipitation over California (CA) during the cool season using Integrated Forecasting System (IFS) reforecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF). Fan et al. (2023) used a multilayer perceptron to postprocess the National Oceanic and Atmospheric Administration (NOAA)'s Climate Forecast System (CFS; Saha et al. 2014) week 3–4 forecasts of precipitation and a 2-m temperature (t2m). Horat and Lerch (2024) used variations of convolutional neural networks to postprocess combined weeks 3–4 and weeks 5–6 and global IFS tercile forecasts of t2m and precipitation.

Here, we propose a deep learning architecture that leverages decades of NOAA Global Ensemble Forecast System (GEFSv12) reforecasts to postprocess subseasonal ensemble precipitation forecasts over the contiguous United States (CONUS) for each season. Zhou et al. (2022) showed that GEFSv12's CONUS-wide skill drops below that of climatology by day 16 lead times for precipitation amounts >1 mm day^{-1} and by day 11 for events over 20 mm day^{-1}. These findings motivate the need for postprocessing to improve the quality of the forecasts so that they are more skillful than a simple climatological forecast.

Our postprocessing algorithm builds upon the ResUnet (Zhang et al. 2018) to account for intricacies of subseasonal prediction (see section 3c). The scope of this paper is to introduce and demonstrate the skill of the new network using a simple set of inputs (i.e., forecasted pressure and moisture variables, geographic variables, and time-of-year embeddings) with the objective of improving subseasonal 7- and 14-day accumulated precipitation outlooks for NOAA's Climate Prediction Center.

Section 2 describes the reforecasts and gridded observation data as well as the data-splitting and cross-validation strategies used to train and evaluate the postprocessing models. Section 3 outlines the new ResUnet, feature-wise linear modulation (FiLM), and Climatological-Offramp (RUFCO) network, and section 4 explains the conventional benchmark postprocessing method. Section 5 discusses RUFCO's hyperparameter tuning strategy. Probabilistic verification metrics are discussed in section 6, while section 7 details the results, including a sensitivity experiment of the FiLM component. Finally, sections 8 and 9 include an example forecast outlook and summary, respectively.

2. Data

a. *Reforecasts*

Training neural networks typically requires large datasets to learn complex tasks. Therefore, we use the 20-yr GEFSv12 reforecasts (Guan et al. 2022; Hamill et al. 2022; Amazon Web Services 2023) which are forecasts retrospectively generated from a consistent model version of GEFSv12. GEFSv12 is a global atmospheric model ran operationally at NOAA's National Centers for Environmental Prediction for lead times up to 35 days. GEFSv12 was implemented in September 2020 after demonstrating its improved performance over previous versions, including for forecasts of precipitation. Details about the model configuration, including the cloud microphysics, radiation, and convection schemes, are given in Zhou et al. (2022), Guan et al. (2022), and Hamill et al. (2023).

The reforecasts use the same model system as the real-time forecasts, but their ensemble size and initialization dates are more limited. We retrieved reforecasts for dates between 1 January 2000–31 December 2019 that were initialized at 0000 UTC and executed to +29 days lead, with data output at 6-h increments. The reforecasts were initialized once per week (i.e., a total of 1040 dates) and included 11 ensemble members. Data for the first 10 days were available at a 0.25° rectilinear grid spacing; we upscaled to 0.5° with the conservative regridding algorithm (Jones 1999), which is suitable for remapping spatially sparse variables, via the xESMF Python-based regridding library (Zhuang et al. 2023). Data for leads 10+ days were already available at a 0.5° grid spacing.

We retrieved data for a spatial domain surrounding CONUS from 134° to 65°W longitude and 23° to 53°N latitude for each of the weather-predictor variables: precipitation accumulation, total column water (TCW), and geopotential height at 500-hPa (Z500). The domain has a total of 139 × 61 (8479) grid points. Precipitation totals at each grid point were then summed over 7 days for week-1 and week-2 forecasts and over 14 days for the combined week-3 and week-4 forecasts. We accumulated precipitation amounts starting at 1200 UTC to be consistent with the start of the accumulation period of the observed dataset (see section 2b). Therefore, a week-1 forecast included precipitation that fell between 12- and 180-h (0.5–7.5 days) lead times, a week-2 forecast included 180–348-h (7.5–14.5 days) lead times, and the combined weeks-3 and -4 forecasts included hours 348–684-h (14.5–28.5 days) lead times. TCW and Z500 were averaged over the same 7 or 14 days and were selected as predictors given their representation of the larger-scale patterns associated with precipitation.

b. *Gridded precipitation analyses*

The GEFSv12 reforecasts are calibrated and verified with gridded precipitation analyses upscaled from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM; Daly et al. 1997) precipitation dataset. PRISM data are available over land as 24-h accumulations starting at 1200 UTC for years 1981–2019. We again apply conservative regridding to upscale and match PRISM's original 4-km rectilinear grid to the 0.5° grid of the reforecasts (a total of 3371 CONUS grid points) and accumulate over 7- and 14-day periods for each grid point.

c. *Data-splitting strategy*

For fair evaluation and comparison of the various postprocessing methods presented below, the data of forecast-observation pairs need to be split into two independent datasets. One set will be used to train the algorithms and is referred to in the paper as the "training" set and the other set, called the "test" set, will only be used for final evaluation to assess the methods' performance on unseen data. We maximized the use of the data by implementing the splits in a leave-some-out cross-validation procedure described below. For the benchmark postprocessing algorithm (censored, shifted gamma distribution (CSGD); discussed in section 4), we fit a model separately for each month. Therefore, we leave 1 month and year combination out to make up the test dataset. A particular month's data from the remaining 19 years serve as the training dataset (more details in section 4). The process cycles through all months and years so that eventually every month and year are evaluated.

The splitting strategy (visualized in Fig. SA1 in the online supplemental material) of the new neural-network-based method (RUFCO; discussed in section 3) is different from CSGD's in two ways. First, an entire year serves as the test set for reasons described in section 3 rather than just 1 month. The training set includes data from all months in the remaining 19 years. Second, neural networks require a third split of the data for hyperparameter tuning (see section 5). This third set is a subset of the training set. The procedures to define that subset and to tune the hyperparameters are discussed in section 5 and in section A of the supplemental.

3. Deep neural network for categorical probability predictions

a. *Target variable*

The target variable of the neural network is a vector that specifies the probability of the precipitation amount falling into each $m + 1$ precipitation bins. S20 found that forecast performance was not sensitive to the number of bins and the number mostly translated to either fewer parameters for a network to learn when fewer bins were used or a smoother interpolation between bins in the case of more bins. Therefore, following S20, we set $m = 19$ and use the same binning scheme as S20 for their postprocessing networks for CA cool-season precipitation. Specifically, let B_i represent the ith bin that is bounded by $[b_{i-1}, b_i]$ for $i \in \{0, \ldots, m\}$. The boundary values are determined by first constructing an observed PRISM climatological distribution of 7- or 14-day accumulations specific to each location and day-of-year. Here, for each location, we pool together accumulated observations from all downloaded years (1981–2019) within a ±30-day window centered around each date of interest.

The first bin B_0 has a climatological probability of occurrence denoted by $1 - \text{pop}_{\text{cl}}$, where pop_{cl} represents the climatological probability of nonnegligible precipitation accumulation (>0.254 mm; 0.01 in.). The remaining bins $B_{1,} B_2, \ldots, B_m$ are defined by quantiles of the climatological distribution sampled at probability levels $q_{\text{cl},i} = (1 - \text{pop}_{\text{cl}}) + \text{pop}_{\text{cl}}(i/m)$ for $i \in \{1, \ldots, m\}$. By definition, all but the first (i.e., negligible precipitation) bin for a particular location and day-of-year have equal climatological probabilities of occurrence. This partition scheme helps ensure relatively balanced bins of accumulated precipitation amounts no matter the location or time-of-year. This structure is ideal for the NN's task to simultaneously predict the likelihood of category assignments across the large domain for any date. Figure 1 shows examples of percentiles of the climatological distribution (blue dots) and associated 20 bin partitions (black lines) for locations along a west–east transect over CONUS through 39.5°N.

b. *Input variables*

We use the ensembles of raw accumulated precipitation, TCW, and Z500 from GEFSv12 in addition to geographic variables (latitude, longitude, and elevation) to correct for spatial-dependent precipitation biases. We first smooth each of the raw forecast variables using a neighborhood-smoothing procedure described for the CSGD method in section 3a of S20. Then, for each of the smoothed variables, we calculate an extreme forecast index (EFI; Zsótér 2006). The EFI quantifies the departure of the raw ensemble forecast distribution from the cumulative distribution function (CDF) of its own model climatology. A detailed description of the formulation that we used to calculate the EFI is found in appendix A of S20.

The input fields for our neural network are images of the EFI-transformed weather variables, latitude, longitude, and elevation. While the EFI values are inherently defined between [−1, 1], we use a min–max normalization to rescale the remaining inputs between −1 and 1. In addition to the input fields, we include scalar values to condition our network on. These include the one-hot encoded vector of the month-of-year and the logarithm of the climatological probabilities, detailed in sections 3c(2) and 3c(3), respectively.

During preliminary analysis, we also tried using the normalized ensemble mean, standard deviation, and the 10th and 90th quantiles of the ensembles as inputs to the network. The

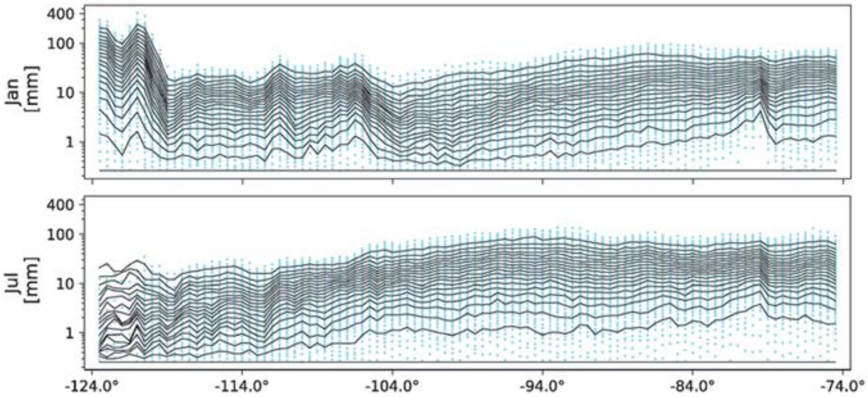

FIG. 1. Example bin boundaries (black) that define the 20 precipitation accumulation categories predicted by the RUFCO network. Categories are based on the 1981–2019 observed PRISM climatology specific to each grid point and day of the year. Boundaries are shown for (top) 15 Jan and (bottom) 15 Jul for grid points along a west–east transect over CONUS at 39.5°N latitude. Equidistant percentiles from 1 to 99 of the climatological distribution are shown by the blue dots.

EFI typically showed lower training loss values in fewer epochs (not shown), so we opted to use EFI as the predictors. Figure 2 shows example inputs to the RUFCO model for a heavy-precipitation event in January 2014 during a North American cold wave.

c. *RUFCO network architecture*

The new postprocessing NN is made up of three components that are trained in parallel: 1) a residual U-Net (ResUnet) (Zhang et al. 2018) to learn nonlinear relationships between predictors and observed precipitation bins, 2) a FiLM layer (Perez et al. 2017) to help the network condition the prediction based on the month-of-year, and 3) a "climatological off-ramp" to help the network revert toward climatology when necessary. Each component is described below, and the overall network shown in Fig. 3 is referred to as the RUFCO model. Table A1 in appendix defines each operation within Fig. 3.

1) RESUNET COMPONENT

The ResUnet is a network that combines the advantages of U-Nets (Ronneberger et al. 2015; Long et al. 2015) and residual learning networks (He et al. 2016) to improve model training and prediction of semantic segmentation "pixel-level" classification tasks. These tasks entail the simultaneous prediction of the most likely category for each pixel in an image rather than predicting a category assignment for the whole image. Thus, this type of network is suitable for our task of simultaneously predicting the probability of each grid box falling within one of many possible precipitation bins.

U-Net is a type of encoder–decoder convolutional network that has been used for a broad range of meteorological-forecasting applications (Chapman et al. 2022; Hu et al. 2023; Badrinath et al. 2023; Lagerquist et al. 2023; Horat and Lerch 2024) that want to leverage spatial dependencies within the data. The architecture consists of repeated building "blocks" that are made up of sequential mathematical operations including convolutions, batch normalization (Ioffe and Szegedy 2015), and nonlinear activation functions (Zeiler et al. 2013; Hara et al. 2015). The encoder blocks learn and extract complex features using "filters" (often called channels) applied across the input images. The result is distilled contextual "feature maps." In our network, the number of learned filters increases by a factor of 2 for each new level of the U-Net. This increase in parameters is offset by decreasing the dimensionality of the feature maps by a factor of 2 via max pooling operations (Boureau et al. 2010; Zafar et al. 2022). The decoder side performs opposing blocks of operations from the encoder side. Between each decoder block, we use upsampling (a simple doubling of rows and columns) followed by a convolutional layer to increase the dimensionality and to reconstruct fine-scale detail within the learned feature maps. A "bridge" containing one block connects the symmetric encoder–decoder at the lowest level, thus giving it a *U* shape. Lastly, skip connections are used between the encoder and decoder sides within each level of the ResUnet.

