JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2021MS002774

Special Section:
Machine learning application to Earth system modeling

Using Simple, Explainable Neural Networks to Predict the Madden-Julian Oscillation

Zane K. Martin[1], Elizabeth A. Barnes[1], and Eric Maloney[1]

[1]Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA

Key Points:

- Simple machine learning models are an efficient, flexible tool to predict and study the Madden-Julian oscillation (MJO)
- Shallow neural networks skillfully predict an MJO index out to ∼18 days in winter and ∼11 days in summer, outperforming linear models
- Varying ANN input and using explainable artificial intelligence methods offer insights into the MJO and key regions for prediction skill

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Z. K. Martin,
zkmartin@colostate.edu

Citation:

Martin, Z. K., Barnes, E. A., & Maloney, E. (2022). Using simple, explainable neural networks to predict the Madden-Julian oscillation. *Journal of Advances in Modeling Earth Systems*, *14*, e2021MS002774. https://doi.org/10.1029/2021MS002774

Received 19 AUG 2021
Accepted 21 APR 2022

Abstract Few studies have utilized machine learning techniques to predict or understand the Madden-Julian oscillation (MJO), a key source of subseasonal variability and predictability. Here, we present a simple framework for real-time MJO prediction using shallow artificial neural networks (ANNs). We construct two ANN architectures, one deterministic and one probabilistic, that predict a real-time MJO index using maps of tropical variables. These ANNs make skillful MJO predictions out to ∼18 days in October-March and ∼11 days in April-September, outperforming conventional linear models and efficiently capturing aspects of MJO predictability found in more complex, dynamical models. The flexibility and explainability of simple ANN frameworks are highlighted through varying model input and applying ANN explainability techniques that reveal sources and regions important for ANN prediction skill. The accessibility, performance, and efficiency of this simple machine learning framework is more broadly applicable to predict and understand other Earth system phenomena.

Plain Language Summary The Madden-Julian oscillation (MJO)—a large-scale, organized pattern of wind and rain in the tropics—is important for making weather and climate predictions weeks to months into the future. Many different numerical models have been used to study the MJO, but few works have examined how machine learning and artificial intelligence methods can predict and understand the oscillation. In this work, we show how two different types of machine learning models, called artificial neural networks, perform at predicting the MJO. We demonstrate that simple artificial neural networks make skillful MJO predictions beyond 1–2 weeks into the future, and perform better than other statistical methods. We also highlight how neural networks can be used to explore sources of prediction skill, via changing what variables the model uses and applying techniques that identify regions important for skillful predictions. Because our neural networks perform relatively well, are simple to implement, are computationally affordable, and can be used to inform scientific understanding, we believe these methods are more broadly applicable to study other important climate phenomena aside from just the MJO.

1. Introduction

The Madden-Julian oscillation (MJO), a planetary-scale, eastward-propagating coupling of tropical circulation and convection (Madden & Julian, 1971, 1972; Zhang, 2005), is a key source of subseasonal-to-seasonal (S2S) predictability (H. Kim et al., 2018; Vitart et al., 2017). Skillful MJO prediction has important societal implications (H. Kim et al., 2018; Meehl et al., 2021; Vitart et al., 2017), and extensive research has explored using both statistical models and initialized dynamical forecast models to predict the MJO (e.g., H. Kim et al., 2018; Meehl et al., 2021; Vitart et al., 2017; Waliser, 2012; and references therein). Before the late 2000s, statistical models showed superior MJO prediction skill (∼2 weeks; Kang & Kim, 2010; Waliser, 2012) compared to dynamical models, but S2S forecast models have continually improved and several now skillfully predict the MJO beyond 1 month (H. Kim et al., 2018; Vitart, 2014, 2017).

In contrast, statistical MJO modeling has stagnated in recent years. Compared to dynamical models, statistical MJO models have the advantage of being computationally inexpensive and are often simpler to formulate and understand. To date, the most common statistical MJO models use linear methods (e.g., H. Kim et al., 2018; Jiang et al., 2008; Kang & Kim, 2010; Maharaj & Wheeler, 2005; Marshall et al., 2016; Seo et al., 2009), and applying new statistical tools to study or predict the MJO, including especially non-linear machine learning (ML) techniques, remains a nascent research topic. ML techniques have proven skillful at predicting a variety of other climate and weather phenomena (Gagne et al., 2014; Ham et al., 2019; Lagerquist et al., 2017; Mayer & Barnes, 2021; McGovern et al., 2017; Rasp et al., 2020; Weyn et al., 2019), and application of ML methods to

study the MJO may thus improve the ability to forecast the oscillation or related S2S processes (e.g., Mayer & Barnes, 2021).

Studies using ML to study the MJO have identified the MJO (Toms et al., 2019), reconstructed past MJO behavior (Dasgupta et al., 2020), or bias-corrected dynamical model output of MJO indices (H. Kim et al., 2021), but only a few studies have examined MJO prediction solely using ML (Hagos et al., 2021; Love & Matthews, 2009). It is thus timely to establish ML frameworks for predicting the MJO and quantify ML model performance compared to other statistical and dynamical models. A further goal of this study is to demonstrate how simple ML models may be used for more than just prediction. While prediction skill is an undeniably important metric for model performance, simple ML models are also flexible tools that invite experimentation and can inform physical understanding of climate processes like the MJO. We highlight this under-appreciated aspect of ML modeling here through experiments changing model input, through exploration of both deterministic and probabilistic ML model architectures, and through application of tools from the field of explainable AI (XAI; Mamalakis et al., 2021; McGovern et al., 2019; Toms et al., 2020).

This paper addresses three aspects of using ML to study the MJO: (a) developing ML frameworks, (b) analyzing ML model performance, and (c) demonstrating how ML can inform scientific understanding. We prioritize simple techniques (i.e., shallow, fully-connected artificial neural networks [ANNs]) to establish a benchmark for future ML modeling, to ensure our approach is broadly accessible to the climate community, and to facilitate applying XAI tools. We view this work as a starting point upon which future ML studies focused on the MJO may build. Further, the concept and methods we describe are widely transferable to other areas in Earth science, and may help inform simple ML modeling of other climate phenomena. Section 2 describes the data used in this study. Section 3 describes the ANN models, an ANN explainability method, the linear models we compare the ANN to, and how model skill is assessed. Section 4 describes our results, and Section 5 provides a summary and conclusion.

2. Data

The predictors of our ANN models are latitude-longitude maps of processed tropical variables from 20°N to 20°S. The predictand is the observed Real-time Multivariate MJO index ("RMM"; Wheeler & Hendon, 2004) which tracks the MJO using an empirical orthogonal function analysis of outgoing longwave radiation (OLR), and zonal wind at 850 and 200 hPa. The index consists of two time series ("RMM1" and "RMM2") that represent the strength and location of the MJO. Plotted on a 2-D plane, the RMM phase angle describes the location, or "phase," of the MJO (e.g., Figure 1), while the RMM amplitude ($\sqrt{\text{RMM1}^2+\text{RMM2}^2}$) measures MJO strength. RMM has known limitations (Roundy et al., 2009; Straub, 2013) and other MJO indices exist (e.g., Kikuchi et al., 2012; Kiladis et al., 2014; Ventrice et al., 2013), but RMM represents a logical starting point in this work as it is a widely-used, benchmark MJO index suitable for real-time forecasts.

The tropical input data are from three sources: OLR is from the NOAA Interpolated OLR data set (Liebmann & Smith, 1996), sea-surface temperature (SST) is from the NOAA OI SST V2 High Resolution data set (Reynolds et al., 2007), and all other variables are from ERA-5 reanalysis (Hersbach et al., 2020). Additional data from the ERA-20C data set (Poli et al., 2016) is used in the Supporting Information S1, as described therein. We use daily mean data from 1 January 1979 (1982 for SST) to 31 December 2019 that are interpolated onto a common $2.5° \times 2.5°$ grid.

The input data are divided into training, validation, and testing periods. Training data is used to find the weights/coefficients of the statistical models presented below, validation data is used when tuning model performance, and test data is set aside until the final models are settled upon. Here the training period is from 1 June 1979 to 31 December 2009; the validation data is from 1 January 2010 to 31 December 2015; and the testing is from 1 January 2016 to 30 November 2019. Results assessing the model performance discussed below are evaluated over the testing period, or in a few instances the testing and validation data are used together to increase the sample size, where explicitly noted.

