Understanding Predictability of Daily Southeast U.S. Precipitation Using Explainable Machine Learning

KATHY PEGION,[a,b] EMILY J. BECKER,[c] AND BEN P. KIRTMAN[c]

[a] *Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, Virginia*
[b] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[c] *Cooperative Institute for Marine and Atmospheric Studies, Rosenstiel School, University of Miami, Miami, Florida*

(Manuscript received 7 February 2022, in final form 16 September 2022)

ABSTRACT: We investigate the predictability of the sign of daily southeastern U.S. (SEUS) precipitation anomalies associated with simultaneous predictors of large-scale climate variability using machine learning models. Models using index-based climate predictors and gridded fields of large-scale circulation as predictors are utilized. Logistic regression (LR) and fully connected neural networks using indices of climate phenomena as predictors produce neither accurate nor reliable predictions, indicating that the indices themselves are not good predictors. Using gridded fields as predictors, an LR and convolutional neural network (CNN) are more accurate than the index-based models. However, only the CNN can produce reliable predictions that can be used to identify forecasts of opportunity. Using explainable machine learning we identify which variables and grid points of the input fields are most relevant for confident and correct predictions in the CNN. Our results show that the local circulation is most important as represented by maximum relevance of 850-hPa geopotential heights and zonal winds to making skillful, high-probability predictions. Corresponding composite anomalies identify connections with El Niño–Southern Oscillation during winter and the Atlantic multidecadal oscillation and North Atlantic subtropical high during summer.

KEYWORDS: Precipitation; Climate prediction; Climate models; Climate variability; Subseasonal variability

1. Introduction

Predicting when, where, and how much precipitation will fall is critical for decision-making and responding to the impacts of extreme rainfall across a wide range of sectors and time scales (Balmaseda et al. 2020). The predictability of subseasonal-to-seasonal (S2S) precipitation stems from many different sources, varies by season, and is changing with anthropogenic climate change (Balmaseda et al. 2020). S2S prediction of total precipitation, particularly in the extratropics and during the warm season by state-of-the-art coupled Earth system models (ESMs) is overall very poor, because the current generation of ESMs have large biases that profoundly limit the ability to simulate and predict the circulation patterns that lead to much of the spatial–temporal variability of precipitation (Balmaseda et al. 2020).

The skill of predicting even total weekly and monthly precipitation anomalies over North America remains poor (e.g., Pegion et al. 2019; Becker et al. 2020). To add to the challenge, actionable predictions require the ability to predict precipitation and its extremes at higher temporal resolution and at regional scales. Prolonged dry spells or the rains from a single event can have devastating impacts.

In particular, S2S precipitation variability in the southeastern United States (SEUS) has significant impacts on water resources and agriculture in the region. Figure 1b shows the daily precipitation anomalies in the SEUS region (Fig. 1a) for December–February (DJF) and June–August (JJA). Daily precipitation over a large region such as the SEUS is impacted by local factors such as low-level jets, diurnal variability, and soil moisture, among others. In addition to local sources, a suite of large-scale climate mechanisms across a range of time scales have been identified as potential sources of variability and predictability of North American precipitation, which includes the SEUS, on S2S time scales (Table 1). This paper explores the contribution of these large-scale sources of predictability to SEUS precipitation.

The combined impact of large-scale SST anomalies (SSTA) in the Pacific and Atlantic Oceans on monthly and annual precipitation anomalies in North America is well documented (e.g., Schubert et al. 2009). Specifically, warm Pacific SSTAs and cold Atlantic SSTAs are associated with greater than normal precipitation while cold Pacific and warm Atlantic SSTAs are associated with less than normal precipitation. The Pacific SST anomalies primarily impact North American precipitation during winter through changes in circulation patterns associated with El Niño–Southern Oscillation (ENSO) (Ropelewski and Halpert 1986). In the SEUS, these circulation changes lead to an increased number of winter storms in the land areas bordering the Gulf of Mexico and Atlantic Oceans during El Niño winters and a decreased number of storms in the regions bordering the Gulf of Mexico during La Niña winters (Schubert et al. 2009). On subseasonal time scales, the Madden–Julian oscillation (MJO) impacts cold season precipitation frequency and intensity over North America through the extension and contraction of the midlatitude jets (Stan et al. 2017; Becker et al. 2011; Zhou et al. 2011). Large-scale circulation anomalies associated with preferred patterns of winter midlatitude variability on S2S time scales known as weather regimes in the Pacific–North America region have also been shown to impact moisture flux, atmospheric rivers, and precipitation over North

Corresponding author: Kathy Pegion, kpegion@ou.edu

DOI: 10.1175/AIES-D-22-0011.1 e220011

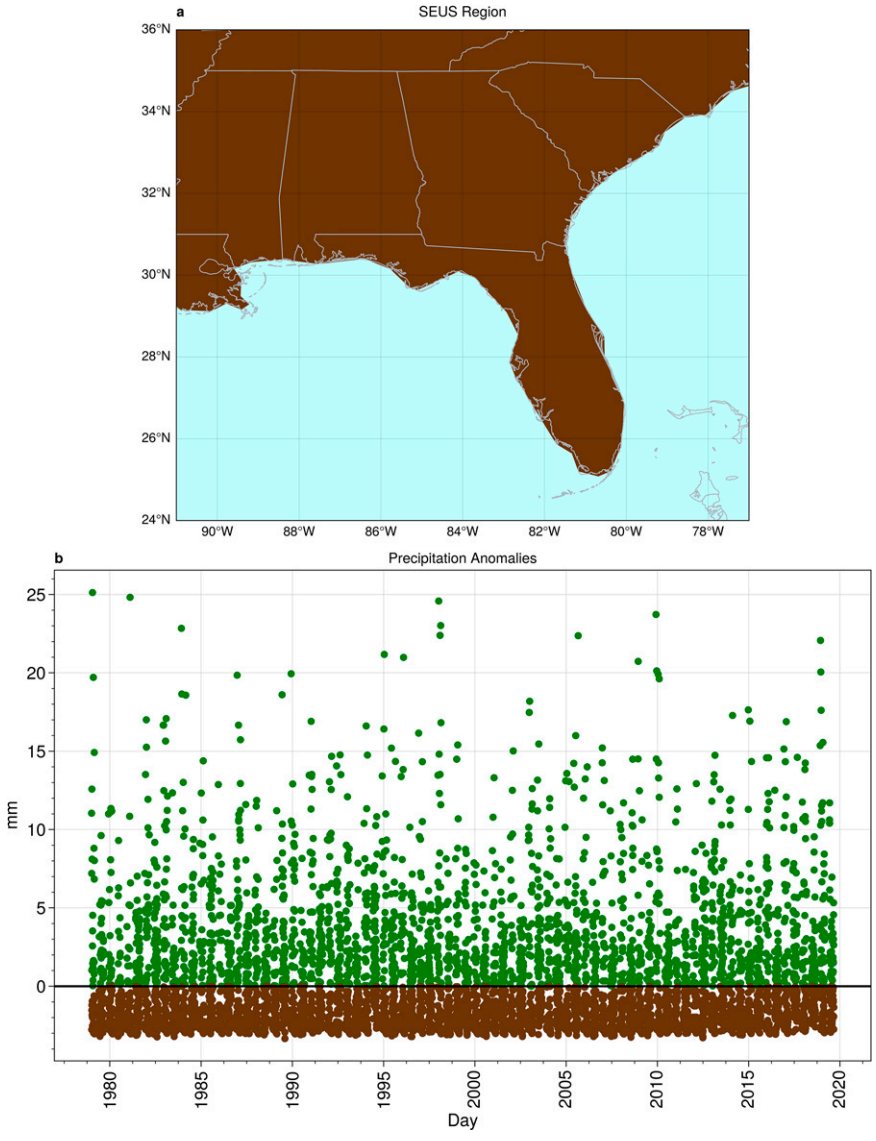

FIG. 1. (a) Map of the southeastern U.S. region, and (b) average precipitation anomalies (mm) over land areas in this region for each day in DJF and JJA. Green indicates positive anomalies, and brown indicates negative anomalies. During DJF there are 1096 positive anomalies and 2573 negative anomalies. During JJA there are 1482 positive anomalies and 2290 negative anomalies.

America (Amini and Straus 2019; Robertson et al. 2020; Cassou 2008). Recently, Stan and Krishnamurthy (2019) identified large-scale periodic midlatitude variability that impacts North American precipitation, including a 120-day midlatitude seasonal oscillation (MLSO).

The impact of large-scale climate variability on summer precipitation in the SEUS is much less explored and understood. Variability of Atlantic SST anomalies is thought to affect summer SEUS precipitation through the Atlantic multidecadal oscillation (AMO) and its impact on the variability of the North Atlantic subtropical high (NASH) (Li et al. 2019, 2012). See also Zhang et al. (2022) for a discussion of how the NASH impacts decadal SEUS rainfall. When the AMO is anomalously cold, the NASH is stronger and shifted westward, with large meridional extension that impacts most of North America (Hu et al. 2011). During the warm phase of the AMO, the NASH is contracted and shifts northeast of its climatological position (Hu et al. 2011). When the NASH is shifted northwest of its climatological position, the SEUS is anomalously dry during summer, while a southwest shift of the NASH is associated with anomalously wet summers in the SEUS (Li et al. 2012). The strength and sign of the MLSO is also identified as impacting summer SEUS precipitation, although the physical mechanisms are not well understood (Manthos et al. 2022). The impact of summer weather regimes over the Pacific–North America region has also not received much attention.

TABLE 1. Table of index predictors.

Predictor	Values	Time scale	Variables	Source
AMO	Continuous index	Monthly	SST	NOAA/ESRL/PSL
PDO	Continuous index	Monthly	SST	NOAA/ESRL/PSL
Niño-3.4	Continuous index	Monthly	SST	NOAA/ESRL/PSL
NAO	Continuous index	Monthly	Z500	NOAA/ESRL/PSL
MJO phase	Integer 0–7	Daily	U200, U850, and OLR	CAWCR
MJO amplitude	Continuous Index	Daily	U200, U850, and OLR	CAWCR
NASH phase	Integer 0–3	Daily	Z850 and U850	Calculated from ERA-Interim
NASH amplitude	Continuous index	Daily	Z850	Calculated from ERA-Interim
PNA weather regimes	Integer 0–5	Daily	Z500 and U250	Calculated from ERA-Interim
MLSO	Continuous index	Daily	Z500	Calculated from ERA-Interim

Previous studies have investigated the large-scale climate factors described above individually or with at most two in combination; also, the emphasis has primarily been on the winter season or annual precipitation, which is dominated by the winter season (e.g., Arcodia et al. 2020; Manthos et al. 2022). Because of the covariability of large-scale climate factors, it is challenging to disentangle how their combinations contribute to precipitation predictability and which ones are most important.

Machine learning (ML) methods, including neural networks (NN), provide a tool to explore these large-scale climate factors as potential sources of predictability. Until recently, NNs have been "black boxes" in which it was difficult to understand what relationships the network learned, rendering them difficult to use for scientific understanding. Recent advances in explainable machine learning (XML) techniques make it possible to understand what the NN learned about the relationships between the predictors and predictand. As a result of this capability, we can utilize these methods for physical understanding of sources of predictability (e.g., Toms et al. 2021a, 2020, 2021b; Barnes et al. 2020; McGovern et al. 2019; Mayer and Barnes 2021). In this paper, we use machine learning and XML to identify and understand sources of predictability for daily SEUS rainfall during winter and summer separately. Specifically, we ask how well we can detect the sign of daily SEUS precipitation anomalies if we know the simultaneous state of the large-scale climate, and what are the most important large-scale climate sources of predictability.