A ResUnet's (Zhang et al. 2018) structure improves training and performance by using "residual [connection] blocks" (ResBlocks) instead of basic convolutional blocks. These blocks propagate information from the block's input to the block's output. A clear distinction between skip and residual connections is that skip connections propagate information by bypassing many layers and operations, while residual connections typically propagate information more locally within the network.

The ordering of convolutional layers, batch normalization, and activation functions within a block can vary. We use a block combination inspired by ResNet34 (He et al. 2016) where two convolutional layers are each followed by a batch normalization layer and then a nonlinear activation function. Then, a residual connection is used to add the block's input directly to the result before a final nonlinear activation function is performed. We add a dropout layer (Srivastava et al. 2014) after the block to further help in regularizing the model.

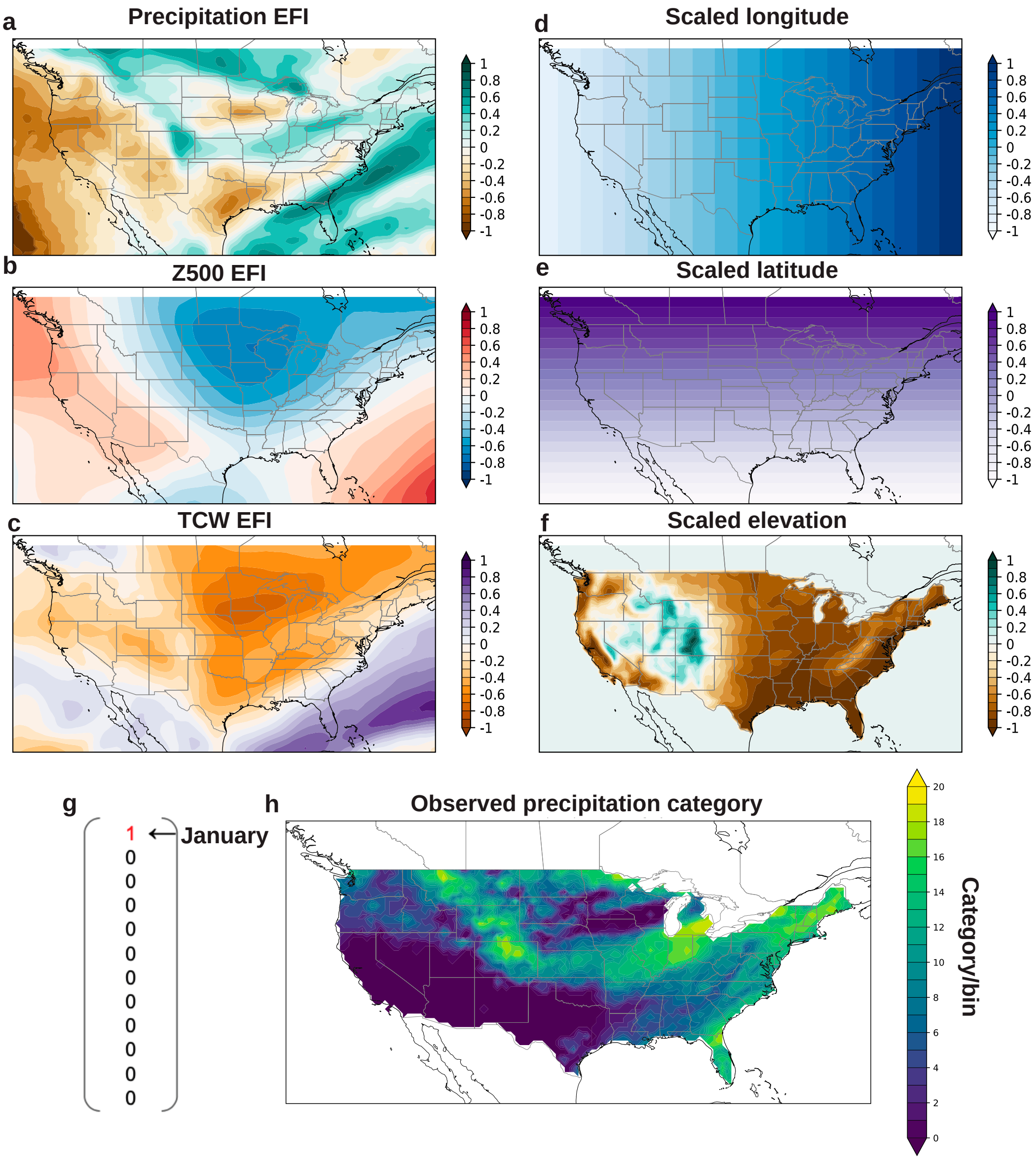

FIG. 2. Example of (a)–(g) input variables to the RUFCO model and the flattened (h) predictand (i.e., the actual predictand is a one-hot encoded array filled with ones in the observed category and 0s in the remaining 19 categories for each grid point). (a)–(f) Image inputs are described in section 3b; the (g) one-hot encoded month-of-year vector is described in section 3c(2). The scaled images represent the week-1 (for days 1–8 Jan 2014) predictors and observed accumulated precipitation for a cold wave event starting on 1 Jan 2014.

We use zero padding so that the output images are the same size as the input images. A schematic of the ResUnet architecture used in our study is illustrated in Fig. 3 (circles 1, 3, and 5).

2) FILM COMPONENT: CONDITIONING THE NETWORK ON THE MONTH-OF-YEAR

To enable the network to learn variations in forecast skill based on the month-of-year, we added a FiLM (Perez et al. 2017) layer, which is a general-purpose conditioning method that has achieved state-of-the-art performance in computer vision tasks. In our implementation, the FiLM layer modulates each pixel of intermediate feature maps within the ResUnet's top-level ResBlocks. The FiLM layer applies a linear affine transformation that can scale the output (e.g., increasing or decreasing each pixel's weight) based on what the network learns about observed precipitation given the month of the year. Specifically, coefficients $\gamma_{i,c}$ and $\beta_{i,c}$ of the FiLM Eq. (1) scale intermediate feature maps $F_{i,c}$, where i and c are the ith sample and the cth feature map within the ResBlock, respectively:

$$\text{FiLM} = \gamma_{i,c}F_{i,c} + \beta_{i,c}. \qquad (1)$$

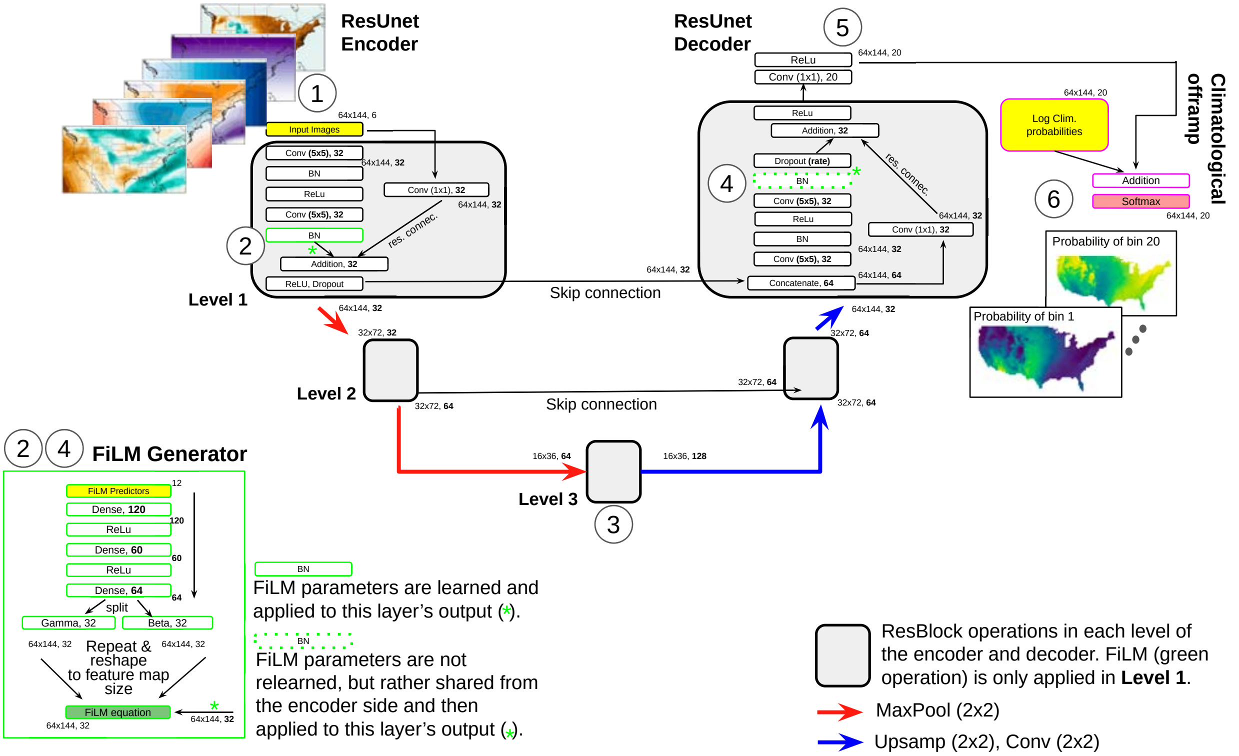

FIG. 3. (top left) Schematic of the RUFCO architecture used in this study as well as example input images and (middle right) the forecasted categorical probability outputs. All operations are performed in order from 1 to 6 as indicated by black-outlined circles. Circles 1, 3, and 5 represent the input variables (see Fig. 2) and the general "ResUnet" component of the network. Circles 2 and 4 represent the FiLM layer applied to the output (indicated by green asterisks) of intermediate layers within the top level of the ResUnet. Circle 6 indicates the climatological off-ramp component. Rectangles and thick arrows indicate sequential mathematical operations except for the yellow-highlighted rectangles, which represent inputs to the three main components of the overall network. Table A1 in appendix defines each operation. Numbers upstream (downstream) of the operations correspond to the spatial size of images input (output) to that layer as well as the number of channels/filters input (output) to the layer. Rectangles without a new size retain the size of the prior-listed size. Bolded values within the schematic indicate values that were tuned. All tuned values shown in this schematic are from one optimized model, which had 3 levels/blocks. The entire network is trained in parallel. A clickable figure may be rotated and enlarged online for enhanced readability.

The FiLM coefficients are learned by a "dense" network called a "FiLM generator function" as shown in Fig. 3 (circles 2 and 4 and the corresponding green-outlined components) that is trained end to end with the other components. Placement of the FiLM layers was guided by ablation studies in Perez et al. (2017). Their findings indicate that the best performance is found when FiLM layers are placed within the ResBlock.

Since the output of the ResBlock has N number of feature maps, both of the FiLM coefficients also need to have N values. In this way, the FiLM layer can modulate each of the feature maps independently based on the month of year. This is achieved by setting the last linear layer of the FiLM generator network to have $N \times 2$ neurons, which are then split in half so that the first $N/2$ learned values are assigned to γ and the last $N/2$ are assigned to β. The input to the FiLM generator is a one-hot encoded array with 12 columns for each training sample where a one indicates the month of the year and zeros indicate the remaining months of the year (see Fig. 2g).

The advantage of the FiLM conditioning approach is that it is a scalable and computationally efficient (Perez et al. 2017) way to incorporate scalar predictors into image-based networks. Only two new parameters are needed to condition each feature map no matter the size of the predictor images (e.g., small regional versus large global images). The FiLM generator itself does create parameters depending on its complexity. To reduce the number of additional parameters, we only learn the FiLM generator and the FiLM coefficients once (on the encoder side) and then reuse those coefficients on the decoder side. This is feasible because each level of the encoder and decoder sides of our ResUnet has the same sized feature maps. This implementation differs from that of Perez et al. (2017) who learned new FiLM generators and coefficients for each ResBlock.

3) CLIMATOLOGICAL OFFRAMP COMPONENT: NETWORK ADJUSTMENTS TOWARD CLIMATOLOGICAL PROBABILITIES

With increasing lead time, the amount of useful information that the predictors can provide decreases and reliable subseasonal forecasts tend to converge toward climatological forecasts.

To help the network learn to revert toward climatology, we include a "climatological offramp" by adding the logarithms of climatological probabilities (as calculated in section 3a) for each bin to the output of the FiLM-ed ResUnet (see Fig. 3, circle 6). This idea was introduced in S20 and does not require any additional parameters for the network to learn. For more details, we point interested readers to section 3b of S20.

d. Loss function

The network learns its parameters by minimizing the modified categorical cross-entropy score (MCCES) introduced by S20:

$$\mathscr{L}_{0,\ldots,\mathrm{m}}(\mathbf{p}, \mathbf{y}) = -\log\left(\sum_{i=0}^{m} \mathbf{y}_i \mathbf{p}_i\right), \tag{2}$$

where $\mathbf{p} = (p_0, \ldots, p_m)$ is a vector of forecasted probabilities associated with each of the $m + 1$ categories. Let $\mathbf{y} = (y_0, \ldots, y_m)$ be the one-hot encoded vector where ones represent the observed category assignment and zeros fill the remaining 19 categories.