ANN input data are pre-processed in a similar way to the RMM input variables (Wheeler & Hendon, 2004). We subtract the daily climatological mean, first three seasonal-cycle harmonics, and a previous 120-day mean from each point. Variables are not averaged latitudinally because we are interested in how the 2-D structure is utilized

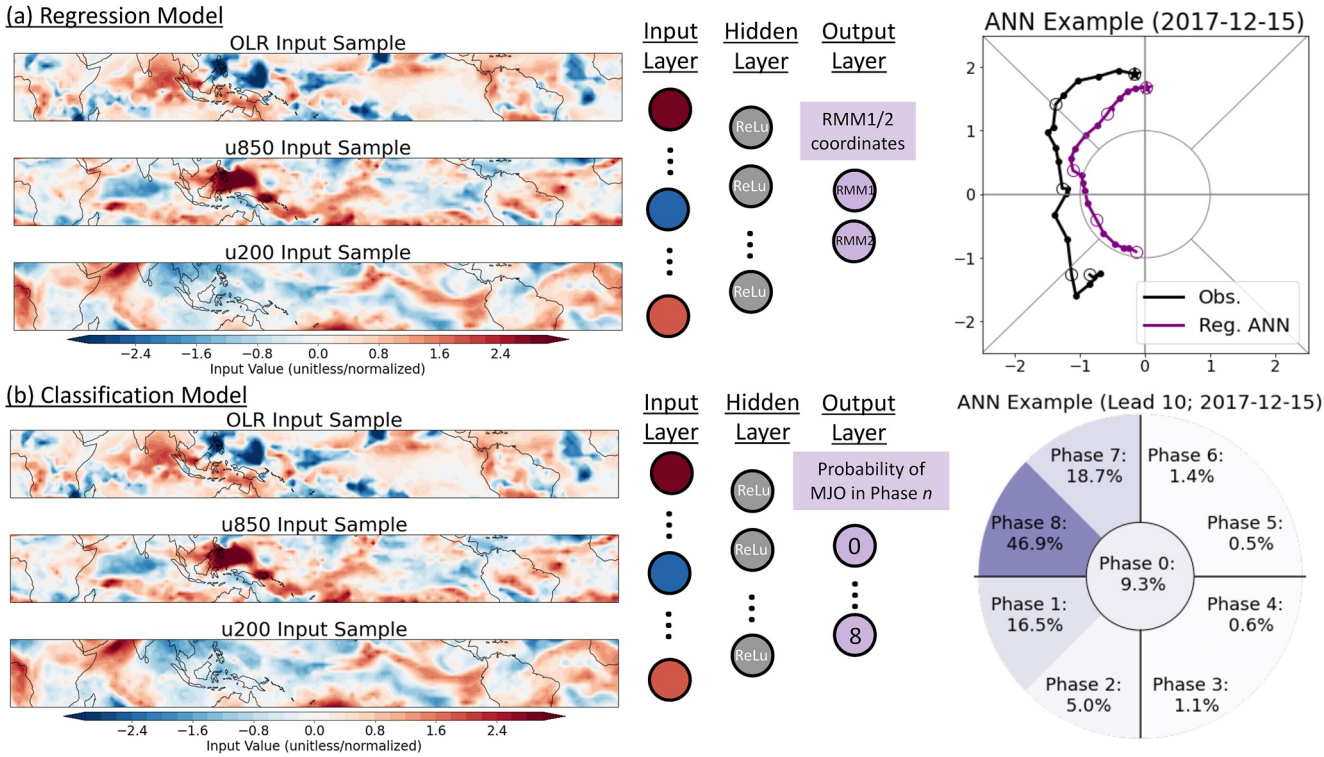

Figure 1. Artificial neural network (ANN) model schematics. (a) The regression ANN; leftmost panels show a sample of processed input outgoing longwave radiation and zonal wind at 850 hPa (u850) and 200 hPa (u200) from 15 December 2017. The input is passed through a 16-node hidden layer with a rectified linear unit ("ReLU") activation function. The regression ANN outputs values of RMM1 and RMM2 at a single lead time, and separate ANNs are trained for leads 0–20 days. An example ANN forecast from lead 0–20 (purple) versus observations (black) is shown in the rightmost panel; dots denote days, with open circles every 5 days. (b) The classification ANN; input is identical to the regression ANN, but the output is the probability the Madden-Julian oscillation is active in Real-time Multivariate MJO index phase 1–8 or is inactive ("phase 0"). An example forecast at a 10-day lead from 15 December 2017 is shown on the right. The model correctly identifies the MJO as in phase 8.

by the ANNs (sensitivity tests exploring latitudinal averaging are discussed in Supporting Information S1). We normalize each variable by subtracting the tropics-wide, all-time mean and dividing by the tropics-wide, all-time standard deviation at each grid point. Tests normalizing each grid point individually showed similar results (not shown). Note that all pre-processing steps (e.g., calculating the climatology, seasonal cycle, and normalization values) are computed only using the training data period, to avoid leakage of information from the validation and testing data into the training data.

In Section 4, the sensitivity of the model to the phase of the stratospheric quasi-biennial oscillation (QBO; Baldwin et al., 2001; Ebdon, 1960; Reed et al., 1961) is assessed. We define the QBO using the monthly, 10°N/S-mean, zonal-mean zonal wind at 50 hPa (U50). Months where U50 is less than the mean minus half a standard deviation are defined as QBO easterly, and months greater than half a standard deviation from the mean are QBO westerly (e.g., Son et al., 2017; Yoo & Son, 2016). Only the November-March period was considered, as a QBO-MJO link is only observed over those months (Martin et al., 2021; Yoo & Son, 2016). To assess whether changes in MJO prediction skill in different QBO phases are statistically significant, we use a bootstrapping test. From all available November-March months in the validation and testing period, we randomly resampled two subsets of months with replacement that have the same size as the observed QBOE and QBOW periods (8 and 12 months, respectively). We compute the change in prediction skill between these two random subsets at each lead time, repeating the process 1,000 times. From that distribution, we assess significance at the 95% confidence level, that is, when the difference between actual QBOE and QBOW periods is less than the 2.5 percentile, or greater than the 97.5 percentile.

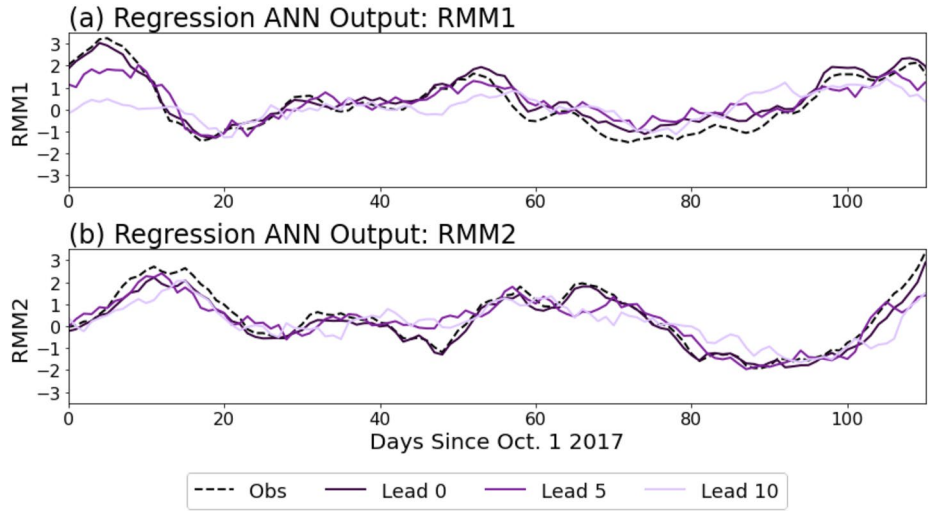

Figure 2. Regression artificial neural network (ANN) example. Example output from the regression ANN during one extended winter season. The observed RMM1 (panel a) and RMM2 values (panel b) are shown in black dashed over a 110-day period beginning 1 October 2017. The regression ANN prediction for each day at a lead of 0, 5, and 10 days are shown in shades of purple.

3. Machine Learning and Linear Statistical MJO Models

We first discuss the two types of ANNs and an ANN explainability technique used in this study. We then describe three conventional statistical MJO models used in prior studies (Jiang et al., 2008; Kang & Kim, 2010; Maharaj & Wheeler, 2005; Marshall et al., 2016) that we compare to the ANNs. We conclude with a brief discussion of how model forecasts are evaluated.