2. Method

We attempt to predict the sign of the area-aggregated daily precipitation anomaly over land in the SEUS (Fig. 1) based

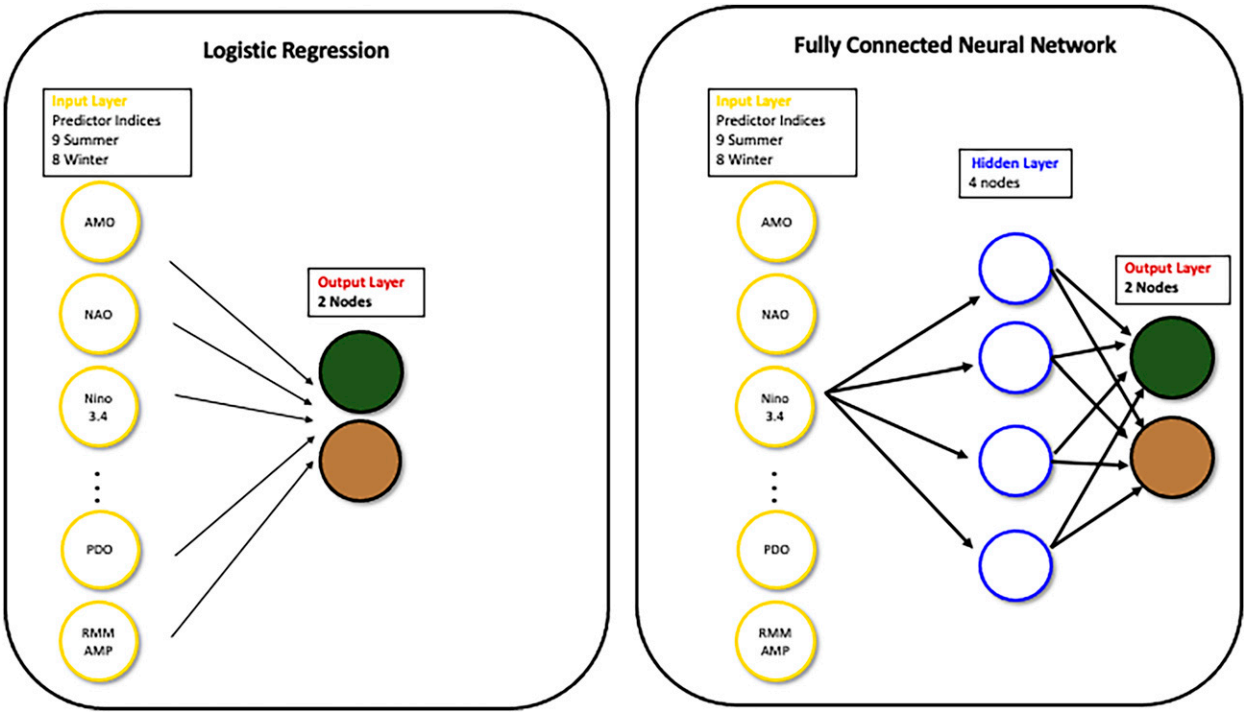

FIG. 2. Schematic of (left) LR and (right) FC-NN architectures.

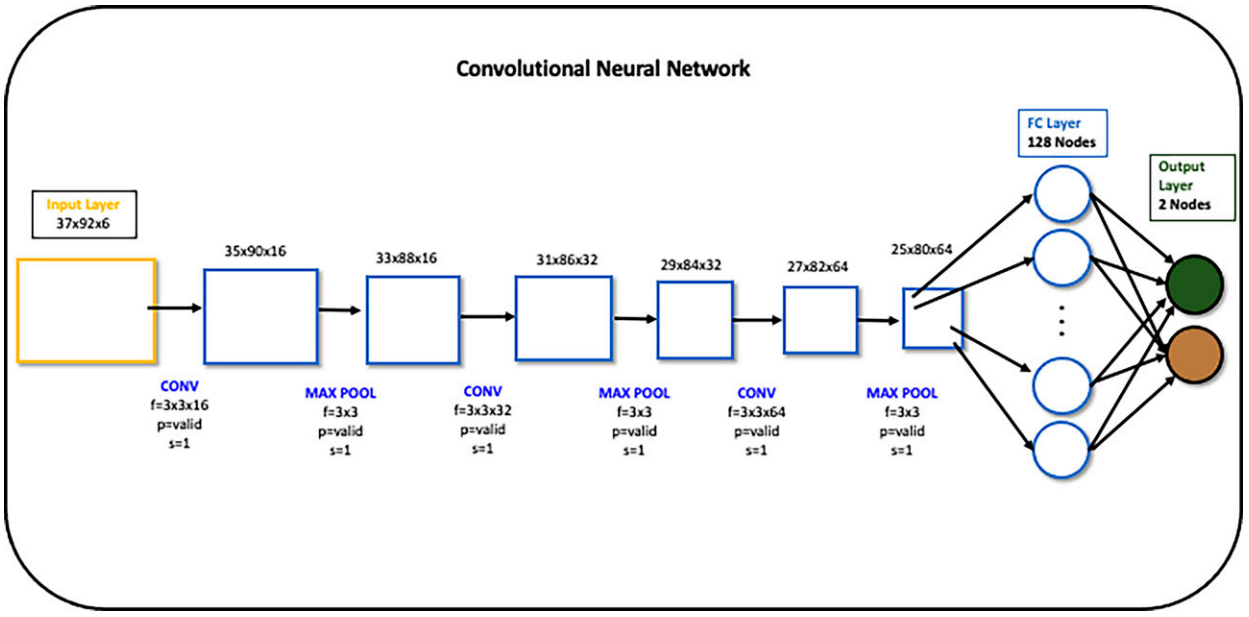

FIG. 3. Schematic of CNN architecture. The input layer (yellow) consists of coarse-grained fields of SST, OLR, U200, U850, Z200, and Z850 with dimensions of 37 longitude points by 92 latitude points by the 6 fields. Each convolutional layer is followed by a max-pooling layer (blue squares). Numbers above each convolutional or max-pooling layer indicate the dimensions of the data after the convolution or max-pooling is applied. Numbers below each layer indicate the dimensions of the convolutional or max-pooling filter *f*, the padding used *p*, and the stride length of the filter *s*. An FC-NN containing 128 nodes (blue circles) follows the convolutional and max-pooling layers. The output layer consists of two nodes indicating the probability of yes/no for the two categories of positive (green circle) and negative (brown circle) precipitation anomalies.

on the simultaneous state of the large-scale atmospheric circulation, SSTs, and tropical outgoing longwave radiation (OLR). The hypothesis is that knowing the large-scale state of the atmosphere and ocean can provide a significant amount of information about whether anomalous daily rainfall occurs over a large region such as the SEUS. As an initial baseline, we use logistic regression (LR) with indices representing well-known mechanisms of large-scale climate variability as predictors (section 2b). This tests the hypothesis that knowing the state of a particular large-scale climate mechanism can lead to useful prediction of SEUS daily precipitation. Second, we use a fully connected neural network (FC-NN) with the same predictors to determine if a more complex model can better identify the combinations between predictors and make better predictions (section 2c). It is likely that the indices do not capture all of the relevant large-scale variability for predicting precipitation. Therefore, we explore the use of gridded fields of anomalies as predictors with an LR and convolutional neural network (CNN) to see if allowing the models to learn from the data which variables and grid points are most relevant is more skillful and provides better insights into sources of predictability than a priori defining the predictors (also called feature engineering) (section 2d).

a. Data

ERA5-Land (Muñoz-Sabater et al. 2021) is used to define the target daily precipitation anomalies for the period of 1979–2019. Anomalies are determined by calculating and removing the climatology at each grid point. The climatology is calculated as the average value for each day of the calendar year over all years smoothed with a 31-day triangular window, following Pegion et al. (2019). The average precipitation is area aggregated over land points in the SEUS region (24°–36°N; 91°–77°W) and weighted by the cosine of latitude. We also tested ERA-Interim (Dee et al. 2011) precipitation and find consistent results.

Monthly climate indices are used over the same period. The indices used are the AMO, Pacific decadal oscillation (PDO), Niño-3.4 to represent ENSO, and the North Atlantic Oscillation (NAO). A summary of the indices is provided in Table 1. They are obtained from the Physical Sciences Laboratory (PSL) of the NOAA Earth System Research Laboratories (ESRL) and are interpolated to daily values. To represent the MJO, daily amplitude and phase of the real-time multivariate MJO index from the Center for Australian Weather and Climate Research (CAWCR) (Wheeler and Hendon 2004) is used. Daily large-scale circulation in the Pacific–North America region (20°–80°N; 150°–300°E) is identified using weather regimes following Amini and Straus (2019): five distinct regime patterns are identified separately in winter (DJF) and summer (JJA) by applying a *k*-means cluster analysis to Z500 and 250-hPa zonal wind (U250) from ERA-Interim (see appendix A). Each day is assigned an integer (0–4) to represent the identified regime.

The amplitude and phase of the western ridge of the NASH is identified following Li et al. (2015) using Z850 and

U850. The amplitude is the maximum Z850 value in the NASH region (0°–60°N; 120°W–0°). The location of the western edge is determined as the point of intersection between the 1560 gpm isoline and the ridgeline. The ridgeline is defined as the location in the NASH region where the easterly component of the 850-hPa wind reverses to be westerly. An integer value from 0 to 3 is assigned to each day indicating in which quadrant the western NASH edge is located relative to its summer climatological location of 86°W, 27°N (Li et al. 2015). The NASH is only used as a predictor during summer (JJA). The MLSO is also used as an index-based predictor (Stan and Krishnamurthy 2019). It is the 120-day oscillatory pattern determined using multichannel singular spectrum analysis on daily Z500 from 30° to 75°N and calculated using ERA-Interim. We acknowledge that biases in reanalysis datasets used in this study may impact the representation of climate indices and teleconnections (e.g., Belmonte Rivas and Stoffelen 2019).

ERA-Interim fields of anomalous global SST, zonal winds at 200 (U200) and 850 (U850) hPa, geopotential height at 500 (Z500) and 850 (Z850) hPa, and tropical (30°S–30°N) OLR are used as predictors for the field-based models (section 2d). The same fields are used for composites (sections 3d and 3e). The fields are coarse grained to a 5° × 5° grid to emphasize large-scale climate variability as predictors, minimize the likelihood of fitting the models to noise, and for computational efficiency. The grid points in the SEUS region are excluded as predictors.

b. *Logistic regression*

LR is used to represent the linear relationship between predictors and a binary or categorical outcome. It is the simplest model for our problem of detecting the sign of SEUS precipitation anomaly and can be generalized as a neural network with no hidden layers (Fig. 2). The coefficients or weights associated with each predictor are determined by minimizing the cross entropy loss function, which measures the difference between a predicted category and the true category [Eq. (B4)]. The coefficients of each predictor can be easily interpreted to identify their relative contribution to the model prediction. Logistic regression also serves as a baseline with which to compare more complex ML models. A detailed explanation of logistic regression and our implementation is available in appendix B.

c. *Fully connected neural network*

A NN can learn linear and nonlinear relationships between the predictors and target. We use a fully connected NN defined by interconnected layers. The input layer consists of the indices as predictors and the hidden layers of the NN consist of nodes that relate combinations of inputs to the output using weights (Fig. 2). The number of hidden layers, number of nodes, and weights associated with each node are model parameters that are determined via training and validation (see appendix C). The model is trained by minimizing the cross-entropy loss function [Eq. (B4)]. The FC-NN predicts whether SEUS precipitation is above or below normal (i.e., positive or negative anomaly). The output layer of the FC-NN consists of a probability of yes/no for each category. The identified category is then output based on which has the higher probability. Mayer and Barnes (2021) refer to the probability as a "forecast confidence" with a higher probability indicating a more confident prediction. They demonstrate for their NN and target of the sign of Z500 based on tropical OLR that more confident predictions are associated with more accurate predictions and thus identify forecasts of opportunity. Whether this is true for our problem will be explored in section 3. Further details of our FC-NN implementation are provided in appendix C.