MCCES generalizes the standard categorical cross-entropy score (CCES) to include the rare scenario where an observation falls into more than one bin. Ambiguous category assignment is a result of limited precision associated with observed, gridded precipitation datasets and the small spacing between climatological quantiles used to define the categories. In most cases, the assignment is unambiguous and the MCCES reduces to the CCES.

e. Interpolation between categories

Once the network has been trained and forecasted probabilities have been assigned for each precipitation bin, we generate a full predictive CDF F by interpolating between the bins. We follow the interpolation procedures in S20 who found that first transforming the categorical probabilities to $H(x) = -\log[1 - F(x)]$ enabled a simple linear interpolation between bins and a linear extrapolation to values outside of the bins. With the full predictive CDF, we then estimated the probability of exceedance of various climatological thresholds for use in the verification metrics discussed in section 6.

4. Benchmark postprocessing technique (CSGD)

To evaluate the performance of the RUFCO method in the context of established postprocessing techniques, we use the CSGD method (Scheuerer and Hamill 2015; Ghazvinian et al. 2021), which is a state-of-the-art method that has generated skillful and reliable quantitative precipitation forecasts in several studies (e.g., Scheuerer and Hamill 2015; Zhang et al. 2017; Worsnop et al. 2021), even at subseasonal scales (S20). The CSGD method is a parametric, distributional regression technique; it assumes a prescribed distribution family and fits a regression model for the different distribution parameters to a training set of forecast-observation pairs. The specific distribution family used therein, the CSGD, is a modified gamma distribution that allows one to capture characteristics of precipitation such as right skewness and positive probabilities of exactly zero precipitation. The employed regression equations appropriately permit increases in spread with precipitation magnitude (Scheuerer and Hamill 2015).

The CSGD method is typically fitted separately for each grid point and each month or season to account for varying geographical or seasonal biases of the raw forecasts. We implement nearly the same local CSGD fitting procedure and regression equations that S20 used to postprocess cool-season precipitation accumulations over CA. In the present study, we generate forecasts year-round, so we fit a predictive CSGD for each month of the year. Therefore, the test set for the CSGD method consists of accumulations from all reforecast dates in a particular month for a particular year. The training set consists of accumulations from all available reforecast dates within a 91-day window surrounding the 15th of that month but only for the remaining 19 years.

Like S20, we also opted to use ad hoc regression parameters at very dry locations instead of fitting data to the CSGD model. This was done to avoid poor performance caused by overfitting in regions and at times of year when there were few nonzero accumulations in the training dataset. For more details about the ad hoc adjustments, the CSGD method, and specific implementation used in this paper, we point the reader to section 3a of S20.

Limitations to the CSGD method are that its relationship between the predictor and predictand is inflexible and its grid-point-specific approach does not use potentially valuable spatial information from other parts of the domain. Therefore, we use it as a benchmark to evaluate if the proposed network's more flexible and convolutional approach can circumvent these limitations and produce better performance.

5. Hyperparameter tuning

Several decisions related to the architecture, learning and optimization, and regularization of the network are not learned by the network itself and therefore need to be set a priori by the implementer. Supplemental A outlines the cross-validated hyperparameter (HP) tuning approach and software that we employed to select optimal HPs (listed in Table A2 of the appendix). Fifty different combinations of HPs were tested for each of the 20 different networks, one network for each of the 20 left-out-(test)years. The best combination (based on criteria in supplemental A) was used to tune and train the final "optimized model." Figure A1 and associated discussion detail the tuning results and provide insights into which parameters RUFCO are most sensitive to. Networks were tuned and trained separately for each lead time.

6. Probabilistic verification metrics

Precipitation outlooks are typically presented as probabilistic forecasts of categorical departures from normal. For weeks 3–4, our partners at NOAA's Climate Prediction Center (CPC) only issue two-category outlooks for below- and above-normal events where "normal" is defined as the 50th percentile of the observed climatological distribution. CPC uses the median instead of the mean (like they use for weeks 3–4 outlooks of temperature) because it gives a better representation of the central tendency/split of the precipitation distribution; precipitation distributions are typically heavy tailed so that observed amounts most often fall on the dry side of the mean. Probabilities of exceeding the terciles (33rd and 67th percentiles) and 85th percentile of climatology are also of interest for precipitation forecasting. Therefore, we use all of these percentiles [0.333, 0.50, 0.667, 0.85] to calculate ranked probability skill scores [RPSSs; Epstein 1969; Murphy 1971; Wilks 2011, Eq. (8.52)] to quantify the calibration and sharpness of the raw, CSGD- and RUFCO-based probabilistic forecasts. We also construct reliability diagrams (RDs; Wilks 2011, chapter 8.4.4) and the corresponding Brier skill scores [BSSs; Wilks 2011, Eq. (8.37)] for exceedances of tercile events (0.333 and 0.667). As an additional postprocessing comparison, we calculate the same verification metrics for what we refer to as a simple "bias-corrected" forecast. For this forecast, the threshold percentiles are determined from the raw model climatological distribution instead of from the observed climatology. For example, a bias-corrected forecast for the above-normal events will be based on the probability of the raw forecast ensemble exceeding the 66.7th percentile of the raw model climatology. Therefore, the bias-corrected forecast is not a direct adjustment to the raw forecast but rather achieves inherent bias correction by adjusting the exceedance threshold values based on the raw model's own climatology.

A reference forecast is required to calculate skill scores for the raw forecast and for each of the three postprocessing methods. We construct a climatological forecast to serve as the shared benchmark. The climatological forecast, separately for each grid point, is made up of observed accumulations for years 2000–19 except for the year associated with the date of interest (i.e., the left-out test year). The dates included are those that fall within ±30 days centered around the date of interest for which there were both observations and forecasts available. BSS or RPSS values of 1.0 indicate the largest possible improvement compared to the climatological forecast; values of 0.0 indicate the same skill as climatology, and negative values indicate worse skill than climatology.

7. FiLM sensitivity experiment and verification of the probabilistic forecasts

a. *FiLM sensitivity*

The FiLM layer herein helps the network learn the seasonality of precipitation accumulation. We demonstrate this effect by tuning and training two new networks, one with and one without the FiLM layers. To isolate the FiLM's ability to condition a network, in this case on the month-of-year, we run this sensitivity experiment without the climatological offramp (i.e., the same architecture as in Fig. 3 but without the sixth component). This is done because the FiLM's impact is likely to be somewhat masked, especially at longer lead times when the network is more likely to lean on the climatological offramp. We focus results for this experiment on lower tercile events (≤33.3rd percentile) in winter and summer months since these seasons demonstrate strong seasonality across CONUS. Figure 4 shows that the inclusion of the FiLM layer can change the forecast probability of precipitation being in the lower tercile by 25% or more. The FiLM increases the probability of precipitation being in the lower tercile in regions that are typically dry for the time of year (i.e., the FiLMed network produces forecasts that are more likely to be dry in the central United States during the winter and the western United States in the summer). Conversely, the FiLM layer modulates forecast probabilities so that it is more likely to be wetter in the western United States in the winter and central and southeast United States in the summer, which is consistent with precipitation patterns during these seasons. While there are some exceptions (red-shaded areas in Fig. 4), these changes in probability translate to improved Brier scores (blue-shaded areas), with the most improvement occurring in regions with at least a 5% change in the probability of the lower-tercile event. Similar conditioning effects are found for upper-tercile events (Fig. SC1).

b. *Reliability of forecasts for above- and below-normal events: lead-time variations in forecast reliability in winter and summer*

We demonstrate the reliability of forecasts issued for a range of forecast probabilities by showing RDs (Figs. 5 and 6) for precipitation events falling within the lower (≤33.3rd) and upper (>66.7th) percentiles of climatology. We generated confidence intervals around the reliability estimates via bootstrapping by random resampling, with replacement, of the forecast-observation pairs within the pooled time dimension (20 test years, 3 months) 5000 times. Using the resampled distributions, we calculated a bootstrapped distribution of observed relative frequencies for each forecast probability bin and then plotted the confidence interval between the 5th and 95th percentiles.

The RDs, inset histograms, and corresponding BSS in Figs. 5 and 6 reveal several characteristics of the different forecast methods for winter and summer:

- Relative to the postprocessed forecasts, the raw forecast ensemble is sharp; many forecasts (as seen in the inset histograms) are issued near 0% or 100% forecast probability of falling below the 33.3rd or exceeding the 66.7th percentiles of climatology. However, the large departure below the diagonal 1–1 line indicates that the raw forecasts lack reliability and are overconfident, especially for forecasts issued with high probability at longer lead times. Even more, the resolution of the raw forecasts is poor as seen by the relatively gentle slope of the black line compared to the 1–1 line; the raw forecast is not able to strongly discern different observed outcomes even when the issued forecast probabilities are substantially different. The resolution is especially poor at weeks 3–4 as indicated by the nearly flat black line, which means that the raw forecast is not able to issue reliable forecasts when forecasted probabilities are much smaller or larger than the climatological expectation.
- The simple bias-corrected forecast has slightly better reliability and BSS (solid gray in Figs. 5 and 6) than the raw forecast at all lead times, but the improvement is limited.

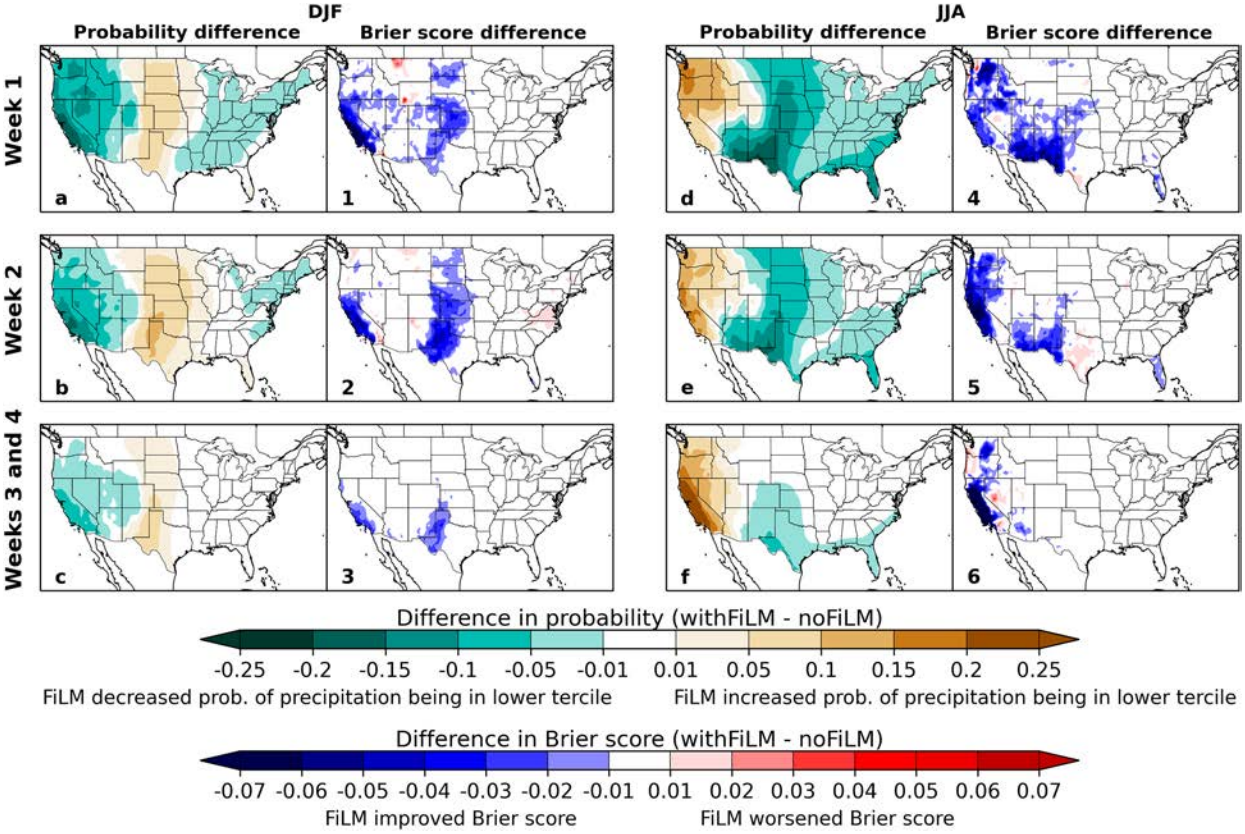

FIG. 4. Differences in forecast probabilities (a)–(f) and Brier scores (1–6) for lower-tercile precipitation events between networks tuned and trained with and without the FiLM layer. Differences are averages from data initialized in December–February (DJF) or June–August (JJA) and all test years. Three lead times (week 1, week 2, and combined weeks 3 and 4) are shown. Implications of the color shading are written underneath the colorbars.