3.1. Artificial Neural Networks

3.1.1. ANN Input, Output, and Architecture

We explored two ANN architectures to study the MJO: a "regression model" and a "classification model" (see summary schematic Figure 1). Both ANN architectures input processed latitude-longitude maps from a single day, and output information about the RMM index N days into the future (Figure 1). Note that inputting tropical maps into the ANN is distinct from the majority of statistical MJO models, which typically input values of the RMM index or a limited number of principal components (Jiang et al., 2008; Kang & Kim, 2010; Waliser, 2012). Using the ANNs in the present manner allows the 2-dimensional structure of a range of different combinations of input variables to be used in the model. In this work we focus on ANNs that input between 1 and 3 different variables. In particular, in this section and Section 4.1 we use ANNs that input three variables simultaneously: OLR, zonal wind at 850 hPa, and zonal wind at 200 hPa (Figure 1). This combination is among the best-performing across the experiments we conducted and uses the variables that comprise RMM. Exploration of other variables is described in more detail in Section 4.2.

For both regression and classification ANN architectures, a separate ANN is trained for each lead time N from 0 to 20 days. The difference between the regression and classification ANNs is the nature of their outputs. The regression ANN (not to be confused with a linear regression model) outputs RMM1 and RMM2 values (i.e., a vector of two real numbers). Examples of regression ANN output are shown in Figures 1a and 2. Figure 1a shows an example prediction in RMM phase space for a 20-day forecast in the ANN compared to observations. Figure 2 shows lead 0, 5, and 10-day predictions on each day over a particular winter period for RMM1 and RMM2.

In contrast to the regression model, which is deterministic, the classification ANN provides probabilistic forecasts. The classification ANN outputs the probability that the MJO at a given lead time is in each of nine classes (e.g., Figures 1b and 3): either active (RMM amplitude ≥ 1) in one of the eight canonical RMM phases (Wheeler & Hendon, 2004) or weak ("phase 0"; RMM amplitude <1). The class with the highest probability is considered

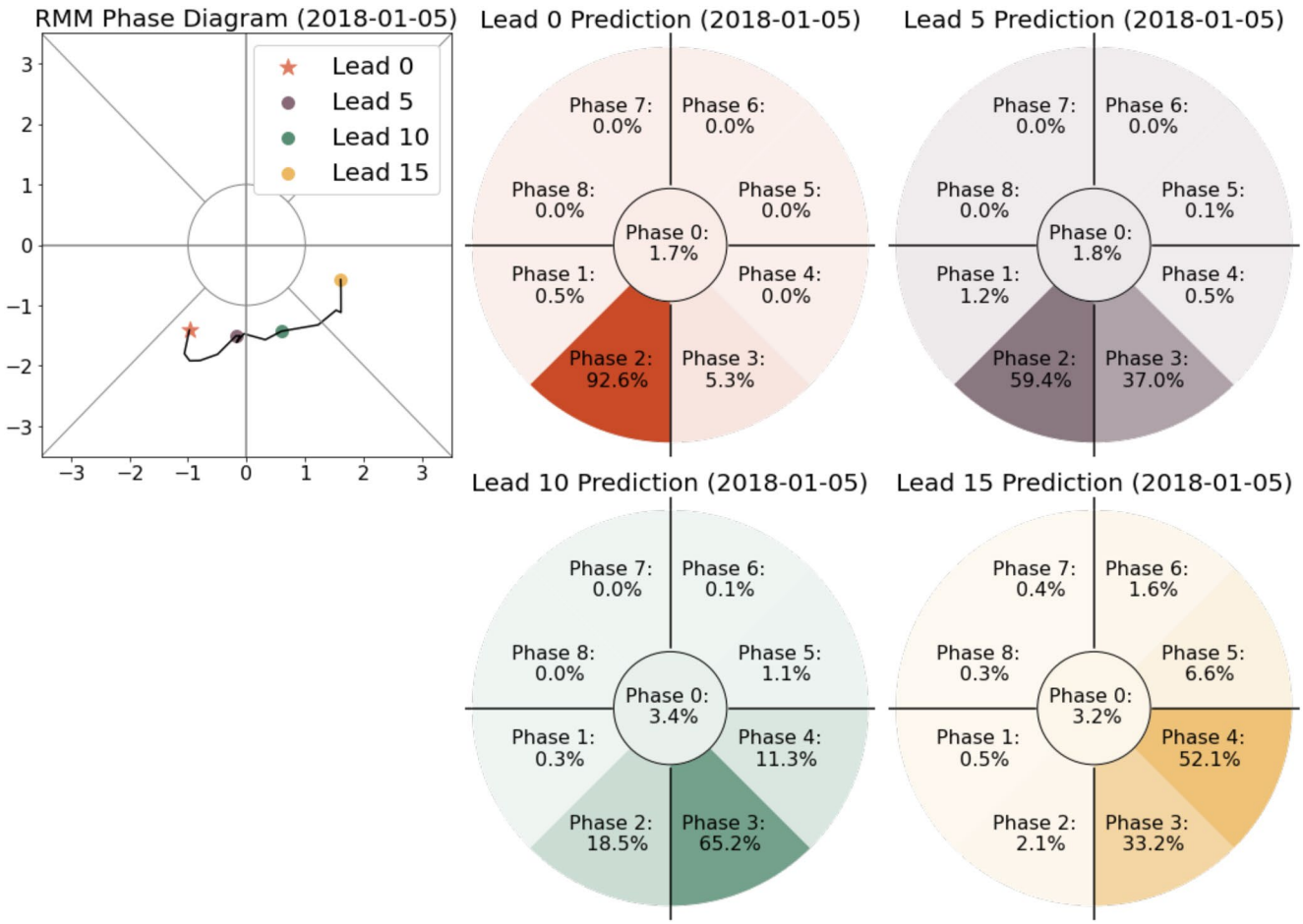

Figure 3. Classification artificial neural network (ANN) example forecast. Output from the classification ANN for lead times of 0, 5, 10, and 15 days. The left panel shows the observed Real-time Multivariate MJO index for 20 days beginning 5 January 2018. The right four panels show the classification ANN confidence for each of the 9 Madden-Julian oscillation phases at the indicated lead time. The class with the highest probability is considered the prediction; in this example predictions are phase 2 (lead 0; correct), phase 2 (lead 5; correct), phase 3 (lead 10; correct), and phase 4 (lead 15; correct).

the predicted class. An example of the classification ANN output for one initialization date at four different lead times is shown in Figure 3 alongside the observed RMM index.

Both the regression and classification ANNs are simple, shallow, fully-connected neural networks. Both architectures have one layer of 16 nodes that use a rectified linear activation function ("ReLU"). For the regression ANN, the loss function is the mean-squared error, while the classification ANN loss function is the categorical cross-entropy, with a softmax operator applied to the output to normalize class probabilities so predictions sum to 1. To help prevent overfitting, both ANN architectures use ridge regularization (an L_2 norm penalty) to limit the weights of the hidden layer. Both architectures also use early-stopping during training, which monitors the loss on the validation data and stops training once the validation loss plateaus (or increases) for a specified number of epochs. The total number of epochs used to train a given ANN varies by model type and lead time. In general, the regression model at short leads tended to train for approximately 150 epochs while at longer leads trained for around 30–40 epochs. For the classification model, training at short leads took approximately 180 epochs, and at longer leads stopped after around 100 epochs.

For the classification ANN, since weak MJO days are the most common class (~39% of all days) we avoid class imbalance by randomly subsampling weak MJO days during training so they are 11% of all training days. Weak days are not subsampled over the validation or testing periods. Values of key hyperparameters used in both architectures and additional model details are listed in Table 1. Sensitivity tests varying ANN parameters and input

Table 1
Regression and Classification Neural Network Model Architecture Details and Key Hyperparameters Used in This Study

ANN model details and hyperparameters		
Name	Regression ANN value	Classification ANN value
Winter/summer training samples	5,560/5,612	3,990/3,726
Winter/summer validation and test samples	1,093/1,098	1,093/1,098
Hidden layer size	16 nodes	16 nodes
Activation function	ReLU	ReLU
Optimizer	Stochastic gradient descent	Stochastic gradient descent
Loss function	Mean-squared Error	Categorical cross-entropy
Learning rate	0.0005	0.0005 (0.001 for 1-variable models)
Batch size	32	32
Ridge penalty	0–5 day leads: 0.25 6–10 day leads: 1 11+ day leads: 3	0.25 (all leads)
Early-stopping patience	8 epochs	4 epochs

Note. Sensitivity tests to various aspects of these and other aspects of the ANN models are discussed in the Supporting Information S1. ANN, artificial neural network.

data were explored, and while the present configuration was optimal across the tests conducted, results from a subset of our sensitivity tests are discussed in the Supporting Information S1.