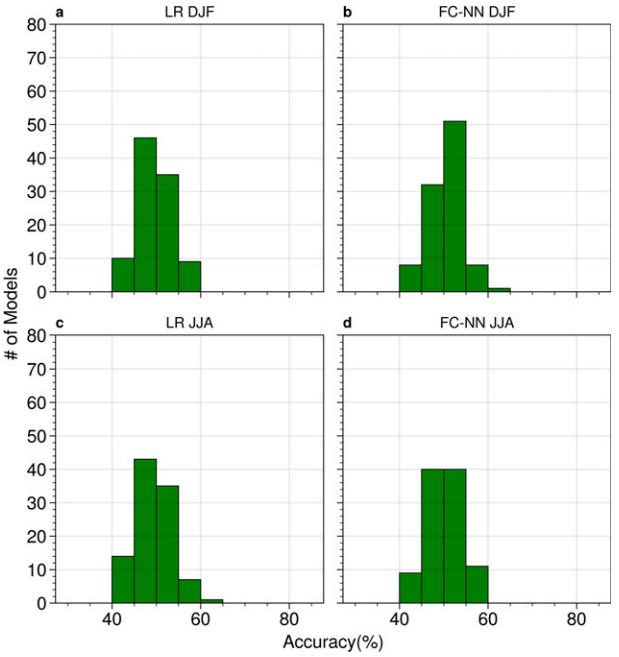

FIG. 4. Histogram of test accuracy of 100 (left) logistic regression models and (right) fully connected neural network models for (a),(b) winter and (c),(d) summer.

d. *Convolutional neural network*

We hypothesize that predicting the sign of SEUS precipitation anomalies using a priori defined indices as predictors is not likely to be skillful or reliable because these predictors do not capture the full range of large-scale variability that may impact SEUS precipitation. Therefore, we also explore the use of relevant gridded fields as predictors. The idea is that the data may be able to tell us more specifically what features are most important for SEUS precipitation and provide new insights into understanding and detecting the associated mechanisms. For example, Mayer and Barnes (2021) use a fully connected NN to predict the sign of North Atlantic 500-hPa geopotential height on subseasonal time scales using vectorized fields of tropical OLR. While using a single gridded input field works well in their case of predicting 500-hPa height, many mechanisms are hypothesized to contribute to SEUS precipitation predictability (see section 1), requiring a large number of input fields. With so many predictors relative to the sample size, the problem becomes intractable since weights

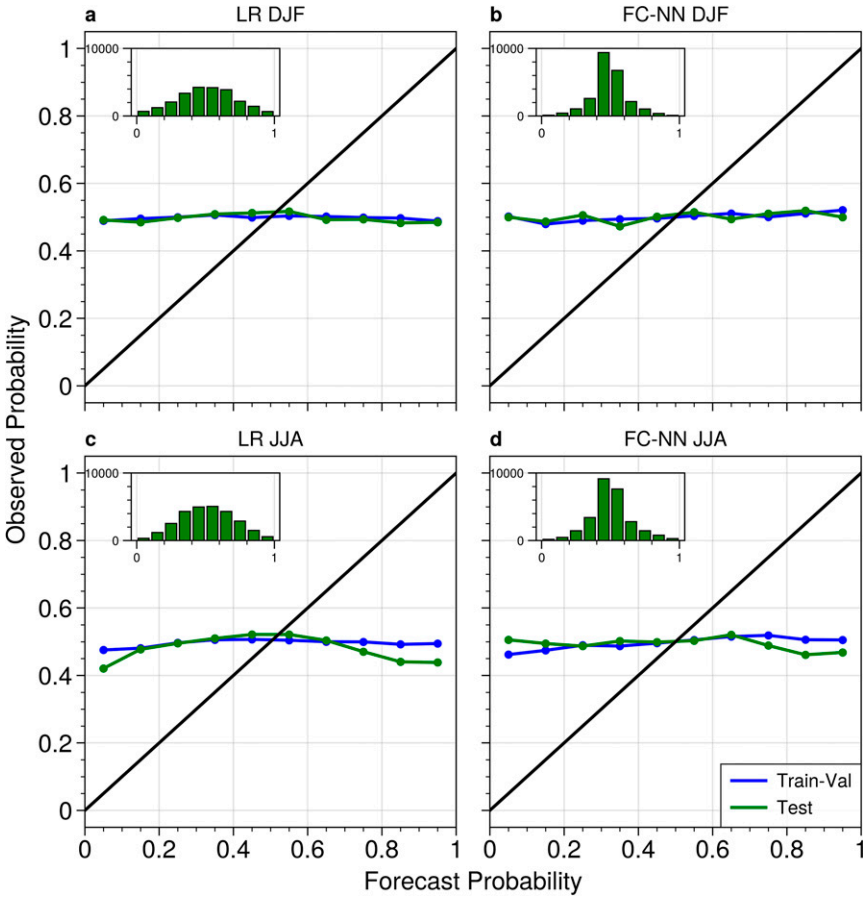

FIG. 5. Reliability diagrams for positive precipitation anomalies using the index models for (left) LR and (right) FC-NN during (a),(b) DJF and (c),(d) JJA. The *x* axis indicates the predicted frequency, and the *y* axis indicates the observed frequency. Test data are shown in green; training and validation data are shown in blue. Insets show the number of forecasts in each frequency bin for test data. Because there are two mutually exclusive categories, reliability is shown for the positive precipitation anomaly category. Negative precipitation reliability is 1 − positive precipitation reliability.

must be learned for every predictor connected to every node. A common approach in machine learning to address this is to use a CNN, which reduces the number of connections and weights to be learned in training the model by taking advantage of the spatial coherence of the data to learn low-level predictors (Krizhevsky et al. 2012). CNNs are often used in image detection as a feature selector, then a fully connected NN is used for the final layer (e.g., Krizhevsky et al. 2012; Zeiler and Fergus 2013). In our implementation, the gridded fields associated with each of the hypothesized predictor fields are used as input to the CNN, thus each grid point of each field constitutes a predictor. Our CNN consists of three convolutional layers, each of which are followed by max-pooling layers (Fig. 3). The third convolutional layer is followed by one fully connected layer and finally the output layer, which is the same as the FC-NN output layer (Fig. 3). Further details of our CNN implementation are provided in appendix D. The CNN is compared with an LR model that follows the same architecture as described in section 3b and appendix C using vectorized values of the same fields input to the CNN.

e. *Training, validation, and testing*

To train and test the models, the data are split into training, validation, and testing sets. The training data are used by the model to learn the weights associated with each predictor and node that can identify the known outcome. Validation data are used to tune the parameters of the models such as number of layers and nodes, regularization, and learning rate in order to identify a model configuration that is both skillful and is not overfit to the training data. The years 1979–2016 are used for training and validation, with 90% ($N_{\mathrm{DJF}} = 3087$; $N_{\mathrm{JJA}} = 3146$) used for training and 10% ($N_{\mathrm{DJF}} = 383$; $N_{\mathrm{JJA}} = 350$) for validation. Testing data are held out until a model configuration is decided on and fully trained to ensure that the model generalizes to data it has not yet seen. The years 2017–19 are used for testing ($N_{\mathrm{DJF}} = 240$; $N_{\mathrm{JJA}} = 276$).

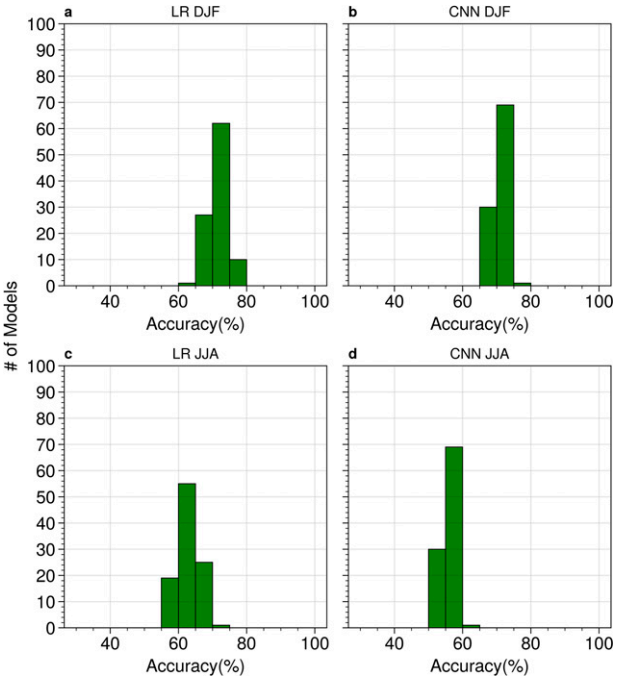

FIG. 6. Histogram of test accuracy of 100 (left) LR models and (right) CNN for (a),(b) winter and (c),(d) summer.

f. *Layerwise relevance propagation*

Following Barnes et al. (2020), Toms et al. (2021a, 2020, 2021b), and Mayer and Barnes (2021), we use layerwise relevance propagation (LRP) (Montavon et al. 2019; Bach et al. 2015) to understand what the NNs have learned about the relationships between the predictors and SEUS precipitation. Once a NN is trained, an input can be given to it to make a prediction. Using LRP, that prediction is then passed back through the trained network following specific rules to determine the relevance of each of the input predictors to the final prediction. The result is a "heat map" for each prediction that indicates a unitless relevance of each predictor to the output. Mayer and Barnes (2021) apply LRP to their NN to identify the OLR patterns most important for making successful and confident predictions of the sign of North Atlantic Z500. Toms et al. (2020) use LRP to produce heat maps of El Niño and La Niña patterns correctly identified by an FC-NN using SST anomalies as predictors. Toms et al. (2021b) use LRP to show how an FC-NN can accurately identify the known MJO spatial patterns when learning to detect the phase of the MJO based on clouds and circulation fields. Recent work has demonstrated that certain variations of LRP can potentially be misleading for geoscience applications (Mamalakis et al. 2022). Therefore, we test two variations of LRP. The LRP-$\alpha\beta$ rule used by Mayer and Barnes (2021) can determine a relevance for inputs that contribute both positively (α) and negatively (β) to the model prediction. We also test the LRP$_z$ rule, which was shown to be the most accurate rule using a benchmark climate dataset in Mamalakis et al. (2022). We find little difference between the two rules for our particular problem and model. We apply LRP to our FC-NN and CNN to identify sources of predictability for SEUS precipitation using the LRP-$\alpha\beta$ rule with $\alpha = 1$ and $\beta = 0$ to select only inputs that contribute positively to the predicted outcome of the model. Details of LRP and its implementation are provided in appendix E.

g. *Software and code implementation*

The models used in this paper are written using Keras and Tensorflow. LRP is applied using the innvestigate package (Alber et al. 2019). Data preprocessing uses xarray (Hoyer and Hamman 2017; Hoyer et al. 2021), numpy (Harris et al. 2020), scipy (Virtanen et al. 2020), and scikit-learn (Pedregosa et al. 2011). Plotting is done using matplotlilb (Hunter 2007) and proplot (Davis 2021). Reliability is calculated using xskillscore (https://xskillscore.readthedocs.io/en/stable/). The codes used in this paper are publicly available in GitHub (https://github.com/kpegion/ml-precip).