Like the raw forecasts, the bias-corrected forecasts have negative BSS in JJA at week 2 and for both seasons at weeks 3–4 and still suffer from poor resolution. This indicates that a more sophisticated postprocessing method is needed to generate reliable forecasts.

- The CSGD and RUFCO postprocessing methods improve the reliability and BSS at all lead times and for both seasons. In the winter, the RUFCO and CSGD BSSs for both event types indicate an approximate 41%, 10%, and 3% improvement over a climatological forecast for week 1, week 2, and week 3–4 lead times, respectively. Summer shows less of an improvement compared to DJF, with values of 29%, 6%, and 2% for those lead times. The two methods generally have comparable near-perfect reliability for upper-tercile events in both seasons for week 1 and week 2. The RUFCO model, compared to CSGD, shows superior reliability for lower-tercile events at week 2, especially for JJA, likely a result of CSGD's difficulty fitting the parametric distribution or estimating the ad hoc parameters at really dry locations (see section 4). Unlike the raw and bias-corrected forecasts, the sophisticated postprocessing methods, particularly for the summer, are able to generate reliable forecasts at weeks 3–4 even for probabilities that differ from the climatological expectation.
- These RDs indicate that both the CSGD and new RUFCO model are able to produce relatively reliable forecasts at all lead times during the winter and summer. RDs for the spring and fall are provided (Figs. SD1–2) and show similar characteristics to that of winter and summer.

c. *Seasonality and regionality of ranked probability skill scores for each lead time*

Aggregate statistics in the form of CONUS-wide RPSSs for each season and lead time (Table 1) are presented. RPSSs account for all four threshold exceedances (0.333rd, 0.50th, 0.667th, and 0.85th percentiles) and quantify the combined calibration and sharpness of the forecasts compared to that of a climatological forecast.

By week 2, the raw forecast is worse than climatology in all but the winter months. The bias-corrected forecast, on average, provides a 1%–6% improvement over climatology at week 2 during fall through spring but is no longer advantageous during the summer. Conversely, the CSGD and RUFCO methods are skillful at week 2 during all seasons and provide a 6%–10% improvement over climatology, with the highest gains coming from the RUFCO model during winter and spring. RPSSs for weeks 3 and 4 show a similar result, except that the amount of improvement over climatology is reduced to 2%–3%, again

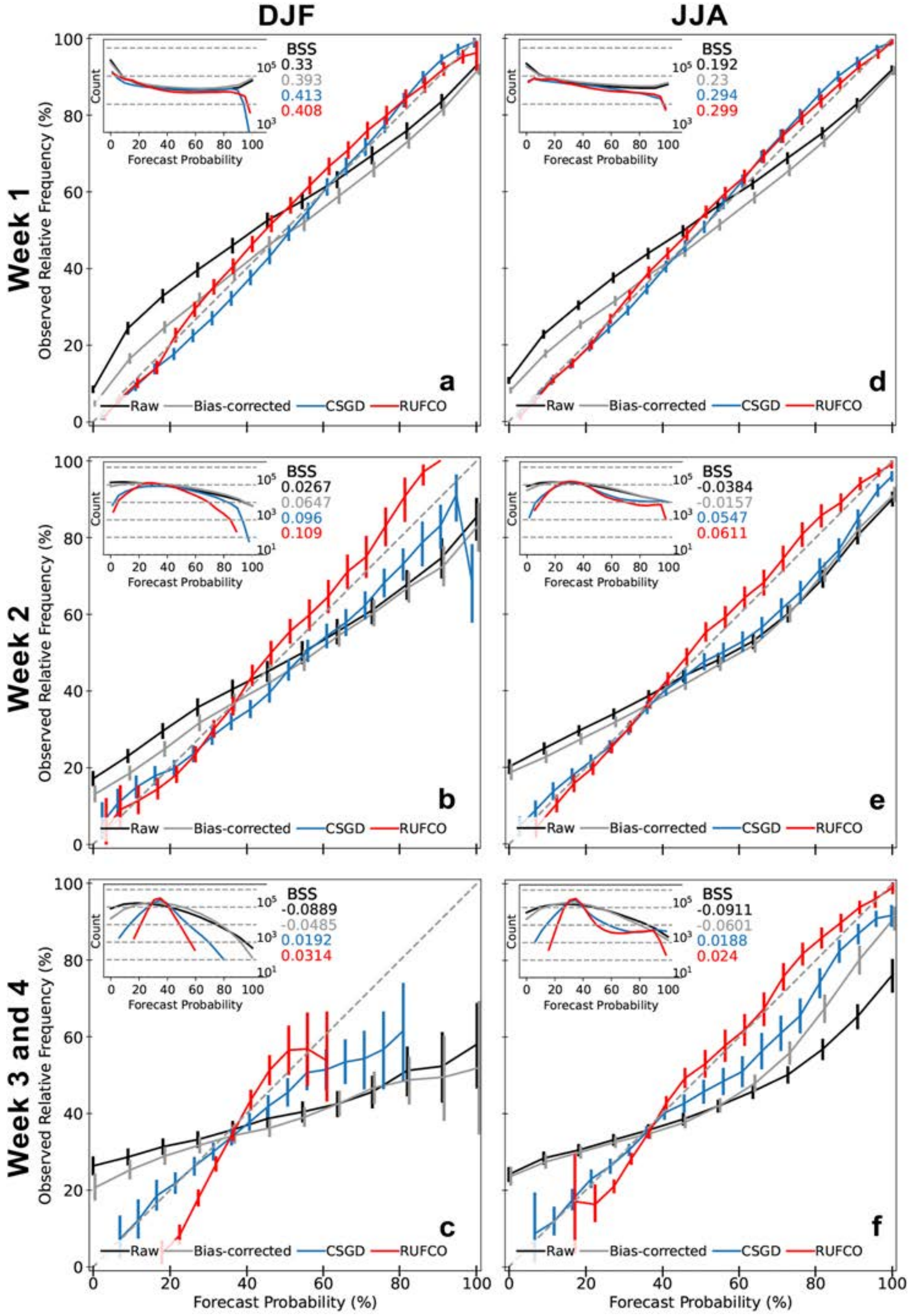

FIG. 5. RDs, BSSs, and inset relative frequency of occurrence histograms for forecasts of 7- or 14-day precipitation accumulations ≤33.3rd percentile of climatology. Data were pooled over the CONUS domain, 20 test years, and for forecasts initialized in (a)–(c) DJF or (d)–(f) JJA. Forecasts generated from the raw (black), bias-corrected (gray), CSGD (blue), and RUFCO (red) models valid at (a) and (d) week 1, (b) and (e) week 2, and (c) and (f) weeks 3 and 4 are shown. A line marker for each probability bin was only plotted if the probability bin contained at least 30 samples. Perfectly reliable forecasts exhibit a 1–1 relationship (gray, dashed line) between the forecast probability and the observed relative frequency. Confidence intervals (vertical lines) represent the 5th and 95th percentiles of a 5000-sample bootstrapped distribution with replacement (see section 7b). More details about diagram calculations are available in supplemental B.

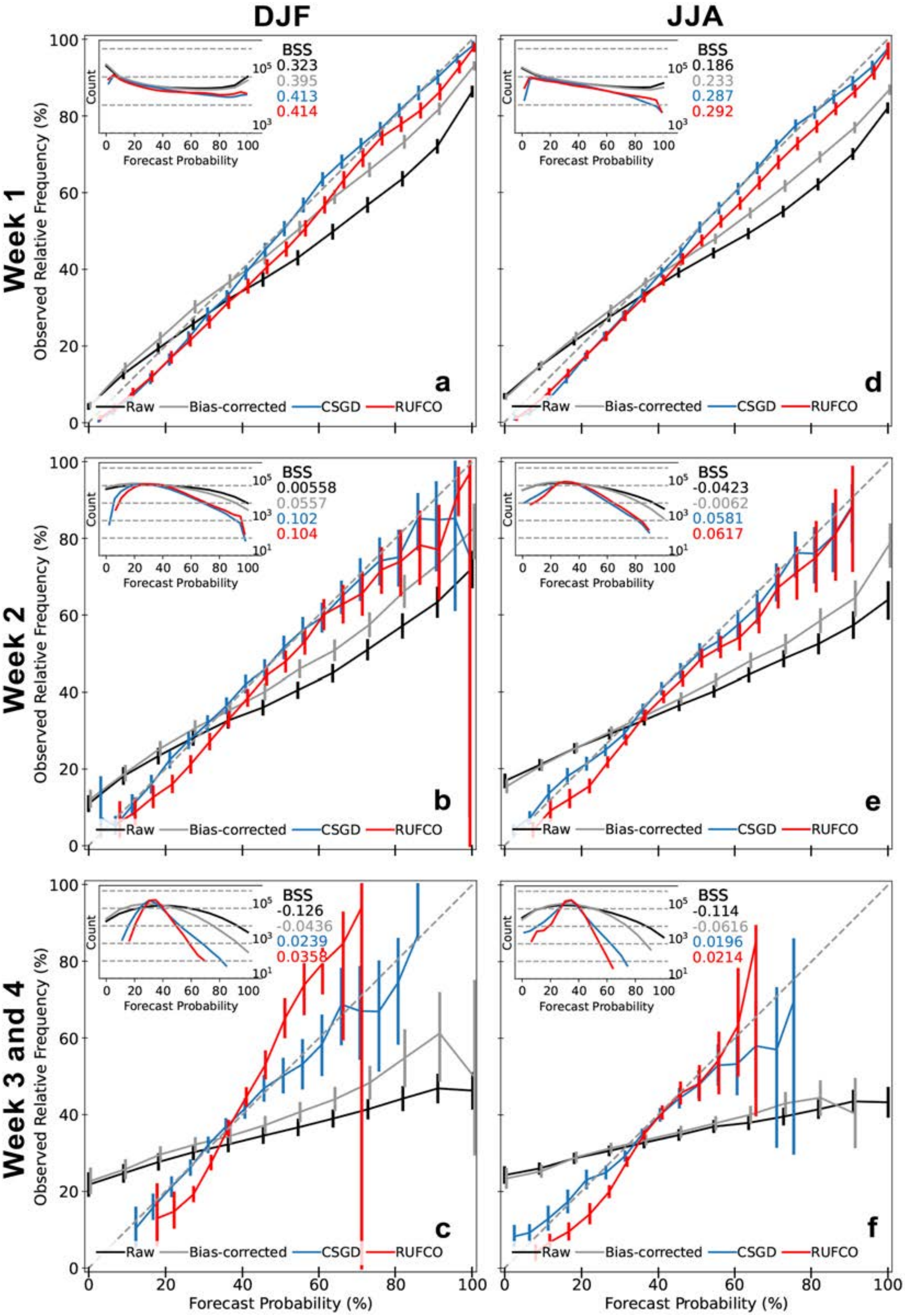

FIG. 6. As in Fig. 5, but for forecasts of 7- or 14-day precipitation accumulations >66.7th percentile of climatology.

TABLE 1. Skill (RPSS) averaged over the entire CONUS domain, each season, and all 20 test years. RPSSs were calculated using the corresponding observed climatological forecasts as the benchmark forecast. Bold values indicate the best-performing forecast method for each season and lead time.

	SON	DJF	MAM	JJA
Week 1				
Raw ensemble	0.348	0.321	0.228	0.181
Bias corrected	0.376	0.392	0.318	0.227
CSGD	0.392	**0.410**	0.348	0.285
RUFCO	**0.397**	0.408	**0.351**	**0.290**
Week 2				
Raw ensemble	−0.003	0.010	−0.035	−0.044
Bias corrected	0.010	0.058	0.023	−0.010
CSGD	0.058	0.098	0.077	0.056
RUFCO	**0.070**	**0.103**	**0.086**	**0.060**
Weeks 3 and 4				
Raw ensemble	−0.100	−0.118	−0.137	−0.109
Bias corrected	−0.073	−0.046	−0.063	−0.061
CSGD	0.008	0.022	0.014	0.019
RUFCO	**0.022**	**0.034**	**0.024**	**0.022**

with the highest gain coming from the RUFCO model. In aggregate, the RUFCO method outperforms the other methods in its ability to generate forecasts that perform better than a climatological forecast.

We next present maps of RPSSs (Figs. 7–9) at each grid point to show how the skill varies for each forecast method, season, and lead time across different regions of the CONUS domain. We only plot grid points that have a statistically significant improvement in RPS compared to those of the reference climatological forecast. The significance of improvement was calculated separately for each grid point, using a one-sided, paired t test (motivated by suggestions in Hamill 1999). Following S20, we first accounted for serial correlation that may exist between the samples of RPS differences by applying modifications to the variance used in the denominator of the test statistic [Eq. (2.15) of Jones 1975 and described further in S20]. Test multiplicity was accounted for by controlling the false discovery rate (FDR; Benjamini and Hochberg 1995) at $\alpha_{\mathrm{FDR}} = 0.1$, following the implementation around Eq. (5.37) of Wilks (2011).