ANN performance is slightly improved if the models are trained separately on different seasons (Figure S1 in Supporting Information S1), which allows the ANNs to learn more season-specific patterns. This is likely important for the MJO due to its seasonal shifts in behavior, strength, and structure (Hendon & Salby, 1994; Hendon et al., 1999; Zhang & Dong, 2004), and we found splitting the data into two 6-month periods (October-March, or herein "winter," and April-September, or "summer") provided a good trade-off between seasonal specificity and number of training samples.

Finally, in some instances we trained multiple ANNs for the same season and lead time, creating an "ANN ensemble." The ANNs in the ensemble are distinct only in the random initial training weights; otherwise the training data and architecture is the same across all ANNs. The ensemble thus allows us to check for convergence of our results during ANN training, and quantifies sensitivity to ANN initialization.

3.1.2. Layer-Wise Relevance Propagation

To demonstrate how the classification ANN correctly captures regions of importance for predicting the MJO, we use an ANN explainability technique called layer-wise relevance propagation (LRP; Bach et al., 2015; Montavon et al., 2019; Samek et al., 2016). LRP has been used in Earth science as a tool for understanding the decision-making process of ANNs (Barnes et al., 2020; Madakumbura et al., 2021; Mamalakis et al., 2021; Mayer & Barnes, 2021; Toms et al., 2019, 2020), and here, we provide a high-level overview.

Broadly, LRP is an algorithm applied to a trained ANN. After a particular prediction is made, LRP back-propagates that prediction's output through the ANN in reverse. Ultimately, LRP returns a vector of the same size as the input (here a latitude-longitude map of one or more variables), where the returned quantity, termed the "relevance," shows which input points were most important in determining that prediction. By construction, LRP relevance maps are unique to each input sample, not each output class.

We use LRP to analyze output from the classification ANN. There are several different implementation rules for LRP, which differ in the details of how they back-propagate information (see Bach et al., 2015; Mamalakis et al., 2021; Montavon et al., 2019; Samek et al., 2016). Based on results in Mamalakis et al. (2021) assessing various implementations of LRP in a synthetic data set, we use the "LRP_z" method, which in their case performed well compared to other implementations of LRP. The LRP_z method returns both positive and negative relevance values, but because we are interested in regions that positively contribute to correct predictions, we take

only regions of positive relevance in each sample. Overall conclusions are not changed if negative relevance is included (not shown). To ensure each sample contributes equally to the composite plots in Section 4.2, we normalize each LRP heat map by dividing by its maximum.

3.2. Traditional Linear MJO Models

We compare ANN performance to three established, statistical MJO models: a persistence model, a vector autoregressive (VAR) model, and a multi-linear regression (MLR) model.

The persistence model is often used as a minimal benchmark for statistical MJO model performance, and forecasts RMM1 and RMM2 values by persisting the initial condition. For a forecast beginning at time t_0, at each lead time τ the persistence model forecasts:

$$[\mathrm{RMM1}(t_0+\tau), \mathrm{RMM2}(t_0+\tau)] = [\mathrm{RMM1}(t_0), \mathrm{RMM2}(t_0)]$$

While useful as a benchmark of statistical MJO models, the persistence model is unrealistic, in that it neglects the well-understood propagating nature of the MJO. Thus, we focus primarily on comparing the ANN to two more complex statistical models which better account for the propagating nature of the MJO.

The VAR model (Maharaj & Wheeler, 2005; Marshall et al., 2016) is a linear model which inputs RMM values for a given day and predicts RMM values 1 day into the future. Following Maharaj and Wheeler (2005), this is formulated as:

$$[\mathrm{RMM1}(t_0+1), \mathrm{RMM2}(t_0+1)] = L_{\mathrm{var}}\,[\mathrm{RMM1}(t_0), \mathrm{RMM2}(t_0)]$$

L_{var} is a matrix calculated using a multiple linear regression fit from the training data. As with the ANNs, and following Maharaj and Wheeler (2005), we compute L_{var} separately for winter and summer periods using the same training period as the ANNs. Coefficients of L_{var} match closely with those described in the literature (Maharaj & Wheeler, 2005; Marshall et al., 2016), differing slightly due to our different training period and definition of winter and summer. VAR model forecasts are initialized with the observed RMM1/2 values, and then the initial conditions are stepped forward 1 day at a time out to a lead time of 20 days.

The final simple model, the MLR model (Jiang et al., 2008; Kang & Kim, 2010; Wang et al., 2019), generally follows Kang and Kim (2010), who showed across several statistical models that the MLR model performed best at predicting RMM. The model can be written as:

$$[\mathrm{RMM1}(t_0+\tau), \mathrm{RMM2}(t_0+\tau)] = L_{MLR,\tau}\,[\mathrm{RMM1}(t_0), \mathrm{RMM2}(t_0), \mathrm{RMM1}(t_0-1), \mathrm{RMM2}(t_0-1)]$$

$L_{MLR,\tau}$ is a matrix of coefficients calculated using a multiple linear regression fit from the training data. The main differences from the VAR model are the MLR model inputs RMM values on the initial day and 1 day prior, and predicts the RMM1/2 values at a specified lead time of τ. As with the ANNs, we train separate MLR models for each lead time and in winter and summer.

3.3. Model Assessment Metrics

To assess model skill in the regression ANN, we utilize the bivariate correlation coefficient (BCC; e.g., H. Kim et al., 2018; Vitart et al., 2017), with a value greater than 0.5 used to denote skill. In the classification ANN, skill is measured using the model's accuracy as well as probability-based skill scores. Following Marshall et al. (2016), who examined probabilistic MJO forecasting in a dynamical model framework, we assess skill at predicting MJO phase using the ranked probability skill score (RPSS). We first calculate the ranked probability score (RPS) for a given statistical model for each lead time as:

$$\mathrm{RPS}_{\mathrm{model}} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{M-1}\sum_{m=1}^{M}\left[\left(\sum_{k=1}^{m} p_k\right) - \left(\sum_{k=1}^{m} o_k\right)\right]^2\right\}$$

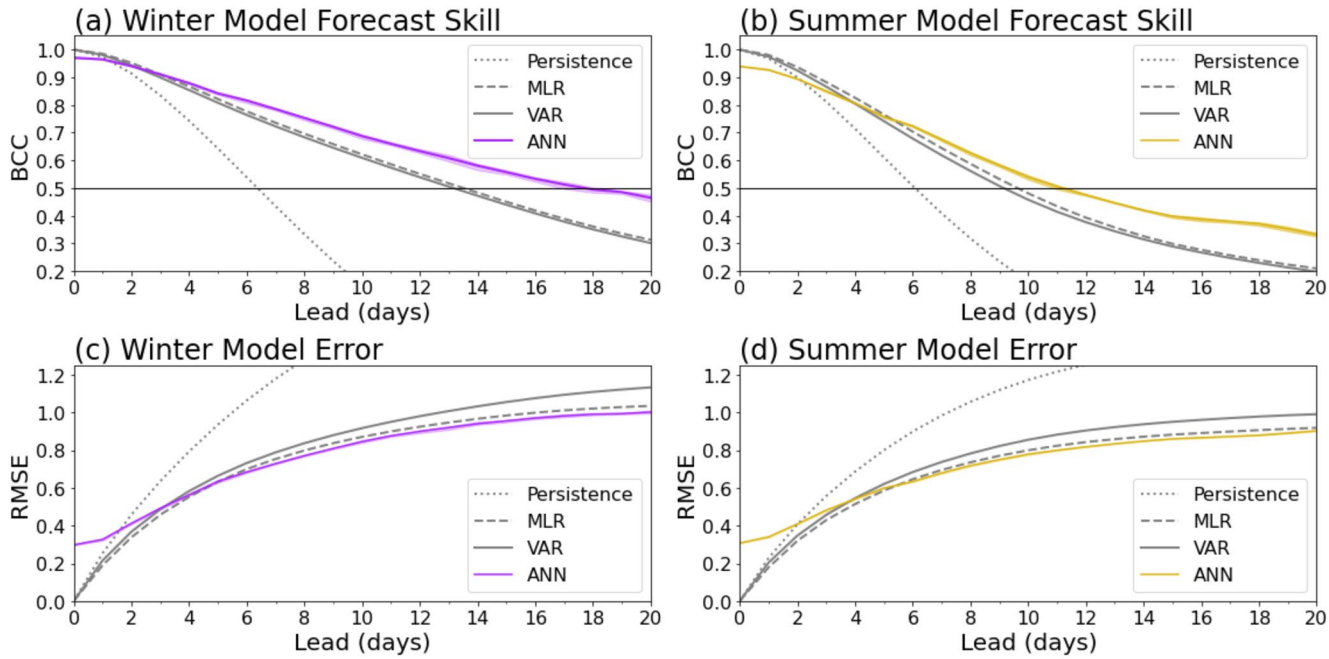

Figure 4. Regression artificial neural network (ANN) overall performance. Real-time Multivariate MJO index prediction skill (a and b) and root-mean-square error (c and d) for the regression ANN (purple/gold) and other simple statistical models (gray). Skill in the top panels is measured via the bivariate correlation coefficient; a threshold of 0.5 denotes skill.