3. Results

a. *Skill and reliability of index-based models*

Despite the same architecture, the stochastic nature of the initialization and minimization procedure produces models with different weights, thus 100 models are trained and tested using the indices noted in Table 1. The histogram of the accuracy for the LR and FC-NN models is shown in Fig. 4. The accuracies demonstrate that the FC-NN and LR have similar skill. In both cases, the average skill is about 50%, indicating that the models are no better than random chance in identifying the sign of SEUS precipitation anomalies using these indices as predictors. Even if overall skill is poor, perhaps there are forecasts of opportunity that can be identified based on the probabilities provided by the models, that is, perhaps the forecasts with higher probabilities have a higher correct identification of the outcome (e.g., Mayer and Barnes 2021). This is quantified using reliability diagrams (Fig. 5) (Murphy and Winkler 1977). Reliability tells us how well the predicted probabilities of an event occurring corresponds to their observed frequencies. Neither the LR nor FC-NN can reliably identify the sign of SEUS precipitation. These results are underscored by the fact that there is little consistency from forecast to forecast or across the models in the weights of the LR predictors or the predictors identified as most important using LRP with the FC-NN (not shown). This tells us that these predictors are not sufficient to accurately or reliably identify the sign of daily SEUS precipitation anomalies. It also tells us that a complex model that can learn nonlinear relationships does not produce better predictions than the logistic regression model when using the a priori defined, index-based predictors.

b. *Skill and reliability of grid-based models*

Next, we test the skill and reliability of models using gridded ERA-Interim fields of anomalous global SST, U200, U850, Z500, Z850, and tropical OLR as predictors rather than defining them a priori as climate indices. One hundred CNNs are trained using randomly initialized weights. We also train and test 100 randomly initialized LR models using vectorized

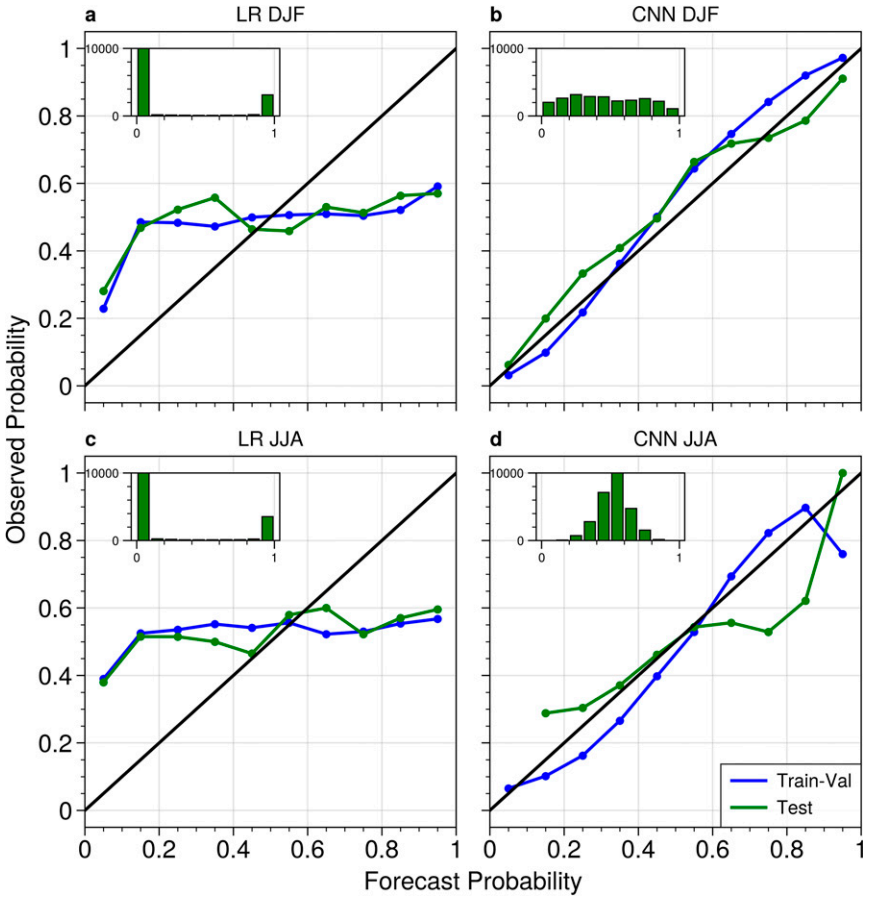

FIG. 7. Reliability diagrams for positive precipitation anomalies using the grid-based models for (left) LR and (right) CNN during (a),(b) DJF and (c),(d) JJA. The *x* axis indicates the predicted frequency, and the *y* axis indicates the observed frequency. Test data are shown in green; training and validation data are shown in blue. Insets show the number of forecasts in each frequency bin for test data. Because there are two mutually exclusive categories, reliability is shown for the positive precipitation anomaly category. Negative precipitation reliability is 1 − positive precipitation reliability.

gridded fields as input predictors as a baseline comparison for the CNN. For winter and summer, both the LR and CNN are more accurate (Fig. 6) than the index-based models, but the CNN is more reliable (Fig. 7). This means that the CNN probabilities can be used as a reliable measure of forecast confidence for any given forecast. We will use these probabilities to explore and understand sources of predictability.

c. *Sources of predictability for positive anomalies of SEUS precipitation*

To explore the sources of predictability for SEUS precipitation, we use LRP to identify the predictors (i.e., grid points and fields) most relevant to the prediction of confident (≥80% probability) forecasts of positive and negative precipitation anomalies. Composites are also used to relate unitless LRP relevance to physical quantities. The relevance and composites for confident, correct forecasts of above-normal SEUS precipitation are shown in Figs. 8 and 9. Relevance values are normalized by the relevance of all other predictors (i.e., grid points and fields) such that the grid point over all fields with the highest value in any given forecast is one. The red and blue contours are positive and negative composite anomalies of the specified field. Relevance is indicated by the gray shading and is only shown for positive values meaning that a given variable and grid point contributes positively to the CNN prediction.

The anomalous low 850-hPa geopotential heights over the Gulf of Mexico and associated 850-hPa zonal wind anomalies have the strongest relevance during winter (Figs. 9a,c). The circulation associated with these height anomalies brings moisture from the Gulf of Mexico into the SEUS and is coincident with El Niño–related SST and OLR anomalies in the tropical Pacific (Figs. 8a,b). The SST and OLR anomalies show no relevance in comparison with the circulation fields (i.e., there are very few points with gray shading). The winter relevance and composites show results consistent with well-known winter teleconnection patterns associated with ENSO (e.g., Ropelewski and Halpert 1986). This provides further evidence that the CNN can learn and LRP can identify physical

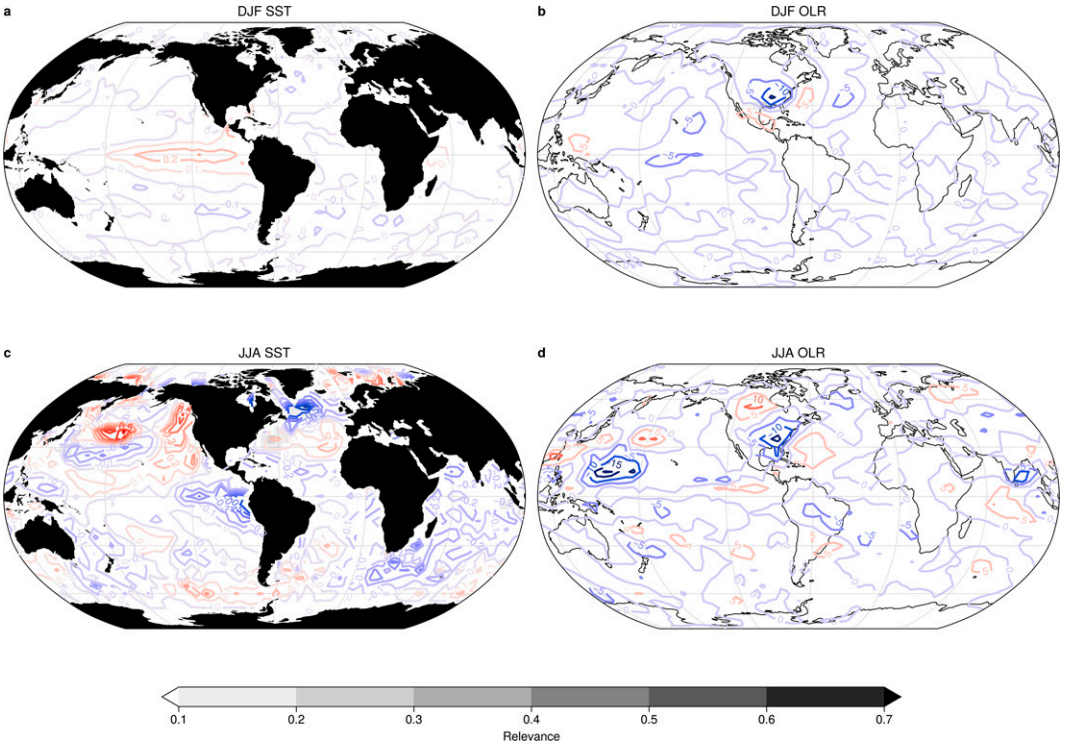

FIG. 8. Composite anomalies of (left) SST (°C) and (right) OLR (W m^{-2}) (contours) and normalized relevance (grayscale shading) for confident (≥80% probability) and correct forecasts of positive precipitation anomalies over training, validation, and test data for (a),(b) DJF and (c),(d) JJA. Contour intervals are 0.1°C for sea surface temperature and 5 W m^{-2} for outgoing longwave radiation. Relevance values may be too small to appear on the map.

relationships. Previous studies have identified a correlation between DJF SST anomalies in the Gulf of Mexico and convective precipitation in which warm Gulf of Mexico SST anomalies contribute to severe convection in the SEUS (e.g., Edwards and Weiss 1996; Molina et al. 2016, 2018; Molina and Allen 2019; Molina et al. 2020). However, we do not find any SST relevance in the Gulf of Mexico during the winter season in our analysis.

During the summer, the strongest relevance is associated with positive 850-hPa height anomalies and corresponding 850-hPa zonal winds. This circulation is consistent with a southwest shift of the NASH and wet conditions over the SEUS (Fig. 9c). There is also relevance about 10%–20% of the maximum relevance associated with SST anomalies in the NASH region and in the North Atlantic, consistent with the AMO–NASH hypothesis for SEUS precipitation (Fig. 8c). Molina et al. (2016) find a relationship between Gulf of Mexico SST and severe thunderstorm occurrence in the SEUS during March–May, but no SST relevance for the Gulf of Mexico is found during summer in our analysis.

d. *Sources of predictability for negative anomalies of SEUS precipitation*

Confident, correct forecasts for negative precipitation anomalies in the SEUS are shown in Figs. 10 and 11. The locations and corresponding composites are similar to the positive precipitation forecasts but with opposite sign. The 850-hPa zonal winds and geopotential heights enhance flow of continental air over the SEUS during winter (Figs. 11a,b). These conditions occur during weak La Niña SST anomalies and near zero tropical OLR anomalies (Figs. 10a,b). The relevant 850-hPa heights and winds during summer are consistent with a northwestward shift of the NASH. (Figs. 11c,d). Corresponding summer SST anomalies are cold in the central tropical Pacific and strong in the subtropical and midlatitude Pacific but have weak relevance in comparison with the circulation fields (Figs. 10c,d).

e. *Time scales of predictability*

To further understand these composite patterns as sources of predictability, we investigate their time scales. A time series associated with each pattern is calculated by projecting the original anomalies for each variable onto its composite pattern for confident, successful forecasts of positive and negative precipitation anomalies (i.e., the patterns shown in Figs. 8–11). The projections are calculated separately for DJF and JJA. To quantify the dominant time scales for each of these time series, power spectra are estimated using the Welch (1967) method in scipy.signal (Virtanen et al. 2020), which calculates the power spectra for overlapping segments of data and then averages the spectra for each segment to produce a smooth periodogram. Corresponding red noise spectra are also calculated based on the estimated decorrelation time T_e of the time series of each composite for each variable (Table 2). There is no significant power above a red noise spectra for any of the variables in winter or summer (not shown). This indicates that