Ghazvinian et al. (2022) reported the annual-average raw GEFS skill for varying magnitudes of 24-h precipitation events for various subweek lead times across CONUS. Our week-1 results are largely consistent with theirs in that the greatest relative skill occurs within the western United States, especially along the Pacific Coast, and the lowest relative skill occurs within the Northern Great Plains and southern tier of the United States, especially over Florida. Here, we subdivide the skill performance into seasons (Fig. 7) which shows that the high performance along the western United States mostly stems from cool-season events when precipitation is often a result of large-scale flow interacting with complex terrain (e.g., atmospheric rivers). These types of large-scale events are more explicitly resolved by the model than more convective-scale systems that rely on subgrid-scale parameterization schemes (Hill et al. 2023). This latter point supports the result that low annual-averaged skill mostly stems from the summer months. Gray regions indicating no raw improvement over a climatological forecast are seen over central and southern California and other small regions along the west. This may be due to global models' tendency to overproduce light precipitation (Hamill and Whitaker 2006; Sun et al. 2007; Stephens et al. 2010) and the fact that a climatological forecast of little to no rain is likely hard to beat in these exceptionally dry regions. Florida also shows no improvement by the raw model and that may be a result of GEFSv12's relatively coarse resolution and inability to resolve regular afternoon thunderstorms and seabreezes.

In general, regions with a relatively high (low) skill in the raw forecast also exhibited relatively high (low) skill in the postprocessed forecasts (bottom three rows in Figs. 7–9). At week 1, the CSGD (third row in Fig. 7) and RUFCO (last row in Fig. 7) methods generally produced a more expansive improvement across the domain compared to the raw and simple bias correction methods (first and second rows in Fig. 7) during all seasons. This is demonstrated by the sparsity in the bias-corrected model's skill (row 2 Fig. 7) along the southeastern states in the cool seasons and in the interior West in the summer.

Interestingly, the bias-corrected and CSGD forecasts over the mountains in California produced more skillful forecasts than those generated with the new RUFCO model during the fall and winter. This may indicate that a more grid-point-specific method is beneficial in this region, at week 1, rather than using a convolutional approach that leverages neighboring grid points in the form of 3 × 3 or 5 × 5 kernels to inform predictions. Since the raw skill is already relatively high in this region at week 1, additional information from grid points with dissimilar bias characteristics seems to degrade the relationship between the predictors and predictand. This result over the California mountains is consistent with a study from Ghazvinian et al. (2022) who found that a quantile-mapping postprocessing approach trained with a limited set of similar grid points outperformed, over the Sierra Nevada mountains, a dense neural network, which was trained on data pooled across all grid points within CONUS.

During the summer, the CSGD and RUFCO methods are the only ones that generate an improvement over climatological forecasts over California. Recall that the CSGD method does not fit any model at particularly dry locations and instead applies ad hoc distribution parameters appropriate for dry locations (see section 4). This algorithm characteristic helps avoid overfitting and seems to be a big advantage, at week 1, over California during the summer compared to the raw and simple bias-corrected models. The FiLM layer's representation of seasonality and the use of the climatological off-ramp are likely the sources of good performance in these regions within the RUFCO model.

As expected, week 2 reveals a drastic drop in skill in the raw and postprocessed forecasts (Fig. 8) across all seasons. However, the difference between the simple bias-corrected forecasts and the more sophisticated CSGD and RUFCO methods is more pronounced compared to week 1. A dark purple region over the southwest United States sticks out in

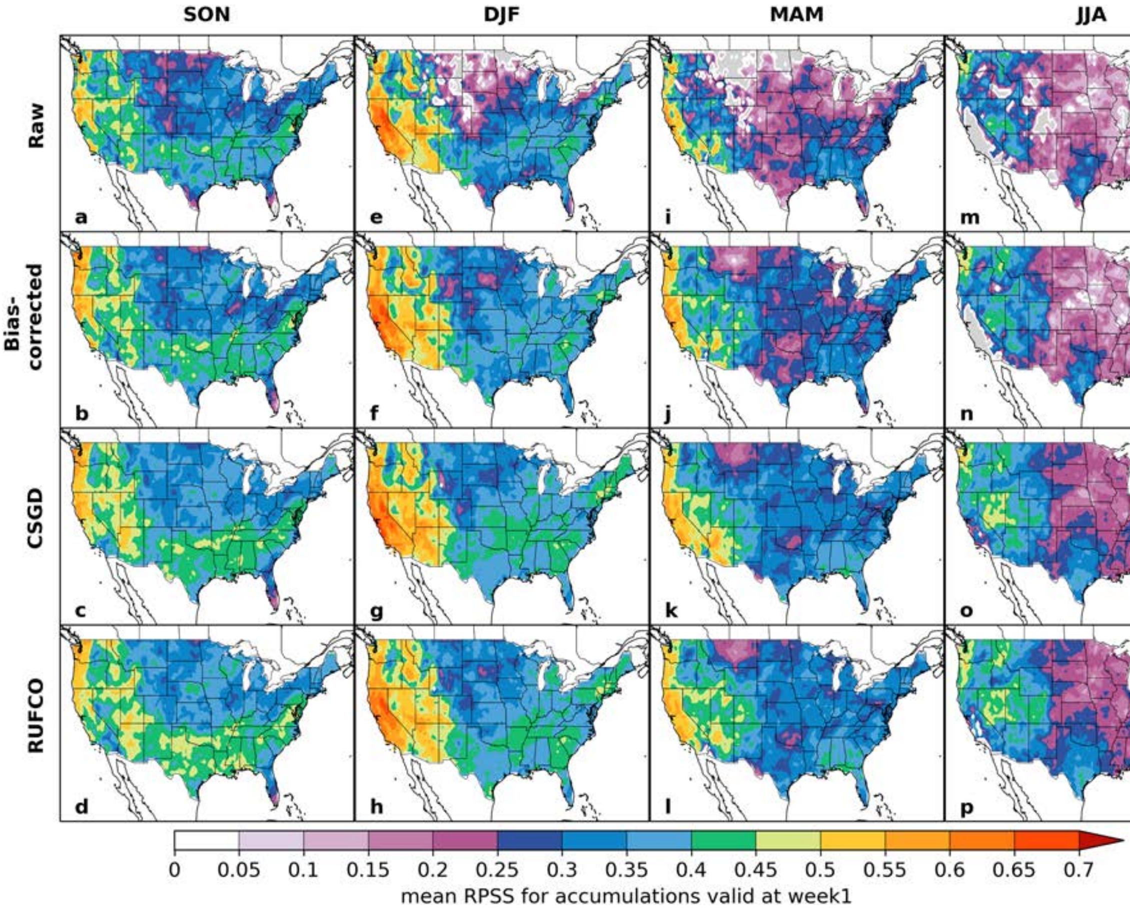

FIG. 7. Skill (RPSS) averaged over all 20 test years for week-1 forecasts (raw, bias-corrected, CSGD, and RUFCO) initialized in four meteorological seasons (a)–(d) September–November (SON), (e)–(h) DJF, (i)–(l) March–May (MAM), and (m)–(p) JJA. Colored and white shading within CONUS indicate grid boxes where the RPSS is a significant improvement over the benchmark climatological forecast at the $\alpha_{\mathrm{FDR}} = 0.1$ level. Gray shading indicates grid boxes where the improvement was not significant.

all of the summer forecast methods, which is likely related to unearthed predictability of the North American Monsoon (Prein et al. 2022). The CSGD's grid-point-specific method produces more high-skilled regions than RUFCO (e.g., northern California into Oregon); however, during all seasons, the number of statistically significant skilled grid points is greater for the RUFCO forecasts compared to any of the other forecasts (Table 2). Since displacement errors tend to increase with lead time, the inclusion of additional grid points to inform prediction may be more beneficial than single grid points. Another likely source for improved skill in the RUFCO model, especially over the western and southwest United States, is that the RUFCO model uses predictor images that expand beyond the CONUS domain (see Fig. 2). Essentially, the inclusion of the expanded domain (including over Mexico and the Pacific Ocean) may help the RUFCO model to better learn any shared covariability and/or sources of predictability that exists over the expanded regions, which can be applied to the CONUS grid points.

For the combined weeks 3–4 lead time, forecast skill reduces substantially for all methods. The raw model has no improved skill over a climatological forecast. In fact, the raw and bias-corrected forecasts exhibit negative skill compared to climatology over most of the United States (not shown).

The biggest differences between the bias-corrected, CSGD, and RUFCO methods appear at this lead time for all seasons. Aside from a few grid points in the bias-corrected forecasts (second row in Fig. 9), the CSGD (third row in Fig. 9) and RUFCO (fourth row in Fig. 9) forecasts are the only ones that show significant skill improvements compared to climatology across the domain. Compared to the CSGD, the RUFCO model again shows consistent skill across a broader area of the domain and across all seasons. The RUFCO model generates approximately 69%, 85%, 65%, and 65% statistically significant grid points in fall, winter, spring, and summer, respectively (Table 2), compared to just 12%, 33%, 16%, and 37% produced by the CSGD model for those respective seasons.

However, some of the RUFCO's expanded skill appears to come at the expense of pinpoint regions of relatively high skill that are seen in the CSGD maps but not in the RUFCO maps

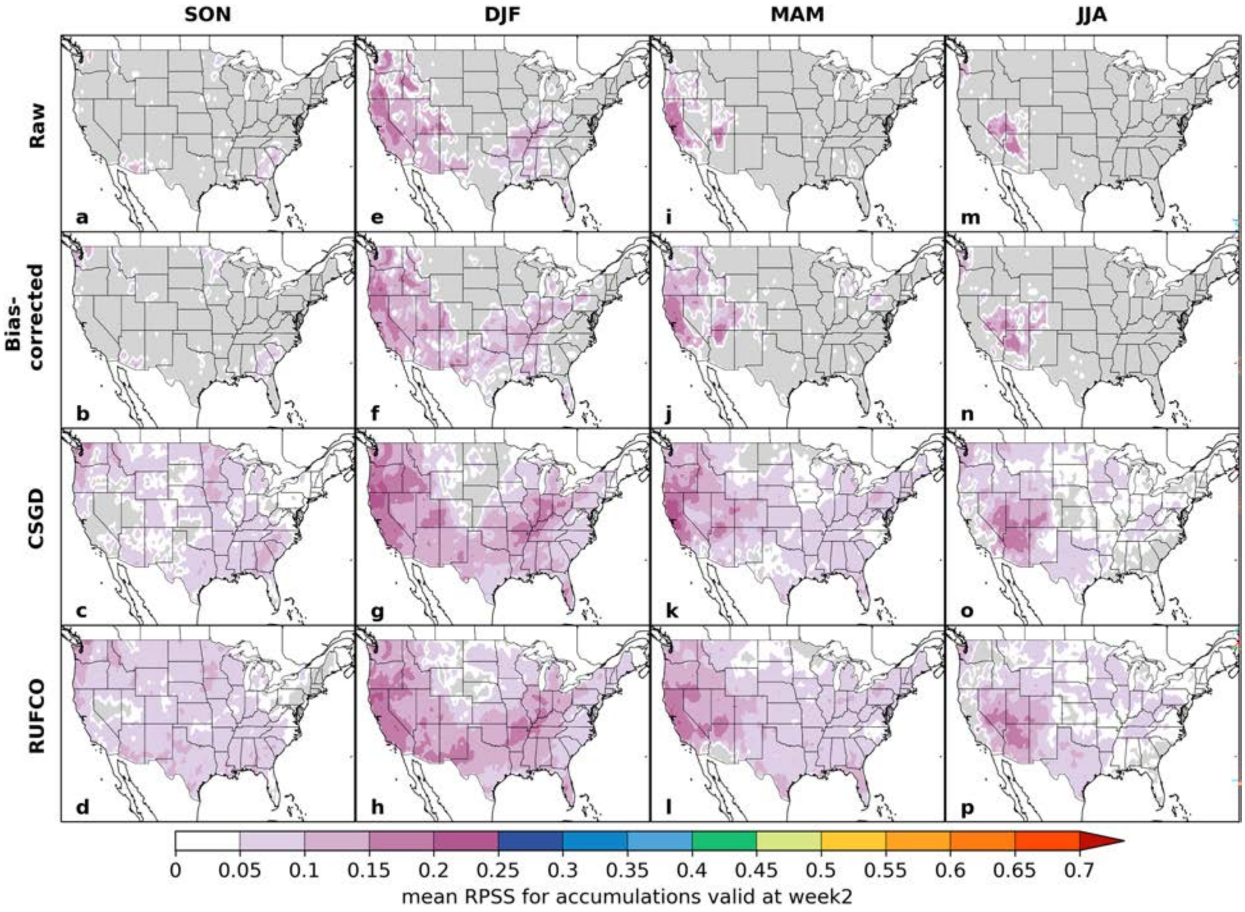

FIG. 8. As in Fig. 7, but for week 2. Note that the magnitude of the colorbar is consistent with Fig. 7 to illustrate the drop in skill from week 1 to week 2.