Here, N is the number of forecasts, M is the number of MJO classes (9), p_k is the forecast probability in a given MJO class, and o_k is the observed probability (i.e., 1 for the observed phase and 0 for all other phases). Following Marshall et al. (2016), we order the m categories from phase 0 to 8, which captures the canonical MJO phase evolution. When the RPS is calculated for the classification ANN, p_k is the model confidence for each phase. For the MLR or VAR model, p_k is 1 for the predicted phase and 0 otherwise.

We compute a climatological reference RPS, denoted RPS_{ref}, by calculating the percentage of days the observed MJO is in phases 0–8 across the training data, and using those percentages as p_k values across all N forecasts. The RPSS for a given model is then computed as:

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS}_{\mathrm{model}}}{\mathrm{RPS}_{\mathrm{ref}}}$$

An RPSS greater than 0 indicates a given model shows better skill than climatology.

4. Results

4.1. Overall Model Performance

In this subsection, we use ANNs that input OLR, zonal wind at 850 hPa, and zonal wind at 200 hPa simultaneously (Figure 1) and initialize forecasts daily.

Overall, the winter and summer regression ANNs show prediction skill, respectively, of ∼18 and ∼11 days (Figure 4), with small spread across a 10-member ANN ensemble. In both seasons, regression ANNs outperform all three of the linear statistical models after 3–4 days in winter and 4–5 days in summer, showing substantially better skill than persistence and modestly better skill than the MLR and VAR models. The ANNs also demonstrate a lower root-mean-square error than other statistical models (Figure 4) indicating that MJO amplitude in both seasons is better captured, though the RMS improvement compared to the MLR model is small. This places simple ANNs at the forefront of statistical MJO prediction techniques, which is impressive given the simplicity of the ANNs and the fact that no explicit information about the RMM index is passed to the networks. The improved performance of the ANN relative to the VAR model further demonstrates that ANNs learn more than just to

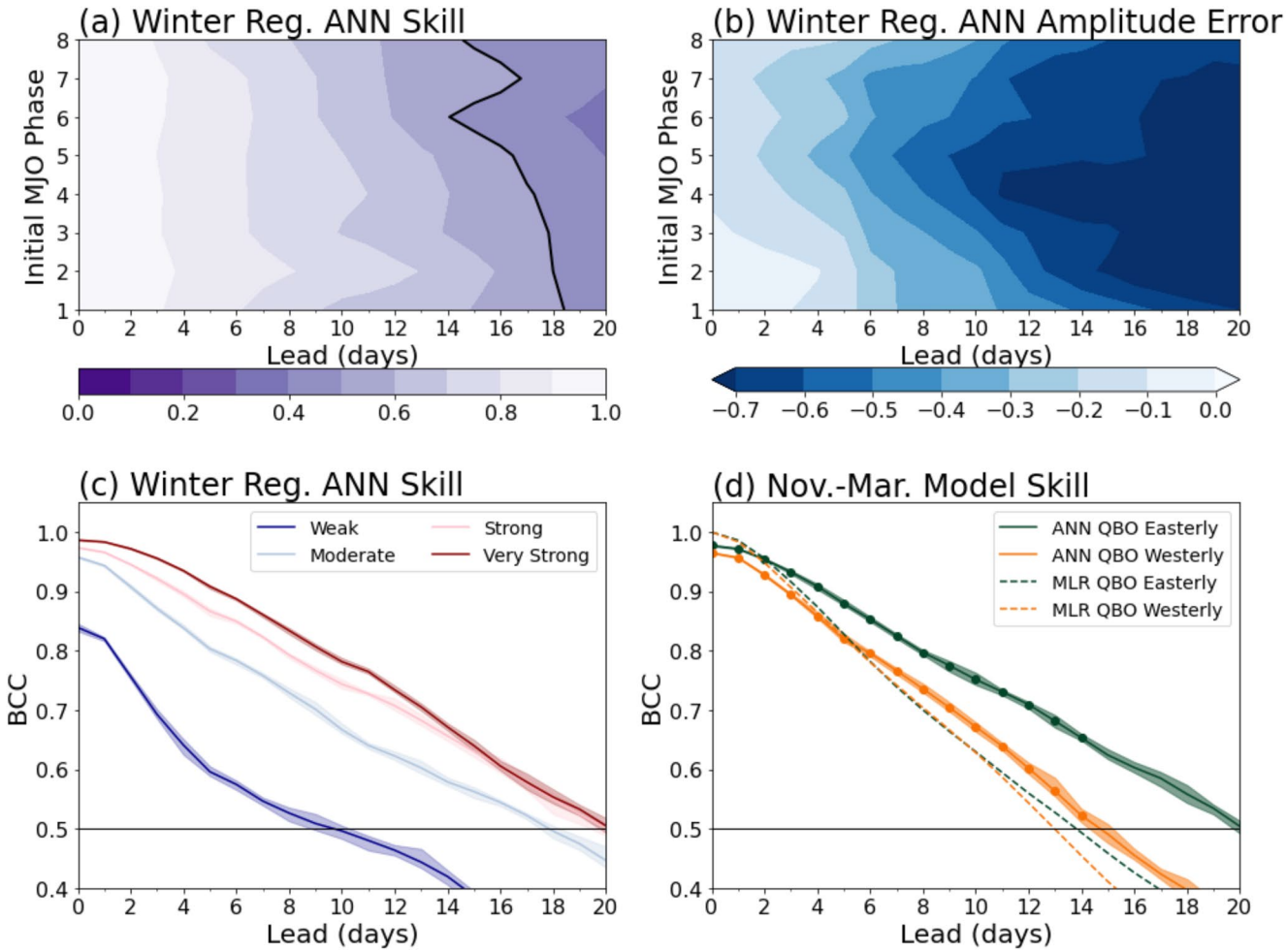

Figure 5. Regression artificial neural network (ANN) detailed performance. (a) The bivariate correlation coefficient (BCC) as a function of initial Madden-Julian oscillation (MJO) phase, without a threshold for MJO activity (i.e., all days are assigned a phase 1–8). Black line denotes a BCC of 0.5. (b) The average Real-time Multivariate MJO index (RMM) amplitude difference between observations and ANN-forecasted events: negative values indicate the ANN prediction is weaker than observed. (c) BCC for winter forecasts binned by observed initial MJO amplitude. Initial RMM amplitude ranges are 0–1 (weak); 1–1.5 (moderate); 1.5–2; (strong) and greater than 2 (very strong). (d) BCC for MJO events in November-March separated by phase of the stratospheric quasi-biennial oscillation. ANN results are solid and multi-linear regression results are dashed; dots in indicate differences are statistically significant at the 95% level via a bootstrapping test. Shading in panels (c and d) denotes the spread across a 10-member ANN ensemble.

identify the MJO in RMM space and then propagate it east. Further, the higher skill in winter versus summer is consistent with results in most dynamical models (e.g., Vitart, 2017), and is one indication that ANNs are able to reproduce aspects of MJO predictability seen in more complex dynamical models. While linear models also show higher skill in winter than summer, the relative increase between the two seasons is larger for the ANN.