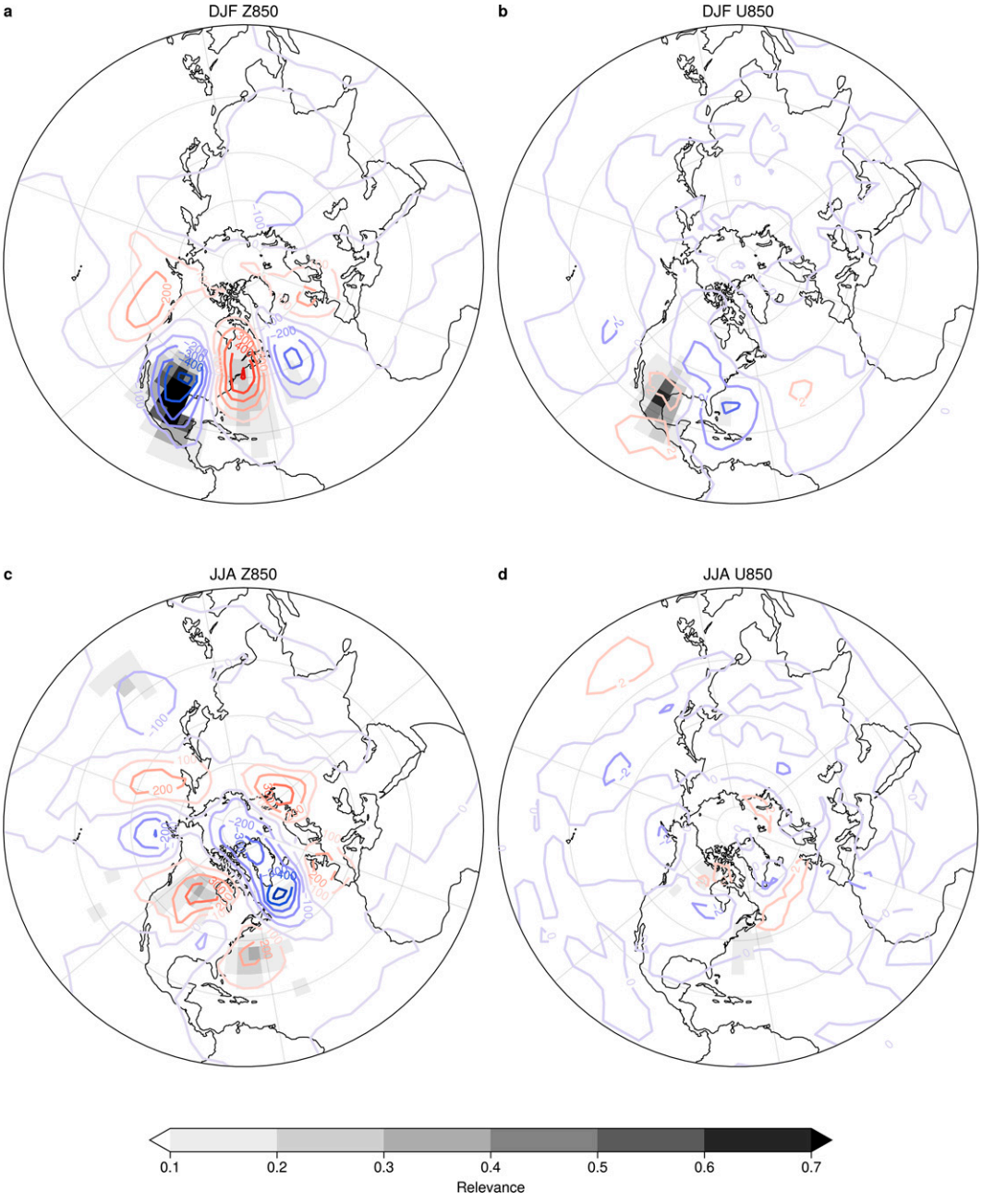

FIG. 9. Composite anomalies of (left) Z850 (m) and (right) U850 (m s^{-1}) (contours) and normalized relevance (grayscale shading) for confident (≥80% probability) and correct forecasts of positive precipitation anomalies over training, validation, and test data for (a),(b) DJF and (c),(d) JJA. Contour intervals are 100 m for Z850 and 2 m s^{-1} for U850.

the time scales of predictability for these patterns are primarily associated with their persistence.

4. Conclusions and discussion

We investigate sources of predictability for daily precipitation in the SEUS with machine learning models. Many studies have provided evidence of large-scale climate phenomena that impact SEUS precipitation; this study explores how well we can predict the sign of daily SEUS precipitation anomalies if these large-scale climate predictors are known simultaneously with the predictand, and which of these predictors are relatively most important. A logistic regression and fully connected neural network trained using indices representing the large-scale climate phenomena as predictors are neither skillful nor reliable. This tells us that these climate indices are not

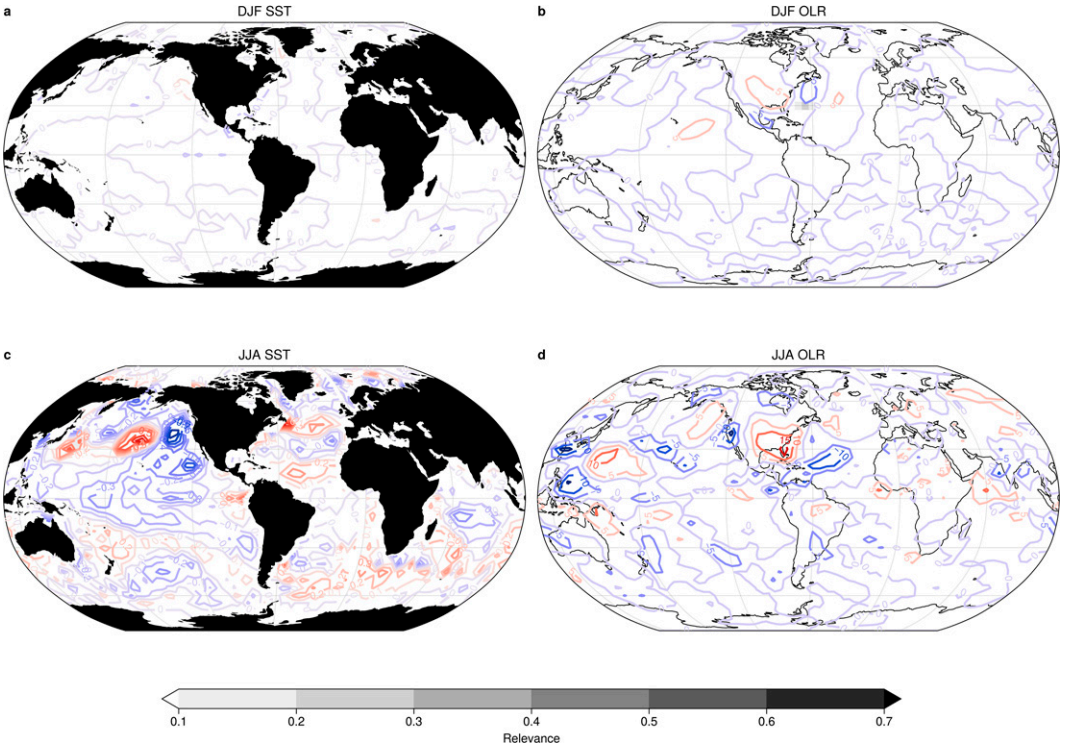

FIG. 10. Composite anomalies of (left) SST (°C) and (right) OLR (W m^{-2}) (contours) and normalized relevance (grayscale shading) for confident (≥80% probability) and correct forecasts of negative precipitation anomalies over training, validation, and test data for (a),(b) DJF and (c),(d) JJA. Contour intervals are 0.1 (°C) for sea surface temperature and 5 (W m^{-2}) for OLR. Relevance values may be too small to appear on the map.

sufficient to predict the sign of daily precipitation anomalies when their simultaneous value with the predictand is known. While an FC-NN is capable of representing nonlinear and linear relationships, the fact that it has similar skill (50% accuracy) and lack of reliability as the logistic regression model indicates that the inability to predict the sign of SEUS precipitation anomalies is because the indices are not good predictors. It is very likely that the key large-scale drivers of SEUS rainfall do not simply map onto our standard climate modes; indeed, indices of climate modes do not capture the full diversity and complexity of large-scale climate variability. Based on this, we can conclude that using indices of large-scale climate phenomena is insufficient to skillfully or reliably predict the sign of daily precipitation anomalies in the SEUS and that the attribution or explanation of precipitation variability or extremes to these climate mechanisms does not necessarily equate to predictability of precipitation in this region.

The finding that neither the logistic regression nor the fully connected neural network with index-based predictors can predict daily precipitation is not an unexpected outcome. While subseasonal and seasonal climate patterns are known to influence the statistics of regional precipitation averaged over weeks or months (e.g., Ropelewski and Halpert 1986, 1987; Higgins et al. 2000), due to the high internal variability of precipitation, we would be surprised if we could predict even the sign of a single-day precipitation anomaly given only several climate indexes. This specific analysis was undertaken to provide a baseline for the prediction of daily precipitation using full atmospheric fields in a CNN, which is found to be much more successful. It also reminds us that even when precipitation events are explained or attributed to large-scale climate mechanisms, that attribution does not necessarily translate to prediction. Future work will test this method in the context of higher-predictability fields, such as the frequency of precipitation during a subseasonal period, the number of dry or wet spells, or subseasonal extreme precipitation. We would also be curious to see how the combination of several climate indexes in an ML context performs with the prediction of daily temperature or heat waves.

Global gridded fields as predictors produce more accurate predictions in both an LR and CNN (70% for winter; 60% for summer) and the predictions from the CNN in both winter and summer are reliable. This allows us to use the probability from the CNN for each forecast category as a forecast confidence to identify forecasts of opportunity, following Mayer and Barnes (2021). We use the forecast confidence and layerwise relevance propagation to identify the most relevant predictors for confident (≥80%) and correct forecasts. For both positive and negative precipitation anomalies, the most relevant predictors are identified as the local zonal wind and geopotential height anomalies at 850 hPa. The patterns identified as relevant by LRP are patterns of large-scale synoptic variability, consistent with Diem (2006) and with the