(e.g., red and orange areas over Texas and New Mexico in winter, yellow areas over southeast Oregon, west Montana, and central Florida peninsula in winter and orange areas over the Colorado–New Mexico border in spring and summer).

The most intriguing difference between the CSGD and RUFCO forecast performances is in winter. In addition to RUFCO's increased number of statistically significant grid points across the whole domain, the relatively high-skilled region fully surrounding the tripoint of California, Nevada, and Arizona and the southern extent of the Colorado River is more pronounced in the RUFCO forecasts than in the CSGD. This region encompasses Lake Mead, a critical reservoir that supplies surrounding states and Mexico with water; having skillful subseasonal forecasts in this region could be a big advantage of the RUFCO model for resource management. This relatively "high-skill" region is analogous to patterns found in an attribution study (Sun et al. 2022) of the skill of subseasonal monthly precipitation forecasts over North America. That study showed a strikingly similar area encompassing the southwest United States and northern Mexico, which had high predictability attributable to surface boundary conditions. The framework of RUFCO, compared to CSGD, appears to have a greater ability to unearth the covariability and/or predictable signal associated with these physically based background states that exist at weeks 3–4. This is likely, in part, to the inclusion of the expanded domain that the network learns through convolutions.

8. Case study of probabilistic categorical outlooks

Aforementioned results have shown that the RUFCO model generally produces competitive calibration and reliability across all seasons compared to traditional approaches, but does the network produce spatially realistic outlooks? We answer this question by providing example week-2 outlooks for an atmospheric river event that took place in January 2019 along the western United States. Figure 10 displays probabilistic forecasts of each method falling within the lower and upper terciles of the observed climatology as well as the observed precipitation amounts and the verified tercile assignment. The RUFCO outlook, on par with CSGD, provides physically realistic and relatively smooth probabilities that are expected at subseasonal lead times. This example shows a case where the RUFCO model well captured the higher probabilities of above-normal precipitation associated with the atmospheric river better than the CSGD forecast. All methods predicted the below-normal event in the south-central United States, but they missed

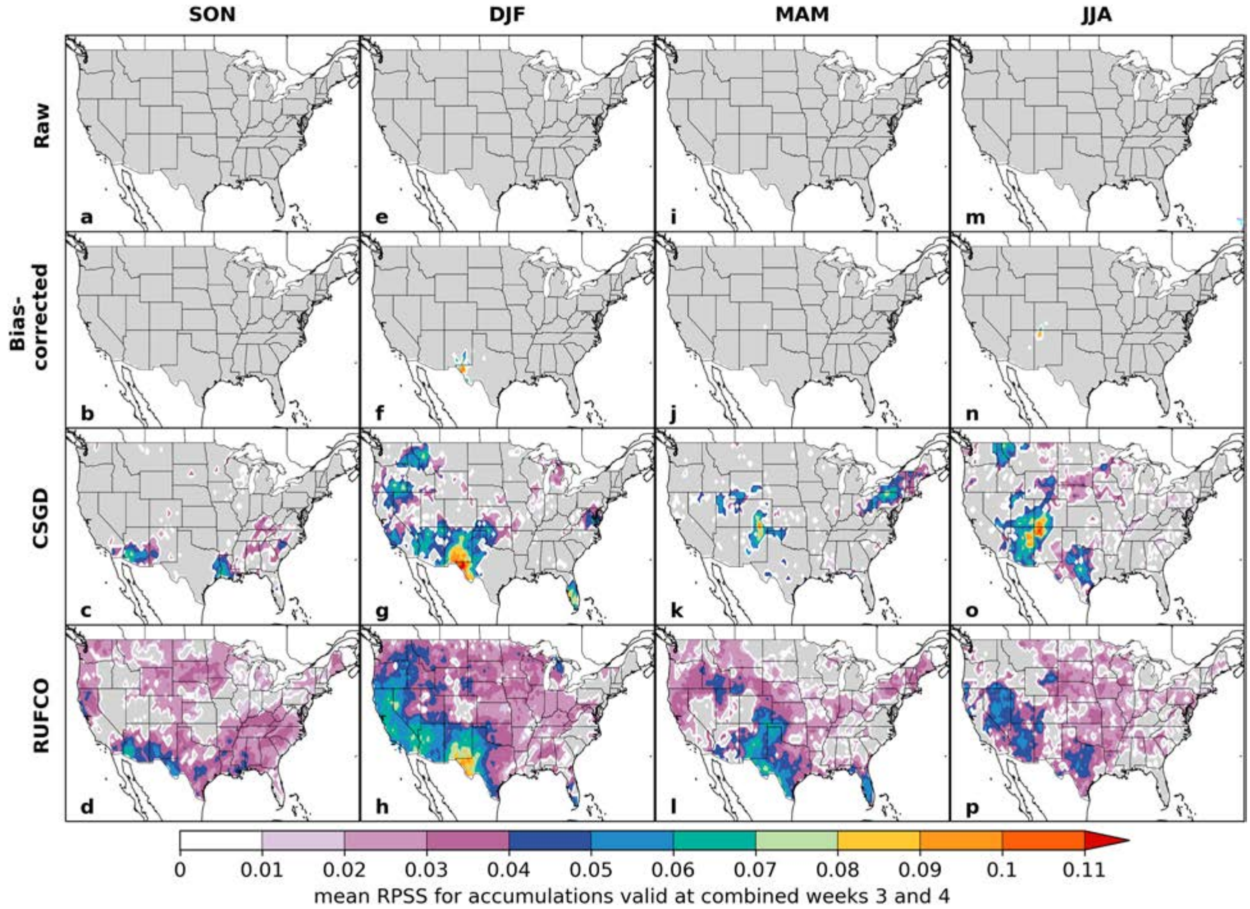

FIG. 9. As in Figs. 7 and 8, but for combined weeks 3 and 4. Note that the magnitude of the colorbar has changed from that used in Figs. 7 and 8. The raw forecasts at this lead time had no grid boxes with significant improvement over climatology (i.e., all areas are shaded gray).

the upper-tercile event (green shading in Fig. 10e stretching from southeastern Texas into Maine). RUFCO issued lower probabilities of a below-normal event in this region than the CSGD did, indicating it was less overconfident about the likelihood of a below-normal event. Readers should keep in mind that we found that the RUFCO model on average has RPSSs that are marginally higher across a wider area than the CSGD, so it is possible that any single case may perform worse. To help

TABLE 2. Number (and percentage) of CONUS grid points with statistically significant improvement over a climatological forecast. Bold values indicate the method(s) with the greatest number of statistically significant grid points for each lead time and season.

	SON	DJF	MAM	JJA
Week 1				
Raw ensemble	3318 (98.43%)	3125 (92.70%)	3001 (89.02%)	2914 (86.44%)
Bias corrected	3368 (99.91%)	**3371 (100%)**	3366 (99.85%)	3238 (96.05%)
CSGD	**3369 (99.94%)**	**3371 (100%)**	**3370 (99.97%)**	**3362 (99.73%)**
RUFCO	**3369 (99.94%)**	**3371 (100%)**	3365 (99.82%)	3351 (99.41%)
Week 2				
Raw ensemble	207 (6.14%)	1157 (34.32%)	360 (10.68%)	202 (5.99%)
Bias corrected	288 (8.54%)	1727 (51.23%)	717 (21.23%)	364 (10.80%)
CSGD	2648 (78.55%)	2938 (87.16%)	3027 (89.80%)	2553 (75.73%)
RUFCO	**3067 (90.98%)**	**3154 (93.56%)**	**3178 (94.27%)**	**2776 (82.35%)**
Weeks 3 and 4				
Raw ensemble	0 (0%)	1 (0.03%)	0 (0%)	0 (0%)
Bias corrected	0 (0%)	27 (0.80%)	1 (0.03%)	11 (0.33%)
CSGD	417 (12.37%)	1128 (33.46%)	538 (15.96%)	1232 (36.55%)
RUFCO	**2336 (69.30%)**	**2849 (84.51%)**	**2196 (65.14%)**	**2197 (65.17%)**

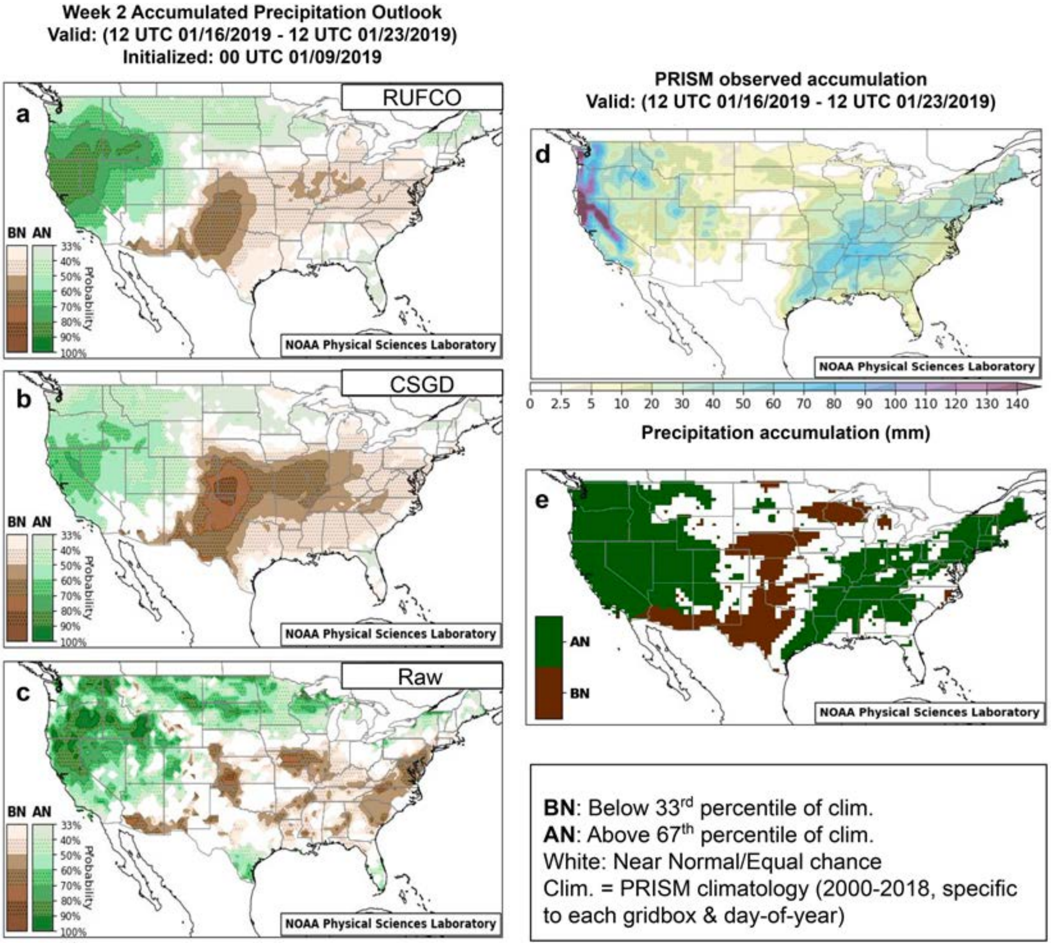
Week 2 Accumulated Precipitation Outlook
Valid: (12 UTC 01/16/2019 - 12 UTC 01/23/2019)
Initialized: 00 UTC 01/09/2019
a
RUFCO
b
CSGD
c
Raw
BN AN
33%
40%
50%
60%
70%
80%
90%
100%
Probability
NOAA Physical Sciences Laboratory
PRISM observed accumulation
Valid: (12 UTC 01/16/2019 - 12 UTC 01/23/2019)
d
0 2.5 5 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Precipitation accumulation (mm)
e
AN
BN
BN: Below 33rd percentile of clim.
AN: Above 67th percentile of clim.
White: Near Normal/Equal chance
Clim. = PRISM climatology (2000-2018, specific to each gridbox & day-of-year)

FIG. 10. (a)–(c) Week 2 probabilistic above-normal (AN; green shading) and below-normal (BN; brown shading) categorical outlooks for RUFCO, CSGD, and raw forecasts initialized on 9 Jan 2019 and valid for dates between 16 and 23 Jan 2019. (d) The verified observed precipitation amount and (e) the above- or below-normal category that the verifying observation belonged to. CSGD and RUFCO forecasts are based on their respective algorithms trained on data from years 2000–18. Figures of the threshold amounts are provided in supplemental E.

identify impactful cases for future research, NOAA's Physical Sciences Laboratory is running the RUFCO model experimentally in near–real time.