To explore regression ANN performance in more detail, we conditioned forecasts based on various aspects of the initial condition, like MJO phase and strength (Figure 5). For this plot, the sample size was increased by including both the validation and testing data in the analysis. The regression ANN skill shows relatively small sensitivity to initial MJO phase (Figure 5a), with somewhat higher skill (~18–19 days) across MJO events initialized in phases 1–3 and lower skill (~14–15 days) for phases 6 and 8. In contrast to the initial phase, the regression ANN shows substantially more sensitivity to initial MJO amplitude: MJO events that are initially strong or very strong (RMM amplitude >1.5) are skillfully predicted out to ~20 days in winter, while skill predicting weak winter events is only ~10 days (Figure 5c). This change in skill based on MJO initial condition is consistent with findings in other statistical and dynamical models (H. Kim et al., 2018). Note that one consequence of the low performance predicting weak MJO events is that the regression ANN struggles to successfully predict MJO initiation. Correctly

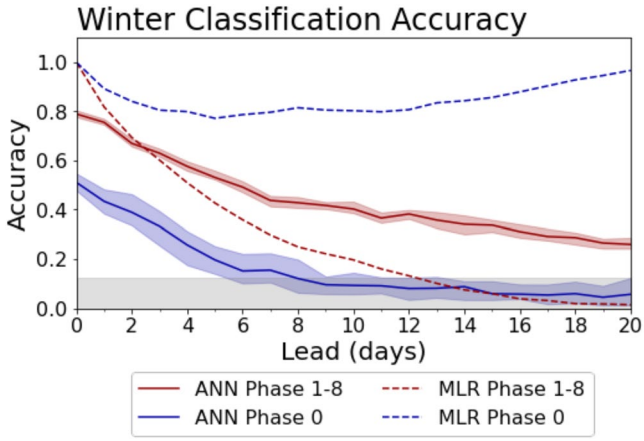

Figure 6. Classification model accuracy. Winter classification artificial neural network (ANN) accuracy forecasting active Madden-Julian oscillation (MJO) days (phase 1–8; red) and weak MJO days (phase 0; blue). Dashed line is the same but for the multi-linear regression model. Gray shading indicates random chance (1/9) or less, assuming all classes are equally likely. Blue and red shading denotes the spread across a 10-member ANN ensemble.

predicting the onset of the MJO is a recognized challenge in MJO research and forecasting in general (Ling et al., 2017), and in particular improving this aspect of model performance may be a target for future work.

ANNs also capture more mysterious aspects of MJO predictability, such as the sensitivity to the phase of the stratospheric QBO (Marshall et al., 2017; Martin et al., 2021). Studies using both dynamical and statistical models have found improved MJO prediction skill in QBO easterly months compared to QBO westerly months during November-March (NDJFM; H. Kim et al., 2019; Lim et al., 2019; Marshall et al., 2017; Wang et al., 2019). Defining the QBO using the U50 index, the regression ANN skill during QBO easterly NDJFM periods is nearly 20 days, whereas during QBO westerly periods skill is only 15 days (Figure 5d). This modulation is quantitatively consistent with findings in dynamical models (H. Kim et al., 2019; Lim et al., 2019), however, it is important to note the number of QBO cycles is limited since only winters from 2010 to 2019 are considered in Figure 5d.

A bootstrap analysis of the change in prediction skill between QBOE and QBOW periods in Figure 5 nevertheless found that differences are statistically significant out to 14 days, but not at longer lead times. This lack of significance at leads beyond 2 weeks is consistent with other studies that have noted the limited significance of a QBO impact on MJO prediction skill at longer leads (e.g., H. Kim et al., 2019). A similar analysis using the MLR shows a much smaller QBO modulation over the 2010–2019 period, which is not significant at any lead time (Figure 5d). This is different from the findings of Wang et al. (2019), who showed a clear QBO modulation of MJO prediction skill in an MLR model, but their study examined a longer 1979–2019 period, and used a different MJO index. Thus, while the results here highlight the possibility for the ANN model to detect signals that simpler linear models may not capture, a more detailed and focused study of the impact of the stratosphere on the MJO in an ANN modeling framework is needed, especially one that considers a longer period of time. One approach may be to fluctuate training and testing periods to allow consideration of more QBO cycles in the ANN and linear models, a method that could be explored in future work.

Overall, a strength of the regression ANN is the quantitative information it provides about MJO phase and strength. Further, the regression ANN may prove an efficient framework in which to continue to examine aspects of MJO predictability discussed above, like sensitivity to initial MJO amplitude and phase of the QBO. But a prevalent source of error in the regression ANN is a decrease in the ANN-predicted MJO amplitude at lead times longer than a few days, especially in phases 4–7 (Figure 5b). Amplitude bias is also an issue in the VAR and MLR model, and continuing to explore ways in which it might be overcome in an ANN model is an open challenge. However, this amplitude bias was one motivation for exploring a classification ANN architecture that focuses more directly on MJO phase. Further, the probabilistic nature of the classification ANN makes it a unique simple statistical tool for MJO forecasting.

Assessed via model accuracy, a 10-member classification ANN ensemble performs well predicting active MJO events in RMM phases 1–8 (Figure 6), outperforming the MLR and VAR statistical models after approximately 2–3 days, with accuracy during days 7–20 approximately 20% higher (Figure 6; only the MLR model is shown as VAR results are similar). At lead 0, where the classification model is identifying the MJO, the phase of active MJO events are correctly predicted with an accuracy of ∼80% (Figure 6), comparable to Toms et al. (2019), despite differences in our input variables, data pre-processing, MJO index, and ANN complexity. The majority of incorrectly predicted active MJO events at short leads are near the boundary between two RMM phases and predictions are often incorrect by only one phase.

While classification ANN skill is substantially better than linear models at predicting active MJO events, it struggles to predict weak MJO days, with an accuracy at short leads of only ∼50%, which falls to near random chance after ∼1 week (Figure 6). This is in part due to the training strategy of the classification ANN: by subsampling weak days during training to prevent class imbalance, the classification model learns not to overemphasize the weak phase. This tendency of the classification ANN to underpredict weak MJO events is in contrast to simple

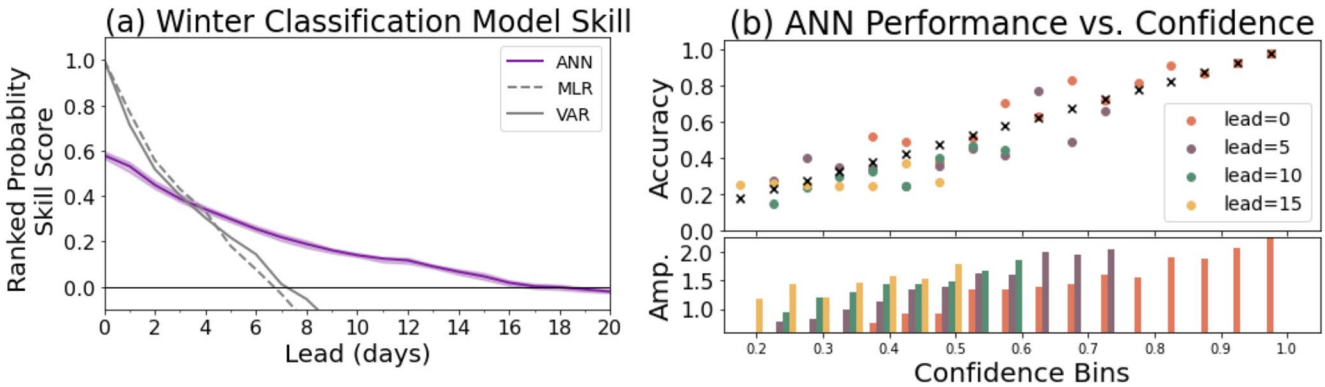

Figure 7. Classification model probabilistic forecasting. (a) The ranked probability skill score in winter for the artificial neural network (ANN), multi-linear regression, and vector autoregressive model predictions relative to climatology; a score greater than zero denotes skill. (b) Winter classification ANN accuracy (top panel) and initial observed Madden-Julian oscillation amplitude (bottom panel) binned by ANN confidence (*x*-axis, in bins of width 0.05) at leads of 0, 5, 10, and 15 days. The black *x*'s in the top panel indicate the one-to-one line.

linear models. The MLR model, for example, has a very high accuracy predicting weak MJO events (Figure 6). At early leads this is because the initial RMM phase is given to the model, and at longer leads the MLR model simply categorizes all MJO events as weak.

Assessing the ANN only via accuracy fails to take full advantage of this model's probabilistic forecasts however. This aspect of the classification ANN is distinct from the deterministic output provided by linear models or even some dynamical models, though Marshall et al. (2016) showed how ensemble runs of dynamical models could be used to provide probabilistic MJO forecasts. Assessing the ANN and linear models via the RPSS (Figure 7a), the classification model performance is clearly superior. The ANN skill remains greater than climatology out to ∼16–18 days in winter (comparable to the regression model skill assessed via the BCC), while the deterministic linear models show skill to about one week. This demonstrates that the classification ANN provides probabilistic information that is useful and adds to the model skill past what deterministic schemes can provide.