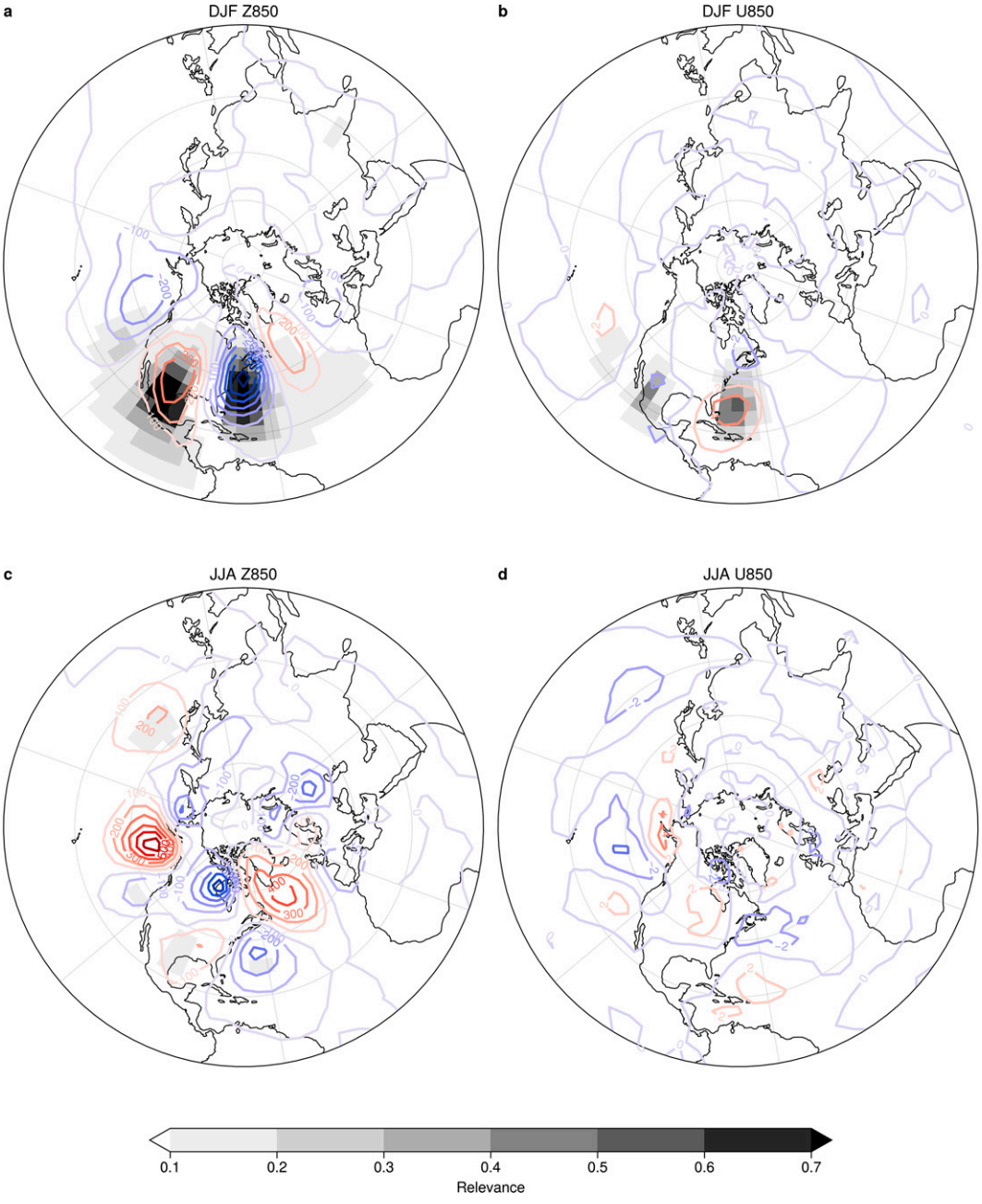

FIG. 11. Composite anomalies of (left) Z850 (m) and (right) U850 (m s^{-1}) (contours) and normalized relevance (gray-scale shading) for confident (≥80% probability) and correct forecasts of negative precipitation anomalies over training, validation, and test data for (a),(b) DJF and (c),(d) JJA. Contour intervals are 100 m for Z850 and 2 m s^{-1} for U850.

persistence time scales of the gridded predictor fields. To correctly predict even the sign of daily precipitation anomalies in the SEUS, skillful predictions of the local synoptic circulation are required. Composites of corresponding SST and OLR anomalies point to a relationship between the circulation and well-known large-scale climate phenomena such as ENSO during winter and the AMO–NASH during summer. However, just using indices of these phenomena did not lead to skillful prediction. This is because there is large uncertainty in the SEUS circulation anomalies associated with large-scale climate variability (e.g., Deser et al. 2018).

Our experiments are idealized–the simultaneous values of these predictors with the predictand would not be known at forecast time in a true forecasting situation. A more realistic prediction would be obtained by using either predicted variables as input to our ML models or observed data at some forecast lead time. We argue that the use of simultaneous predictors and predictand used in this paper show an upper limit to

TABLE 2. Table of decorrelation times. A plus sign indicates time series from positive precipitation anomalies. A minus sign indicates time series from negative precipitation anomalies.

Variable	T_e for DJF (days)	T_e for JJA (days)
SST	+437/−187	+228/−292
OLR	+6/−5	+4/−4
Z500	+7/−5	+7/−6
Z850	+7/−5	+7/−6
U200	+4/−5	+6/−5
U850	+4/−3	+4/−4

predicting the sign of precipitation anomalies in the SEUS based on available training data. Errors in the predictors will produce even less accurate predictions. Future work will explore how errors in the predictor fields lead to errors in precipitation at various lead times. It is also important to note that daily precipitation is an extremely difficult prediction to make and many aspects of it may not be predictable using these predictors (e.g., mesoscale, local soil moisture, tropical cyclones, low-level jets, or diurnal variability).

Acknowledgments. Author Pegion thanks K. Huang for assistance downloading data. Author Kirtman acknowledges the support from NOAA (NA20OAR4320472, NA18OAR4310293), the National Science Foundation (OCE1419569, OCE1559151), and the U.S. Department of Energy (DE-SC0019433).

Data availability statement. Monthly climate indices for the AMO, NAO, and Niño-3.4 are publicly available from NOAA/ESRL/PSL (https://psl.noaa.gov/data/climateindices/list/). The real-time multivariate MJO indices are publicly available from http://www.bom.gov.au/climate/mjo/. ERA-Interim data are available from https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim. ERA5 data are available from the Copernicus Climate Data Store (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5). Codes for this project are available on GitHub (https://github.com/kpegion/ml-precip). Pacific–North America weather regimes, MLSO, and NASH can be calculated from ERA-Interim using the codes provided. The codes used to produce the LR, FC-NN, and CNN models are also provided on GitHub. LRP is applied using the investigate toolkit and is available on GitHub (https://github.com/albermax/innvestigate).

APPENDIX A

Weather Regimes

Weather regimes (also called circulation regimes) are preferred and persistent patterns of the atmospheric circulation (Reinhold and Pierrehumbert 1982; Straus et al. 2007). Their persistence of approximately 7–15 days makes them potential predictors on S2S time scales (Robertson et al. 2020; Vigaud et al. 2018). In this study, they are used as predictors for the index-based models. They are identified using a k-means cluster analysis following Amini and Straus (2019). They applied k-means cluster analysis to the leading 12 principal components (80% variance explained) of combined EOFs of daily winter (DJF) 500-hPa geopotential height and 250-hPa zonal winds from ERA-Interim in the Pacific–North America region. They tested a range of values for the number of clusters (k) and determined that $k = 4$ or $k = 5$ produce a robust and distinct set of clusters. We choose $k = 5$ to define our weather regime predictors and also use ERA-Interim. Composite analysis of the clusters shows that they correspond to well-known large-scale circulation patterns (Fig. A1), including the Arctic high, Arctic low, Alaskan ridge, Pacific wave train, and Pacific trough that occur between about 14% and 25% of the time during the period. We also extend this definition of weather regimes into the summer (JJA), which bear some resemblance to the winter regimes, but are weaker in amplitude (Fig. A2). Similarly, they occur 17%–23% of the time.

APPENDIX B

Logistic Regression Implementation

Logistic regression is used as a baseline, simplest possible model for our two-class problem of predicting yes or no for positive precipitation anomalies and yes or no for negative precipitation anomalies. It can be viewed as a neural network with no hidden nodes. It contains an input layer of predictors (x) and an output layer that predicts the probability of a given output class (e.g., positive or negative anomalies) based on an input x_i as follows:

$$\hat{y} = \sigma(z_i) = P(\text{Pos, Neg}|x_i) \tag{B1}$$

where

$$z_i = w^{\mathrm{T}} x_i + b \quad \text{and} \tag{B2}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{B3}$$

The predictors x for our problem are the indices in Table 1. The weights w and biases b associated with each predictor are determined by minimizing the cross-entropy loss function over all training samples m:

$$J(w, b) = -\sum_{i=1}^{m} y_i \log(\hat{y}_i). \tag{B4}$$

Minimization of the loss function is determined by numerically estimating its derivative, $dJ(w, b)/dw$ through backpropagation. For a given training sample, once the output is determined, the partial derivatives of the loss function due to each node are determined by going backward through the network. These partial derivatives are combined to determine the total errors for a given prediction in a chain-rule like fashion.

The softmax function is applied to the output layer to convert the probabilities to a categorical outcome (i.e., yes/no) for each of the two categories:

$$\hat{y}_i = \frac{e^{z_i}}{\sum e^{z_i}}. \tag{B5}$$

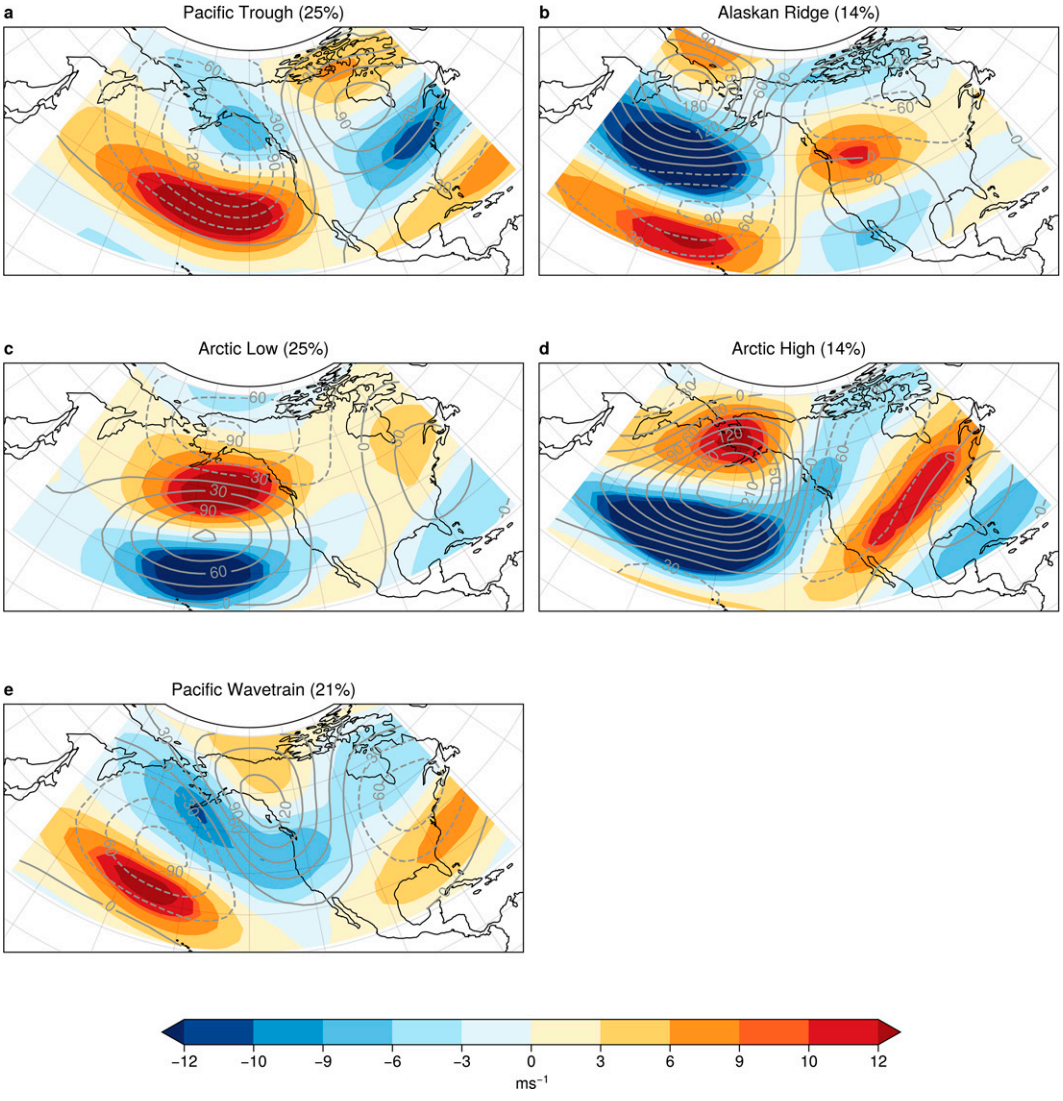

FIG. A1. Winter weather regime composites of 250-hPa zonal winds (m s^{-1}; color-filled contours) and 500-hPa geopotential heights (m; gray contours). Percent occurrence of each regime is indicated in parentheses.