9. Summary and discussion

Three methods are presented to postprocess probabilistic forecasts of the subseasonal precipitation accumulation output from the Global Ensemble Forecast System: 1) a simple bias-corrected forecast achieved by adjusting exceedance thresholds based on the raw model's own climatology, 2) a parametric and established distributional regression-type model (CSGD), and 3) a new nonparametric, deep learning framework (RUFCO) consisting of three unique components tailored for effective subseasonal prediction.

RUFCO's first component is a ResUnet, which uses several layers of convolutions, nonlinear activation functions, and dimensionality reduction operations to predict probabilities of precipitation falling into precipitation bins. The second is the "FiLM layer" which conditions the network on the month of the year, allowing one network to be used for the entire year. A sensitivity experiment showcased the FiLM's ability to learn and condition a network on the month of the year, which resulted in improved Brier scores in regions that exhibit strong seasonality in precipitation amounts. The final component is the climatological off-ramp which was first demonstrated by S20 and ultimately helps the network revert toward climatological probabilities as forecast skill drops to zero. Another unique aspect of the RUFCO model compared to most traditional postprocessing algorithms is its ability to train one model and make predictions simultaneously across all grid points rather than having to train a separate corrective model for each grid point.

We demonstrated that the RUFCO model is an effective and competitive postprocessing algorithm. Its inputs included transformed versions of the raw ensembles of accumulated precipitation, geopotential height at 500 hPa, and total column water as well as geographic and temporal predictors (latitude, longitude, elevation, and month of the year). Reliability diagrams of above- and below-normal events during winter and summer were on par with the CSGD method and showed nearly perfect reliability for weeks 1 and 2 except for the CSGD which issued overconfident forecasts at week 2 in the summer months. Reliability was greatly improved for weeks 3–4 compared to the raw and bias-corrected forecasts.

At week 1, all of the postprocessing methods showed an overall improvement to the raw forecast and therefore would be valid postprocessing options. However, there were some notable trade-offs of each method. The CSGD and RUFCO models were the only methods that produced skillful week-1 forecasts over California during summer, with the best performance coming from the CSGD method. This is likely a result of its prescription of defined distribution parameters at superdry locations rather than (over) fitting a distribution. Even with the overall competitive performance, the RUFCO model failed to produce the highest skillful week-1 forecasts over some mountainous regions in the West. This result is likely due to the learning process of the network, which learns convolutions for the whole domain. In regions with strong regionally specific biases, learned convolutions (filters) seem to be a disadvantage compared to the grid-point-specific techniques.

Conversely, the approach of the RUFCO model became an advantage at week 2 and weeks 3–4. It is at these lead times that the RUFCO model is most differentiated from the other methods in terms of the number of grid points that exhibit statistically significant improvements over climatology. At week 2, RUFCO generated a statistically significant coverage across 82%–94% of the grid points. CSGD generated forecasts that had statistically significant improvements compared to climatology across only 76%–90% of the domain. The disparity is even larger at weeks 3 and 4; RUFCO produced 65%–85%, while CSGD produced 12%–37% statistically significant grid points. The ease of the RUFCO model to learn from a domain that expands beyond the target CONUS domain is likely a contributing factor to its improved performance across the United States, especially along the western and southwestern United States.

We performed extensive hyperparameter tuning and studied which parameters the model is most sensitive to. Some of these insights may be relevant for ResUnet-based postprocessing approaches beyond the specific application studied here.

Lastly, the architecture of the RUFCO framework can be easily extended to add more image predictors such as atmospheric stability, which may help improve performance for convective precipitation events. We hypothesize that incorporating the FiLM layer into each level of the ResUnet may provide even stronger network conditioning to the FiLM predictors. This may be a worthwhile venture for applications using data with higher signal-to-noise ratios and/or larger training datasets to avoid overfitting. The FiLM component also offers ample possibilities to condition the network on climate indices or principal components related to the Madden–Julian oscillation (MJO), North Atlantic Oscillation (NAO), El Niño–Southern Oscillation (ENSO), and sudden stratospheric warming (SSW), which are known contributors of subseasonal midlatitude predictability (Vitart et al. 2012; Hoell et al. 2016; Lang et al. 2020; Merryfield et al. 2020). Additionally, association with certain weather regimes or flow-dependent states can result in more skillful precipitation forecasts (Moore 2023; Lee et al. 2023), even at subseasonal scales (DelSole et al. 2017). Future studies will explore the benefits of explicitly incorporating these additional predictors, which the flexibility of the proposed architecture affords.

Acknowledgments. We gratefully acknowledge three anonymous reviewers and Dr. Kinya Toride (NOAA and the Cooperative Institute for Research in Environmental Sciences) for comments that helped improve this paper. We thank the NOAA Physical Sciences Laboratory and the NOAA/Weather Program Office via the NOAA Cooperative Agreement NA21OAR4590176 for their support. Additionally, we thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC number 2064/1—Project Number 390727645 for supporting J. Schlör.

Data availability statement. GEFSv12 reforecast data are publicly available for download from Amazon Web Services at https://registry.opendata.aws/noaa-gefs-reforecast. Daily gridded PRISM precipitation data are publicly available for download from the PRISM Climate Group website at https://prism.oregonstate.edu/recent/. The RUFCO network was developed in Python using the Keras (Chollet et al. 2018) deep learning application programming interface (API) atop of Tensorflow (Abadi et al. 2015) 2.8.

APPENDIX

RUFCO Architecture and Hyperparameter Tuning

a. Architecture

Table A1 describes each of the operations performed within the RUFCO network as illustrated in Fig. 3.

b. Tuning results

The number of times each candidate HP value in Table A2 was selected as the optimal choice for use in each of the 20 optimized networks is illustrated in the bar charts in Fig. A1. Two main findings are that the model is only sensitive to some HP values and that the optimal choice can vary with lead time. For each lead time, the networks preferred a smaller number of batches. A smaller number of batches translates to passing the network a larger sample size to train and update weights within an epoch. Smaller batch sizes typically result in noisy optimization, which can be overall beneficial for the generalization of the model. However, given the low signal-to-noise ratio inherent in subseasonal forecasts, larger batch sizes may have been preferred to avoid additional stochasticity during training.

Nearly all optimized models had a learning rate in a narrow range between 0.001 and 0.01, with weeks 1 and 2 having a preference toward the higher end of that range. The

TABLE A1. Abbreviations and corresponding descriptions of elements of the RUFCO network shown in Fig. 3. "Specified" refers to a defined input into the parentheses.

Abbreviation	Description
Conv (kernel size), *N* filters	2D convolution performed with "same" padding, stride = 1, and specified kernel size and *N* number of filters
ReLU	Rectified linear unit activation function
Linear	Linear activation function
Softmax	Softmax activation function
BN	Batch normalization
Dropout (rate)	Dropout layer with specified dropout rate
MaxPool (pool size)	2D MaxPooling with "same" padding, stride = 2, and specified pool size
Upsamp (size)	2D upsampling performed with specified size
Dense, *N*-neurons	Dense layer with a specified *N*-number of neurons
FiLM equation	Final equation output from the FiLM generator [refer to section 3c(2)]
res. connec.	Residual connection [refer to section 3c(1)]
Log clim. probabilities	Natural logarithm of the climatological forecast probabilities for each bin [refer to section 3c(3)]

TABLE A2. HPs and their candidate values sampled by Optuna (Akiba et al. 2019) to optimize the RUFCO network (see Fig. 3). The *nfilm* is 12 and corresponds to the size of the one-hot encoded month vector input to the FiLM layer. The *ncl* is 20 and corresponds to the number of classes/bins. The batch size is determined by dividing the training dataset by the number of batches.

HP	Candidate values
Number of batches	[10, 20, 30]
Learning rate	[0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05]
Number of nodes input to the FiLM layer	[*nfilm**5, *nfilm**10, *nfilm**15]
Number of nodes in the FiLM hidden layer	[*ncl**3, *ncl**5, *ncl**10]
L1 regularization in FiLM layers	[0.000 001, 0.000 01, 0.0001, 0.001, 0.01]
Number of ResBlocks	[2, 3, 4]
Number of filters in the first ResBlock	[8, 16, 32]
(Convolution) kernel size	[3 × 3, 5 × 5]
Dropout rate	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5]

network was sensitive to the number of residual blocks, number of filters learned in the first residual block, convolution kernel size, and dropout rate. For week 1 and week 2, the largest number of filters tested (32) was preferred. Week-3 and -4 lead times were evenly split on either 16 or 32 filters. The smallest number of filters (8) was only selected once across all optimized networks and lead times, indicating a clear preference for more filters to learn complex features within the data. Generally fewer blocks (2) were needed to learn features in the data for week 1 compared to that of week 2 and weeks 3 and 4, which mostly used four blocks. A larger kernel size (5 × 5 vs 3 × 3) was the optimal choice for week 2 and weeks 3 and 4, while either kernel size is suitable for week 1. Finally, the smallest dropout rates were selected for week 1, with 0.3 being the most frequently used dropout rate. Optimal rates for week 2 and weeks 3 and 4 varied but were generally higher than those selected for week 1. Since dropout layers help the network prevent overfitting, smaller (larger) rates at week 1 (week 2 and weeks 3 and 4) may indicate that the network had fewer (more) issues with overfitting. The networks were not as sensitive to the choice of nodes in the individual FiLM layers, but we found that the total number of nodes within the FiLM's dense layers ranged from 200 to 300 (out of a possible range of 160–420) in the majority of the networks (not shown), indicating a robustness of the number of FiLM neurons across lead times. This result may be in part a consequence of the inclusion of an L1 regularization parameter within the FiLM generator, which can help reduce overfitting in the event of too many parameters. Week 2 and weeks 3 and 4 overall had larger L1 values than week 1, indicating that more regularization was needed for the longer lead times.

It is likely that more optimal networks can be attained by tuning over a larger pool of candidate HPs. However, we believe that the sensitivity analysis performed in this study provides valuable insight into lead-time dependencies of the HPs and reveals which HPs could be leveraged even more to potentially gain greater performance (e.g., increasing the number of filters).

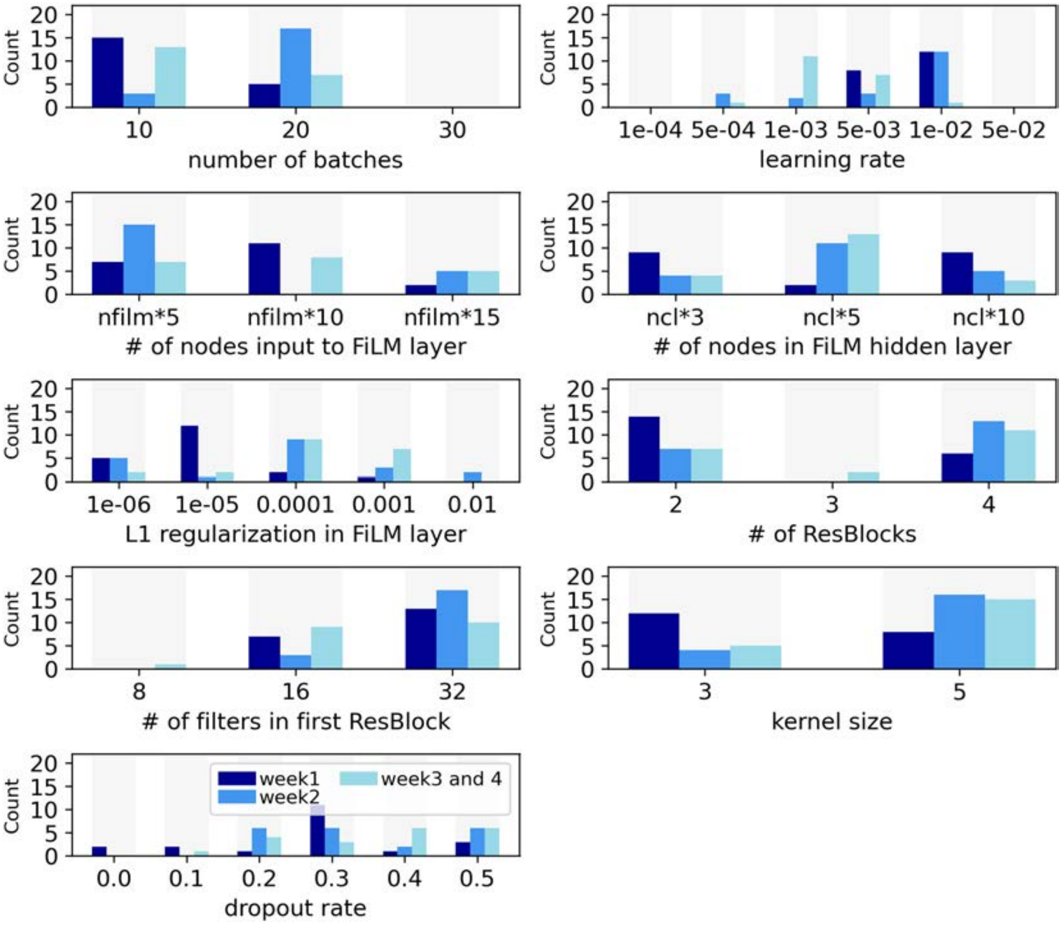

FIG. A1. Bar charts for each lead time indicating the number of times the candidate choices (*x* axes) for each HP (number of batches, learning rate, number of nodes input to the FiLM layer, number of nodes in FiLM hidden layer, L1 regularization in the FiLM layers, number of ResBlocks, number of filters in the first ResBlock, kernel size, and dropout rate) were selected as the optimal HP choice. The candidate choices correspond to those listed in Table A2 and are used in the RUFCO model. Each choice could be selected up to 20 times since 20 networks were optimized for use with the 20 cross-validated test years. Counts close to 20 indicate that the majority of the trained networks selected that choice; low counts indicate it was not a popular choice to achieve an optimal network. Blue shades indicate the forecast lead time listed in the legend, and gray background shading simply separates the choices. Widths of the bars have no meaning and were only adjusted to conserve space.