Model confidence has clear utility for forecasters and could drive future work in probabilistic MJO prediction (Marshall et al., 2016). It further may be useful in improving understanding of MJO predictability. For example, the classification ANNs probabilistic forecasts are reliable—in the sense that ANN confidence corresponds well with model accuracy—which indicates that model confidence is a useful and meaningful output in this work (Figure 7). Furthermore, ANN confidence relates to physical aspects of the MJO: we found ANN confidence is closely associated with initial MJO amplitude (correlation coefficients of ∼0.5–0.7 depending on lead), with higher confidence associated with higher initial RMM amplitude (Figure 7b). Research using ANN confidence to identify predictable states of the atmosphere has recently shown promise, including in the context of MJO teleconnections to the extra-tropics (Barnes et al., 2020; Mayer & Barnes, 2021).

The tradeoffs between the simple classification and regression ANN architectures we explored here make choosing a "better" model difficult, and in presenting both we illustrate their respective strengths and limitations. The regression model outputs more precise RMM information and is more readily comparable to existing models, but struggles to predict strong MJO amplitudes at long leads or MJO initiation. This is true even when the regression model was re-trained using fewer weak MJO days to emphasize strong MJO events: little change in performance was seen (Figure S2 in Supporting Information S1). The classification ANN shows the opposite tendency, overestimating the percentage of active MJO days and struggling to accurately predict weak MJO events. And while the classification ANN cannot provide precise information about MJO strength and location it provides a unique probabilistic output compared to other simple statistical models of the MJO.

Overall, results for both ML architectures show that aspects of the MJO are skillfully predicted by several metrics beyond 2 weeks in winter, and the ANNs outperform existing linear statistical models. A range of sensitivity tests (Text S1 and Figures S3–S5 in Supporting Information S1), including increasing the amount of training data using twentieth-century reanalysis and including additional days in the input, showed comparable performance, though tests were not exhaustive nor explored beyond relatively simple ANN architectures. Also note that while our primary goal here is to introduce and establish a baseline for ML modeling of the MJO, the simple

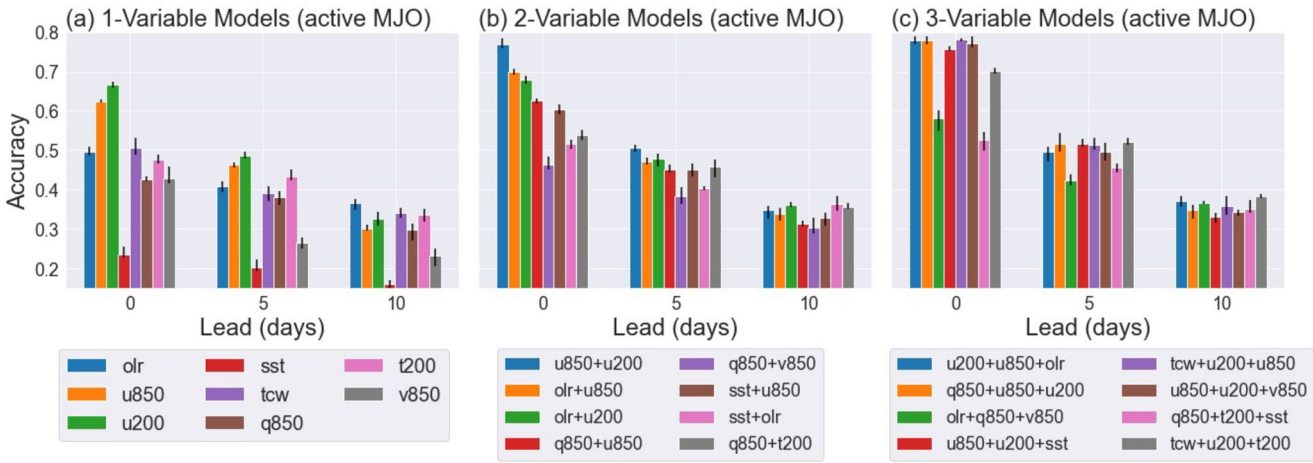

Figure 8. Sensitivity to input variables. Winter classification artificial neural network (ANN) accuracy predicting active Madden-Julian oscillation days at leads of 0, 5, and 10 days given different input variables. 1-variable (panel (a)), 2-variable (panel (b)), and 3-variable (panel (c)) models are shown. For each model, 10 ANNs are trained with different initial random weights (error lines). The legend indicates which variables are used; short-hand refers to zonal wind (u), total column water vapor (tcw), specific humidity (q), temperature (t), sea-surface temperature (SST), and meridional wind (v), with numbers indicating the pressure level where relevant.

ANNs we explored are not yet competitive with most S2S dynamical forecast models (e.g., H. Kim et al., 2018; Vitart, 2017). State-of-the-art dynamic model skill predicting the MJO generally falls between 25 and 35 days when assessed via the BCC (H. Kim et al., 2018; Vitart, 2017), and probabilistic MJO forecasts formed by running ensembles of dynamical models show skill via the RPSS out to approximately 25 days in one S2S model (Marshall et al., 2016). It remains to be seen whether future ML research might improve to the point where it is competitive with dynamical models, but as the next section illustrates, even the simple ANNs introduced here can be used as a tool for more than just prediction, and may help spur new discoveries or generate new hypotheses.

4.2. Experimentation and Explainability of ANN Models

A limiting aspect of many standard MJO statistical prediction models, including the persistence, VAR, and MLR models presented here, is they rely entirely on an MJO index as input. In contrast, the ANNs we utilize learn relationships between latitude-longitude maps of one or more tropical variables and an MJO index, meaning that the statistical relationships they learn connect the spatial patterns and interrelationships of the input variables to the behavior of the MJO at various lead times. This flexible framework allows for more experimentation across input variables and input processing strategies than existing approaches, allowing us to explore the impact of different variables on MJO prediction skill. In addition, this framework in conjunction with XAI techniques further illuminates what aspects and spatial regions of the input variables are most important for the model's predictions.

We first illustrate this through classification ANN experiments inputting various combinations of one to three different variables, targeting leads 0, 5, and 10 days for brevity. Overall, model accuracy varies widely depending on input (Figure 8). For example, across 1-variable ANNs (Figure 8a) 850 hPa meridional wind and SST models show much poorer performance than other inputs. In the case of the SST model, this suggests the ocean state alone (when processed to highlight subseasonal variability) does not contain MJO signals the ANN is able to leverage, consistent with findings that sub-seasonal SST variability does not drive the MJO (e.g., Newman et al., 2009). In the case of meridional wind, while the MJO possesses signals in meridional wind associated with Rossby wave gyres (Zhang, 2005), we hypothesize that skill may be low because these signals lack the global-scale coherence seen in variables like zonal wind and OLR that are captured by RMM.

The most accurate models at short leads are those that input 850 hPa and/or 200 hPa zonal winds (Figure 8). This is consistent with literature showing that MJO circulation tends to drive the RMM index (Straub, 2013; Ventrice et al., 2013), an aspect of RMM the ANN has organically learned. Interestingly, skill identifying the MJO at short leads does not necessarily imply similar performance predicting the MJO at longer leads. For example, at lead 0 the 850 and 200 hPa zonal wind model has the highest accuracy among 2-variable models (Figure 8b), but at lead 5 and 10 its accuracy overlaps with other combinations of variables. Best performing models at longer leads are