The target values of y_i are set to ones and zeros for each category using one-hot encoding. The median of the precipitation anomalies is removed prior to the encoding to ensure balanced target classes.

To train our logistic regression model, we use minibatch gradient descent with a batch size of 25 and 250 epochs. A learning rate of 10^{-5} is used with the Adam optimization method (Kingma and Ba 2017).

APPENDIX C

Fully Connected Neural Network Implementation

The weights w and biases b of the fully connected neural network are trained using the same procedure as in the LR model. However, the FC-NN has hidden layers and nodes that allow it to learn more complex, relationships between the predictors and target. The value of a given node in a specific layer $a(z_j)$ of the FC-NN is determined by minimizing the cross-entropy loss function [Eq. (B4)] and then applying the nonlinear rectified linear unit (ReLU) activation function:

$$a(z_j) = \max(0, z_j). \tag{C1}$$

This is repeated for each layer and node in the FC-NN until the output layer where the softmax function is used to identify the predicted category [Eq. (B5)].

For the FC-NN minimization, the weights of the nodes must be initialized to small random nonzero values to start the gradient descent. The He normal initialization (He et al. 2016) is used to initialize the weights. The same learning rate and Adam optimization were used to train our FC-NN as was used for the LR. A range of hidden layers, nodes, and

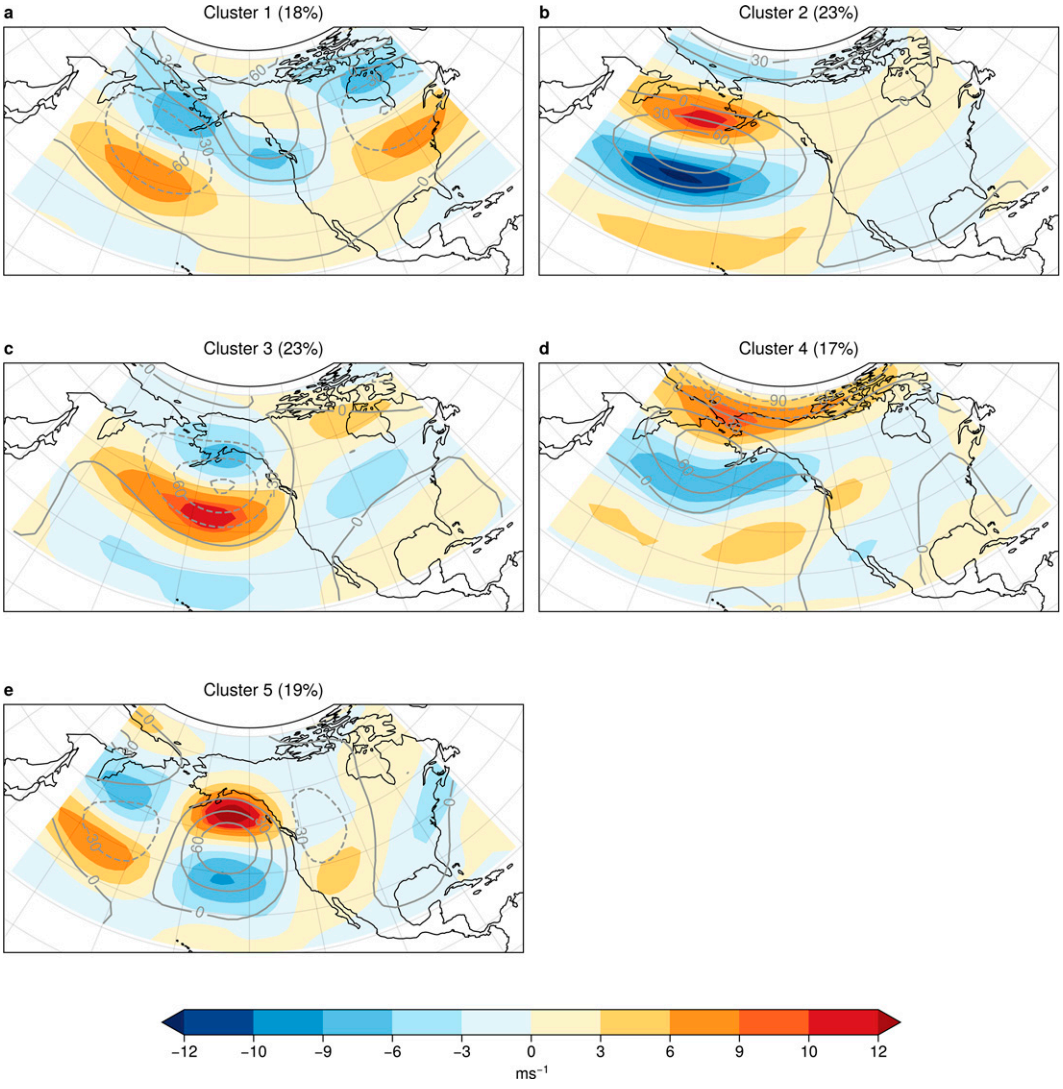

FIG. A2. Summer weather regime composites of 250-hPa zonal winds (m s^{-1}; color-filled contours) and 500-hPa geopotential heights (m; gray contours). Percent occurrence of each regime is indicated in parentheses.

regularization were tested. The selected model produced the most similar accuracy between the train and validation data and the most reliable predictions in training and validation. This model is relatively simple with only one hidden layer containing four nodes and no regularization. More complex models led to overfitting even with regularization.

APPENDIX D

Convolutional Neural Network Implementation

A CNN involves applying convolutional filters to the input data for each node to take advantage of spatial coherence of the data to learn the predictors and reduce the number of weights and biases that must be learned relative to an FC-NN (Krizhevsky et al. 2017). The convolution C of a matrix $\mathbf{A}$ of size $n_h \times n_w \times n_c$ with a filter $\mathbf{F}$ of size $f \times f \times n_c$ is as follows:

$$C_{ij} = (\mathbf{A} * \mathbf{F})_{ij} = \sum_{p=0}^{f-1} \sum_{q=0}^{f-1} \sum_{r=0}^{n_c-1} A_{i+p,j+q,r} * F_{pqr}. \quad \text{(D1)}$$

where the asterisk indicates convolution. The convolution can be applied to the matrix $\mathbf{A}$ with multiple filters. Our CNN architecture uses 16, 32, and 64 filters, respectively. Each filter is size 3×3 with valid padding and a stride of 1.

The value of a node i in a given convolutional layer l of a CNN is calculated as

$$Z_i^l = \mathbf{A} * F_i + b_i. \quad \text{(D2)}$$

Then the nonlinear ReLu activation function is applied to Z_i^l [Eq. (B5)].

Following each convolutional layer, a max-pooling layer is applied with valid padding to select the predictor (e.g., grid point) that has the maximum activation over each 3×3 grid.

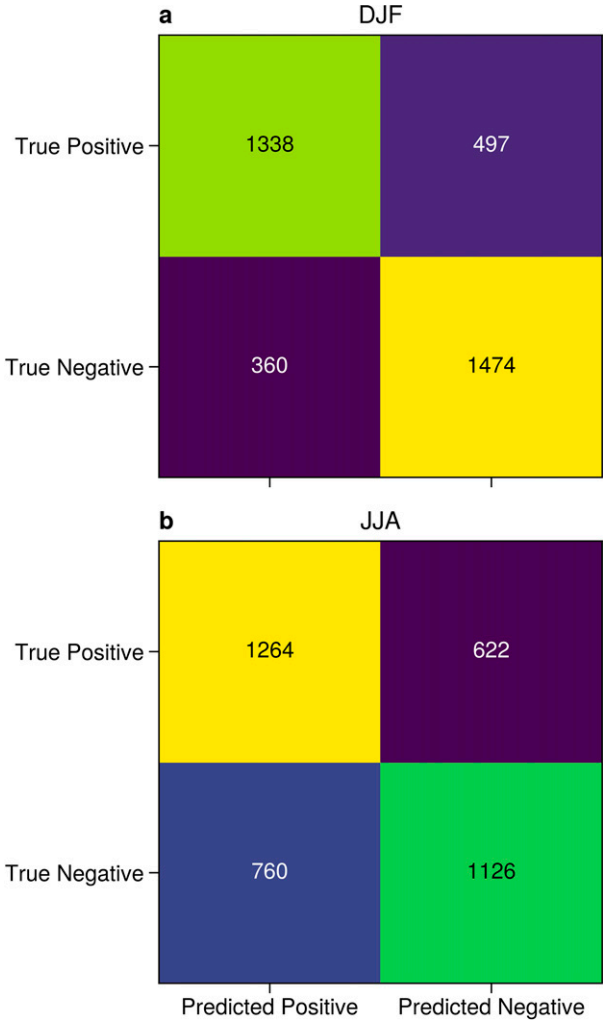

FIG. D1. Contingency tables for a selected model from the (a) DJF and (b) JJA CNNs.

Ridge regularization (L2) is applied in the first and second convolutional layers with a value of $\lambda = 20$ and $\lambda = 10$, respectively. Last, after the third convolutional and max-pooling layer, the outputs are flattened and treated as input for an FC-NN with 128 nodes. No regularization is applied to the third convolutional layer or FC-NN layer. The softmax function [Eq. (B5)] is then applied to identify the yes/no outcome for the two classes of negative and positive precipitation anomalies.

The input features to our CNN are the coarse-grained, global gridded data fields of SST, Z500, Z850, U850, U200, and tropical OLR (zeroed out poleward of 30°). To maintain the periodicity of the predictors in longitude during convolution, the input predictor fields are padded with a periodic halo of $p = 10$ points in the longitude direction. For our input data, $n_h = 37$, $n_w = 72 + 2p$, and $n_c = 6$. To train the CNN, we use a batch size of 25 and 100 epochs with early stopping after the validation loss increases for more than 2 epochs in a row. Training generally takes between 55 and 65 epochs. A learning rate of 3×10^{-5} and Adam optimization method are used. The regularization parameters, number of layers and nodes were determined through training and validation. The selected model produced the most similar accuracy between the train and validation data and the most reliable predictions in training and validation. The contingency table (also called confusion matrix) for an example CNN model is shown in Fig. D1.

APPENDIX E

Layerwise Relevance Propagation Implementation

LRP is applied to the FC-NN and CNN to identify which input predictors are most important to the predicted output of the model for any given prediction. We utilize the innvestigate toolkit (Alber et al. 2019) to apply LRP. For any given input to the model with output y, LRP goes back through the model and determines for each node what input was most relevant to the output until the input nodes are reached. We utilize the LRP-$\alpha\beta$ rule, which can determine a relevance for inputs that contribute both positively (α) and negatively (β) to the model output y. We use $\alpha = 1$ and $\beta = 0$ to select only inputs that contribute positively to the predicted outcome of the model. We also tested the LRP_z rule, which was shown to be the most accurate rule using a benchmark climate dataset in Mamalakis et al. (2022) and find little difference between them for our particular problem and model.

The relevance R of a given input j to an output k for the LRP-$\alpha\beta$ rule is given by the weighted value of the presoftmax activation of a positively contributing node divided by the weighted sum of all previous presoftmax nodes then multiplied by the total relevance of the previous node [Eq. (E1)]. The denominator ensures that total relevance is a conserved quantity:

$$R_j = \sum_j \frac{(a_j w_{jk})^+}{\sum_{0,j}(a_j w_{jk})^+} R_k. \tag{E1}$$

Once we determine the relevance of each predictor for all forecasts in the training and testing set, we then look for consistency between the predictors identified as most relevant to identify sources of predictability.