REFERENCES

Abadi, M., and Coauthors, 2015: Tensorflow: A system for large-scale machine learning. *Proc. 12th USENIX Symp. on Operating Systems Design and Implementation (OSDI'16)*, Savannah, GA, USENIX, 265–283, https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Abatzoglou, J. T., D. J. McEvoy, N. J. Nauslar, K. C. Hegewisch, and J. L. Huntington, 2023: Downscaled subseasonal fire danger forecast skill across the contiguous United States. *Atmos. Sci. Lett.*, **24**, e1165, https://doi.org/10.1002/asl.1165.

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama, 2019: Optuna: A next-generation hyperparameter optimization framework. *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Anchorage, AK, Association for Computing Machinery, 2623–2631, https://doi.org/10.1145/3292500.3330701.

Amazon Web Services, 2023: NOAA's Global Ensemble Forecast System version 12: Reforecast data storage information. AWS NOAA Tech. Doc., 7 pp., https://noaa-gefs-retrospective.s3.amazonaws.com/Description_of_reforecast_data.pdf.

Badrinath, A., L. D. Monache, N. Hayatbini, W. Chapman, F. Cannon, and M. Ralph, 2023: Improving precipitation forecasts with convolutional neural networks. *Wea. Forecasting*, **38**, 291–306, https://doi.org/10.1175/WAF-D-22-0002.1.

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

Boureau, Y.-L., J. Ponce, and Y. LeCun, 2010: A theoretical analysis of feature pooling in visual recognition. *Proc. 27th Int. Conf. on Machine Learning (ICML-10)*, Haifa, Israel, Association for Computing Machinery, 111–118, https://dl.acm.org/doi/10.5555/3104322.3104338.

Breeden, M. L., J. R. Albers, and A. Hoell, 2022: Subseasonal precipitation forecasts of opportunity over central southwest Asia. *Wea. Climate Dyn.*, **3**, 1183–1197, https://doi.org/10.5194/wcd-3-1183-2022.

Chapman, W. E., L. D. Monache, S. Alessandrini, A. C. Subramanian, F. M. Ralph, S.-P. Xie, S. Lerch, and N. Hayatbini, 2022: Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Mon. Wea. Rev.*, **150**, 215–234, https://doi.org/10.1175/MWR-D-21-0106.1.

Chollet, F., and Coauthors, 2018: Keras: The Python deep learning library. Astrophysics Source Code Library ASCL-1806, accessed 1 December 2023, https://ascl.net/1806.022.

Daly, C., G. H. Taylor, and W. P. Gibson, 1997: The PRISM approach to mapping precipitation and temperature. *10th AMS Conf. on Applied Climatology*, Reno, NV, Amer. Meteor. Soc., 10–12, https://idfg.idaho.gov/species/bibliography/1500028.

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.

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.

Fan, Y., V. Krasnopolsky, H. van den Dool, C.-Y. Wu, and J. Gottschalck, 2023: Using artificial neural networks to improve CFS week-3–4 precipitation and 2-m air temperature forecasts. *Wea. Forecasting*, **38**, 637–654, https://doi.org/10.1175/WAF-D-20-0014.1.

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

——, ——, T. M. Hamill, D.-J. Seo, and N. Fernando, 2022: Improving probabilistic quantitative precipitation forecasts using short training data through artificial neural networks. *J. Hydrometeor.*, **23**, 1365–1382, https://doi.org/10.1175/JHM-D-22-0021.1.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

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

——, ——, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

——, and Coauthors, 2022: The reanalysis for the Global Ensemble Forecast System, version 12. *Mon. Wea. Rev.*, **150**, 59–79, https://doi.org/10.1175/MWR-D-21-0023.1.

——, D. R. Stovern, and L. L. Smith, 2023: Improving National Blend of Models probabilistic precipitation forecasts using long time series of reforecasts and precipitation reanalyses. Part I: Methods. *Mon. Wea. Rev.*, **151**, 1521–1534, https://doi.org/10.1175/MWR-D-22-0308.1.

Hara, K., D. Saito, and H. Shouno, 2015: Analysis of function of rectified linear unit used in deep learning. *2015 Int. Joint Conf. on Neural Networks (IJCNN)*, Killarney, Ireland, Institute of Electrical and Electronics Engineers, 1–8, https://doi.org/10.1109/IJCNN.2015.7280578.

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

He, K., X. Zhang, S. Ren, and J. Sun, 2016: Deep residual learning for image recognition. *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, Institute of Electrical and Electronics Engineers, 770–778, https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Hill, A. J., R. S. Schumacher, and I. L. Jirak, 2023: A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. *Wea. Forecasting*, **38**, 251–272, https://doi.org/10.1175/WAF-D-22-0143.1.

Hoell, A., M. Hoerling, J. Eischeid, K. Wolter, R. Dole, J. Perlwitz, T. Xu, and L. Cheng, 2016: Does El Niño intensity matter for California precipitation? *Geophys. Res. Lett.*, **43**, 819–825, https://doi.org/10.1002/2015GL067102.

——, and Coauthors, 2020: Lessons learned from the 2017 flash drought across the U.S. northern Great Plains and Canadian Prairies. *Bull. Amer. Meteor. Soc.*, **101**, E2171–E2185, https://doi.org/10.1175/BAMS-D-19-0272.1.

Horat, N., and S. Lerch, 2024: Deep learning for postprocessing global probabilistic forecasts on subseasonal time scales. *Mon. Wea. Rev.*, **152**, 667–687, https://doi.org/10.1175/MWR-D-23-0150.1.

Hu, W., M. Ghazvinian, W. E. Chapman, A. Sengupta, F. M. Ralph, and L. D. Monache, 2023: Deep learning forecast uncertainty for precipitation over the western United States. *Mon. Wea. Rev.*, **151**, 1367–1385, https://doi.org/10.1175/MWR-D-22-0268.1.

Ioffe, S., and C. Szegedy, 2015: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv, 1502.03167v3, https://doi.org/10.48550/arXiv.1502.03167.

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.

Jones, P. W., 1999: First- and second-order conservative remapping schemes for grids in spherical coordinates. *Mon. Wea. Rev.*, **127**, 2204–2210, https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2.

Lagerquist, R., D. D. Turner, I. Ebert-Uphoff, and J. Q. Stewart, 2023: Estimating full longwave and shortwave radiative transfer with neural networks of varying complexity. *J. Atmos. Oceanic Technol.*, **40**, 1407–1432, https://doi.org/10.1175/JTECH-D-23-0012.1.

Lang, A. L., K. Pegion, and E. A. Barnes, 2020: Introduction to special collection: "Bridging weather and climate: Subseasonal-to-Seasonal (S2S) prediction". *J. Geophys. Res. Atmos.*, **125**, e2019JD031833, https://doi.org/10.1029/2019JD031833.

Lee, S. H., M. K. Tippett, and L. M. Polvani, 2023: A New year-round weather regime classification for North America. *J. Climate*, **36**, 7091–7108, https://doi.org/10.1175/JCLI-D-23-0214.1.

Long, J., E. Shelhamer, and T. Darrell, 2015: Fully convolutional networks for semantic segmentation. *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, Institute of Electrical and Electronics Engineers, 3431–3440, https://doi.org/10.1109/CVPR.2015.7298965.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21A**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Merryfield, W. J., and Coauthors, 2020: Current and emerging developments in subseasonal to decadal prediction. *Bull. Amer. Meteor. Soc.*, **101**, E869–E896, https://doi.org/10.1175/BAMS-D-19-0037.1.

Moore, B. J., 2023: Flow dependence of medium-range precipitation forecast skill over California. *Wea. Forecasting*, **38**, 699–720, https://doi.org/10.1175/WAF-D-22-0081.1.

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.

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.

Pendergrass, A. G., and Coauthors, 2020: Flash droughts present a new challenge for subseasonal-to-seasonal prediction. *Nat. Climate Change*, **10**, 191–199, https://doi.org/10.1038/s41558-020-0709-0.

Perez, E., F. Strub, H. de Vries, V. Dumoulin, and A. Courville, 2017: FiLM: Visual reasoning with a general conditioning layer. arXiv, 1709.07871v2, https://doi.org/10.48550/arXiv.1709.07871.

Prein, A. F., E. Towler, M. Ge, D. Llewellyn, S. Baker, S. Tighi, and L. Barrett, 2022: Sub-seasonal predictability of North American Monsoon precipitation. *Geophys. Res. Lett.*, **49**, e2021GL095602, https://doi.org/10.1029/2021GL095602.

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.

Roberts, N., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteor. Appl.*, **15**, 163–169, https://doi.org/10.1002/met.57.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI, 2015*, N. Navab et al., Eds., Lecture Notes in Computer Science, Vol. 9351, Springer, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

Saha, S., 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.

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

——, M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Slingo, J., and T. Palmer, 2011: Uncertainty in weather and climate prediction. *Philos. Trans. Roy. Soc.*, **A369**, 4751–4767, https://doi.org/10.1098/rsta.2011.0161.

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.

Stephens, G. L., and Coauthors, 2010: Dreary state of precipitation in global models. *J. Geophys. Res.*, **115**, D24211, https://doi.org/10.1029/2010JD014532.

Sun, L., M. P. Hoerling, J. H. Richter, A. Hoell, A. Kumar, and J. W. Hurrell, 2022: Attribution of North American subseasonal precipitation prediction skill. *Wea. Forecasting*, **37**, 2069–2085, https://doi.org/10.1175/WAF-D-22-0076.1.

Sun, Y., S. Solomon, A. Dai, and R. W. Portmann, 2007: How often will it rain? *J. Climate*, **20**, 4801–4818, https://doi.org/10.1175/JCLI4263.1.

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

Vitart, F., 2004: Monthly forecasting at ECMWF. *Mon. Wea. Rev.*, **132**, 2761–2779, https://doi.org/10.1175/MWR2826.1.

——, A. W. Robertson, and D. L. T. Anderson, 2012: Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *WMO Bull.*, **61**, 23–28.

——, and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Worsnop, R. P., M. Scheuerer, F. D. Giuseppe, C. Barnard, T. M. Hamill, and C. Vitolo, 2021: Probabilistic fire danger forecasting: A framework for week-two forecasts using statistical postprocessing techniques and the Global ECMWF Fire Forecast system (GEFF). *Wea. Forecasting*, **36**, 2113–2125, https://doi.org/10.1175/WAF-D-21-0075.1.

Yuan, X., E. F. Wood, and M. Liang, 2014: Integrating weather and climate prediction: Toward seamless hydrologic forecasting. *Geophys. Res. Lett.*, **41**, 5891–5896, https://doi.org/10.1002/2014GL061076.

Zafar, A., M. Aamir, N. M. Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta, and S. Almotairi, 2022: A comparison of pooling methods for convolutional neural networks. *Appl. Sci.*, **12**, 8643, https://doi.org/10.3390/app12178643.

Zeiler, M. D., and Coauthors, 2013: On rectified linear units for speech processing. *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, British Columbia, Canada, Institute of Electrical and Electronics Engineers, 3517–3521, https://doi.org/10.1109/ICASSP.2013.6638312.

Zhang, F., Y. Q. Sun, L. Magnusson, R. Buizza, S.-J. Lin, J.-H. Chen, and K. Emanuel, 2019: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

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.

Zhang, Z., Q. Liu, and Y. Wang, 2018: Road extraction by deep residual U-Net. *IEEE Geosci. Remote Sens. Lett.*, **15**, 749–753, https://doi.org/10.1109/LGRS.2018.2802944.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Zhuang, J., and Coauthors, 2023: Pangeo-data/xESMF: v0.8.2. Zenodo, accessed 1 December 2023, https://doi.org/10.5281/zenodo.8356796.

Zsótér, E., 2006: Recent developments in extreme weather forecasting. *ECMWF Newsletter*, No. 107, ECMWF, Reading, United Kingdom, 8–17, https://www.ecmwf.int/sites/default/files/elibrary/2006/17958-recent-developments-extreme-weather-forecasting.pdf.