REFERENCES

Alber, M., and Coauthors, 2019: iNNvestigate neural networks! *J. Mach. Learn. Res.*, **20**, 1–8.

Amini, S., and D. M. Straus, 2019: Control of storminess over the Pacific and North America by circulation regimes. *Climate Dyn.*, **52**, 4749–4770, https://doi.org/10.1007/s00382-018-4409-7.

Arcodia, M. C., B. P. Kirtman, and L. S. P. Siqueira, 2020: How MJO teleconnections and ENSO interference impacts U.S. precipitation. *J. Climate*, **33**, 4621–4640, https://doi.org/10.1175/JCLI-D-19-0448.1.

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, 2015: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.

PLOS ONE, **10**, e0130140, https://doi.org/10.1371/journal.pone.0130140.

Balmaseda, M., and Coauthors, 2020: NOAA-DOE Precipitation Processes and Predictability Workshop. NOAA Tech. Rep. OAR CPO-9, 48 pp., https://cpo.noaa.gov/Portals/0/Docs/ESSM/Events/2020/NOAA_DOE_PrecipWorkshopReport_July2021.pdf?ver=2021-07-14-160100-057.

Barnes, E. A., B. Toms, J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2020: Indicator patterns of forced change learned by an artificial neural network. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002195, https://doi.org/10.1029/2020MS002195.

Becker, E., B. P. Kirtman, and K. Pegion, 2020: Evolution of the North American multi-model ensemble. *Geophys. Res. Lett.*, **47**, e2020GL087408, https://doi.org/10.1029/2020GL087408.

Becker, E. J., E. H. Berbery, and R. W. Higgins, 2011: Modulation of cold-season U.S. daily precipitation by the Madden–Julian oscillation. *J. Climate*, **24**, 5157–5166, https://doi.org/10.1175/2011JCLI4018.1.

Belmonte Rivas, M., and A. Stoffelen, 2019: Characterizing ERA-interim and ERA5 surface wind biases using ASCAT. *Ocean Sci.*, **15**, 831–852, https://doi.org/10.5194/os-15-831-2019.

Cassou, C., 2008: Intraseasonal interaction between the Madden–Julian oscillation and the North Atlantic oscillation. *Nature*, **455**, 523–527, https://doi.org/10.1038/nature07286.

Davis, L. L. B., 2021: ProPlot. Zenodo, https://doi.org/10.5281/zenodo.5602155.

Dee, D. P., and Coauthors, 2011: The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Deser, C., I. R. Simpson, A. S. Phillips, and K. A. McKinnon, 2018: How well do we know ENSO's climate impacts over North America, and how do we evaluate models accordingly? *J. Climate*, **31**, 4991–5014, https://doi.org/10.1175/JCLI-D-17-0783.1.

Diem, J. E., 2006: Synoptic-scale controls of summer precipitation in the southeastern United States. *J. Climate*, **19**, 613–621, https://doi.org/10.1175/JCLI3645.1.

Edwards, R., and S. J. Weiss, 1996: Comparisons between Gulf of Mexico Sea surface temperature anomalies and southern U.S. severe thunderstorm frequency in the cool season. *Proc. 18th Conf. on Severe Local Storms*, San Francisco, CA, Amer. Meteor. Soc., 2345–2363.

Harris, C. R., and Coauthors, 2020: Array programming with NumPy. *Nature*, **585**, 357–362, https://doi.org/10.1038/s41586-020-2649-2.

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, IEEE, 770–778, https://doi.org/10.1109/CVPR.2016.90.

Higgins, R. W., A. Leetmaa, Y. Xue, and A. Barnston, 2000: Dominant factors influencing the seasonal predictability of U.S. precipitation and surface air temperature. *J. Climate*, **13**, 3994–4017, https://doi.org/10.1175/1520-0442(2000)013<3994:DFITSP>2.0.CO;2.

Hoyer, S., and J. Hamman, 2017: xarray: N-D labeled arrays and datasets in Python. *J. Open Res. Software*, **5**, 10, https://doi.org/10.5334/jors.148.

——, and Coauthors, 2021: Xarray. Zenodo, https://doi.org/10.5281/zenodo.5771208.

Hu, Q., S. Feng, and R. J. Oglesby, 2011: Variations in North American summer precipitation driven by the Atlantic multidecadal oscillation. *J. Climate*, **24**, 5555–5570, https://doi.org/10.1175/2011JCLI4060.1.

Hunter, J. D., 2007: Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95, https://doi.org/10.1109/MCSE.2007.55.

Kingma, D. P., and J. Ba, 2017: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012: ImageNet classification with deep convolutional neural networks. *Proc. Advances in Neural Information Processing Systems 25*, Lake Tahoe, NV, NIPS, 1106–1114, https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

——, ——, and ——, 2017: ImageNet classification with deep convolutional neural networks. *Commun. ACM*, **60**, 84–90, https://doi.org/10.1145/3065386.

Li, L., W. Li, and Y. Kushnir, 2012: Variation of the North Atlantic subtropical high western ridge and its implication to southeastern US summer precipitation. *Climate Dyn.*, **39**, 1401–1412, https://doi.org/10.1007/s00382-011-1214-y.

——, ——, and J. Jin, 2015: Contribution of the North Atlantic subtropical high to regional climate model (RCM) skill in simulating southeastern United States summer precipitation. *Climate Dyn.*, **45**, 477–491, https://doi.org/10.1007/s00382-014-2352-9.

Li, W., T. Zou, L. Li, Y. Deng, V. T. Sun, Q. Zhang, J. B. Layton, and S. Setoguchi, 2019: Impacts of the North Atlantic subtropical high on interannual variation of summertime heat stress over the conterminous United States. *Climate Dyn.*, **53**, 3345–3359, https://doi.org/10.1007/s00382-019-04708-1.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2022: Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environ. Data Sci.*, **1**, e8, https://doi:10.1017/eds.2022.7.

Manthos, Z. H., K. V. Pegion, P. A. Dirmeyer, and C. Stan, 2022: The relationship between surface weather over North America and the mid-latitude seasonal oscillation. *Dyn. Atmos. Oceans*, **99**, 101314, https://doi.org/10.1016/j.dynatmoce.2022.101314.

Mayer, K. J., and E. A. Barnes, 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophys. Res. Lett.*, **48**, e2020GL092092, https://doi:10.1017/eds.2022.7.

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

Molina, M. J., and J. T. Allen, 2019: On the moisture origins of tornadic thunderstorms. *J. Climate*, **32**, 4321–4346, https://doi.org/10.1175/JCLI-D-18-0784.1.

——, R. P. Timmer, and J. T. Allen, 2016: Importance of the Gulf of Mexico as a climate driver for U.S. severe thunderstorm activity. *Geophys. Res. Lett.*, **43**, 12 295–12 304, https://doi.org/10.1002/2016GL071603.

——, J. T. Allen, and V. A. Gensini, 2018: The Gulf of Mexico and ENSO influence on subseasonal and seasonal CONUS winter tornado variability. *J. Appl. Meteor. Climatol.*, **57**, 2439–2463, https://doi.org/10.1175/JAMC-D-18-0046.1.

——, ——, and A. F. Prein, 2020: Moisture attribution and sensitivity analysis of a winter tornado outbreak. *Wea. Forecasting*, **35**, 1263–1288, https://doi.org/10.1175/WAF-D-19-0240.1.

Montavon, G., A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, 2019: Layer-wise relevance propagation: An overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek et al., Eds., Lecture Notes in Computer Science, Vol. 11700, Springer International Publishing, 193–209, https://doi.org/10.1007/978-3-030-28954-6_10.

Muñoz-Sabater, J., and Coauthors, 2021: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data*, **13**, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021.

Murphy, A. H., and R. L. Winkler, 1977: Reliability of subjective probability forecasts of precipitation and temperature. *J. Roy. Stat. Soc.*, **26C**, 41–47, https://doi.org/10.2307/2346866.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830, https://doi.org/10.48550/arXiv.1201.0490.

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.

Reinhold, B. B., and R. T. Pierrehumbert, 1982: Dynamics of weather regimes: Quasi-stationary waves and blocking. *Mon. Wea. Rev.*, **110**, 1105–1145, https://doi.org/10.1175/1520-0493(1982)110<1105:DOWRQS>2.0.CO;2.

Robertson, A. W., N. Vigaud, J. Yuan, and M. K. Tippett, 2020: Toward identifying subseasonal forecasts of opportunity using North American weather regimes. *Mon. Wea. Rev.*, **148**, 1861–1875, https://doi.org/10.1175/MWR-D-19-0285.1.

Ropelewski, C. F., and M. S. Halpert, 1986: North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Wea. Rev.*, **114**, 2352–2362, https://doi.org/10.1175/1520-0493(1986)114<2352:NAPATP>2.0.CO;2.

——, and ——, 1987: Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606–1626, https://doi.org/10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2.

Schubert, S., and Coauthors, 2009: A U.S. CLIVAR project to assess and compare the responses of global climate models to drought-related SST forcing patterns: Overview and results. *J. Climate*, **22**, 5251–5272, https://doi.org/10.1175/2009JCLI3060.1.

Stan, C., and V. Krishnamurthy, 2019: Intra-seasonal and seasonal variability of the Northern Hemisphere extra-tropics. *Climate Dyn.*, **53**, 4821–4839, https://doi.org/10.1007/s00382-019-04827-9.

——, D. M. Straus, J. S. Frederiksen, H. Lin, E. D. Maloney, and C. Schumacher, 2017: Review of tropical-extratropical teleconnections on intraseasonal time scales. *Rev. Geophys.*, **55**, 902–937, https://doi.org/10.1002/2016RG000538.

Straus, D. M., S. Corti, and F. Molteni, 2007: Circulation regimes: Chaotic variability versus SST-forced predictability. *J. Climate*, **20**, 2251–2272, https://doi.org/10.1175/JCLI4070.1.

Toms, B. A., E. A. Barnes, and I. Ebert-Uphoff, 2020: Physically interpretable neural networks for the geosciences: Applications to earth system variability. *J. Adv. Model. Earth Syst.*, **12**, e2019MS002002, https://doi.org/10.1029/2019MS002002.

——, ——, and J. W. Hurrell, 2021a: Assessing decadal predictability in an earth-system model using explainable neural networks. *Geophys. Res. Lett.*, **48**, e2021GL093842, https://doi.org/10.1029/2021GL093842.

——, K. Kashinath, Prabhat, and D. Yang, 2021b: Testing the reliability of interpretable neural networks in geoscience using the Madden–Julian oscillation. *Geosci. Model Dev.*, **14**, 4495–4508, https://doi.org/10.5194/gmd-14-4495-2021.

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

Virtanen, P., and Coauthors, 2020: SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272, https://doi.org/10.1038/s41592-019-0686-2.

Welch, P., 1967: The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.*, **15**, 70–73, https://doi.org/10.1109/TAU.1967.1161901.

Wheeler, M. C., and H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132%3C1917:AARMMI%3E2.0.CO;2.

Zeiler, M. D., and R. Fergus, 2013: Visualizing and understanding convolutional networks. arXiv, 1311.2901, https://doi.org/10.48550/arXiv.1311.2901.

Zhang, W., B. Kirtman, L. Siqueira, B. Xiang, J. Infanti, and N. Perlin, 2022: Decadal variability of southeast US rainfall in an eddying global coupled model. *Geophys. Res. Lett.*, **49**, e2021GL096709, https://doi.org/10.1029/2021GL096709.

Zhou, S., M. L'Heureux, S. Weaver, and A. Kumar, 2011: A composite study of the MJO influence on the surface air temperature and precipitation over the continental United States. *Climate Dyn.*, **38**, 1459–1471, https://doi.org/10.1007/s00382-011-1001-9.