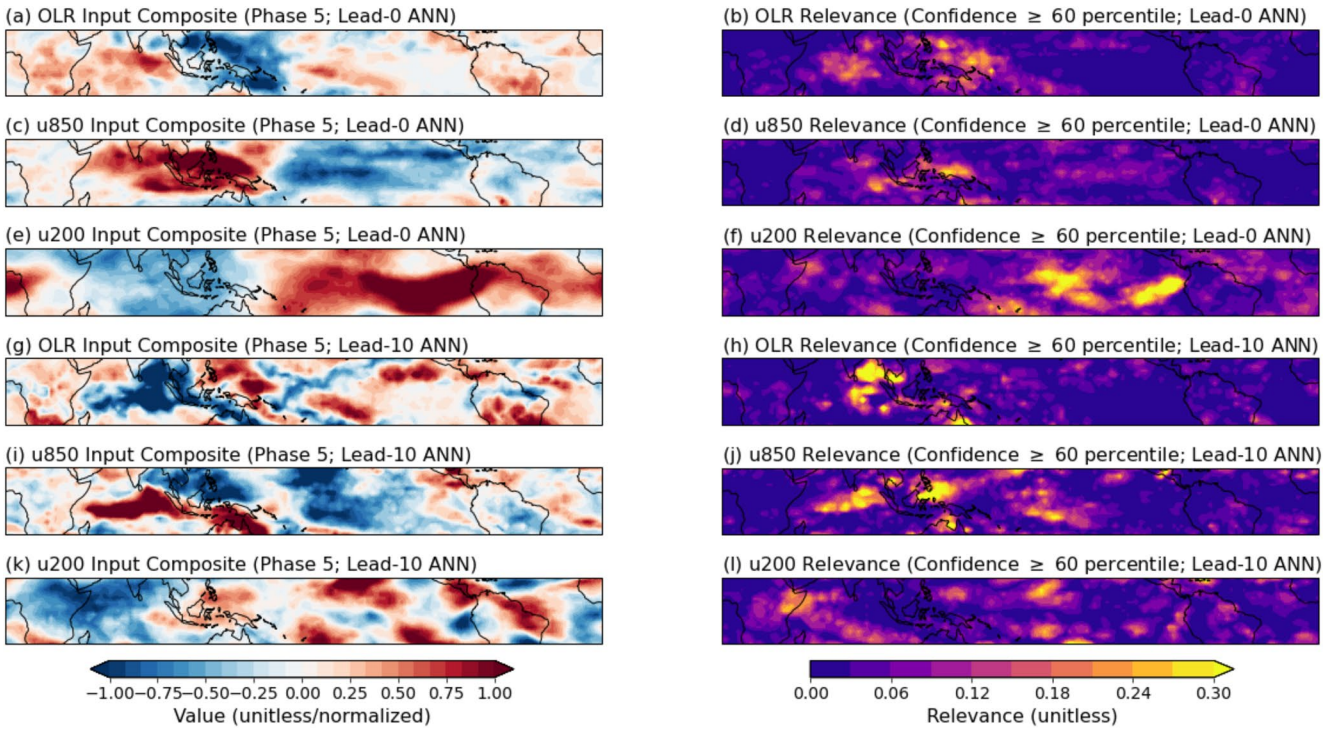

Figure 9. Layer-wise relevance propagation example. Composites of normalized input variables (left column) and layer-wise relevance propagation relevance (right column) for correct classification artificial neural network predictions of Madden-Julian oscillation events in Phase 5 at the time of verification. Only forecasts when model confidence exceeds the 60th percentile are included. Panels (a–f) are the lead-0 model, and (g–l) are the lead-10 model, both inputting three variables: outgoing longwave radiation, and 850 hPa zonal wind (u850) and 200 hPa zonal wind (u200).

those that include information about zonal wind and the large-scale thermodynamic or moisture signature of the MJO, as measured for example, by OLR or column water vapor (Figure 8c). Further, the three RMM input variables are not always clearly best performing at leads of 5 and 10 days: a model with total column water, 200 hPa zonal wind and 200 hPa temperature performs as well as or slightly better than the model with 200 and 850 hPa zonal wind and OLR (Figure 8c).

Finally, while more input variables tend to improve model performance (Figure 8), tests showed no substantial improvement using four or more inputs (Figure S5 in Supporting Information S1), at least among the variables considered here. Whether this is due to the limited complexity of our ANNs, the amount of training data, or because new, meaningful information is difficult to leverage with more variables is not known. Additional variables (perhaps with different preprocessing) will continue to be explored, but these initial tests provide a proof-of-concept for the kind of experimentation that ANNs afford.

Another advantage of ANNs versus other MJO modeling frameworks is the ability to apply XAI tools like LRP (Section 3.1.2), which identifies sources of ANN prediction skill. As a first example, Figure 9 shows wintertime composite LRP maps using the classification ANN from Section 4.1. LRP maps are shown for lead times of 0 and 10 days, composited across correct ANN predictions when the MJO is in phase 5 at the time of verification. Composites are further restricted to those events when model confidence exceeds the 60th percentile (calculated from the full distribution of model confidence for each lead, not the distribution only over correct predictions).

The LRP plots confirm that the classification ANN focuses on regions central to the MJO. At lead 0, OLR relevance highlights suppressed Indian Ocean convection and active conditions around the Maritime Continent (Figures 9a and 9b), whereas wind fields focus on low-level westerly anomalies around the Maritime Continent (Figures 9c and 9d) and upper level signals in the central and east Pacific (Figures 9e and 9f), all of which are hallmark features of a phase 5 MJO. At lead 10, LRP shows how the ANN accounts for eastward MJO propagation: the maximum relevance for OLR is shifted west relative to lead 0, highlighting strong convection in the eastern Indian Ocean (Figures 9g and 9h). The lead-10 model also focuses on a dipole region of strong low-level

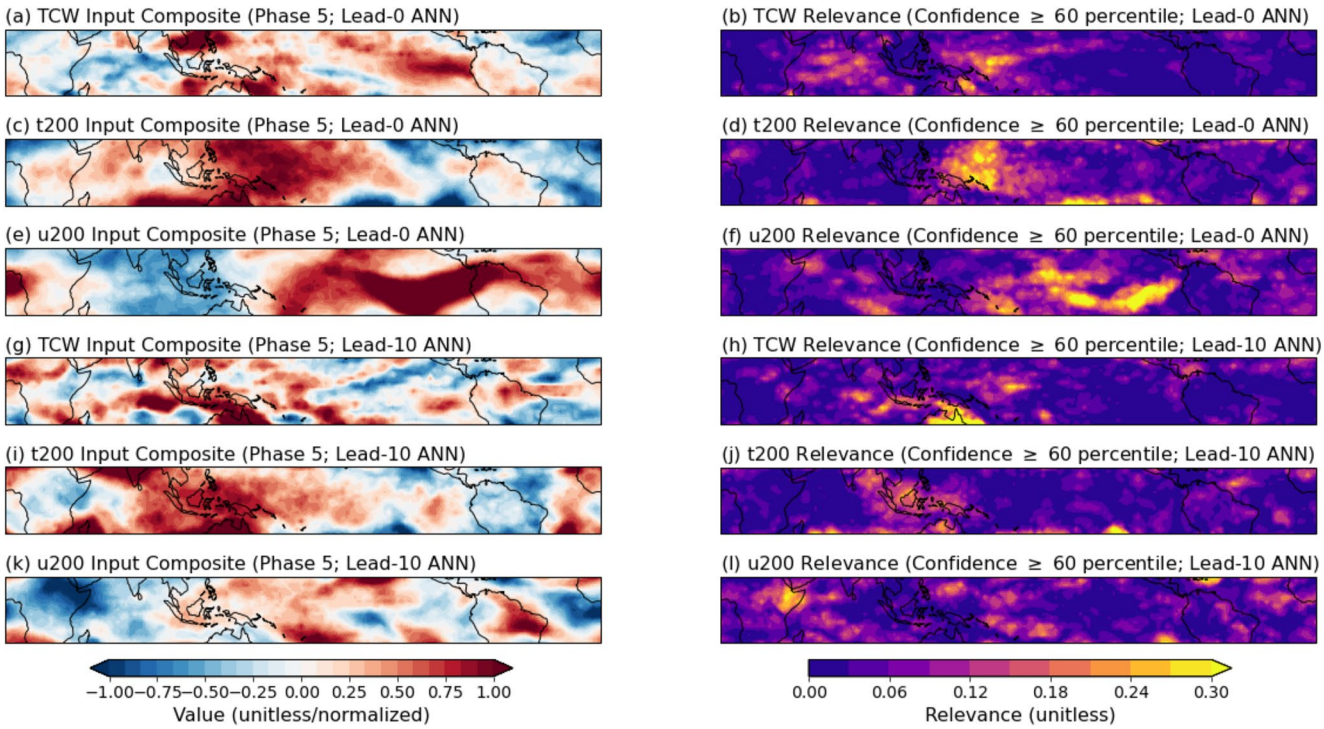

Figure 10. Layer-wise relevance propagation example. As in Figure 9, but for the artificial neural network inputting a different set of variables: total column water vapor, 200 hPa temperature (t200), and 200 hPa zonal wind.

winds near the equatorial Maritime Continent, and upper-level easterly anomalies in the western Indian Ocean (Figures 9i–9l).

Combining both experimentation across model inputs and LRP allows examination of sources of predictability across different variables. For example, while the 3-variable model using total column water vapor, and 200 hPa wind and temperature (gray bar in Figure 8) underperforms the OLR and zonal winds models at lead 0, at lead 10 their performance is comparable; Figure 10 shows the LRP maps from that model. At short leads, total column water vapor relevance matches regions of OLR relevance closely (compare Figures 9b and 10b), and the 200 hPa winds also focus on very similar regions. Upper-level temperatures are most relevant around the western Pacific slightly to the east of enhanced convection, where they show warm anomalies consistent with convective heating in the upper troposphere. In contrast, at 10 day leads the column water vapor shows a clearer difference in relevance compared to the OLR: water vapor signals south of the equator and Maritime Continent, as well as the signals around northern Australia show maxima in relevance. The focus in particular on southern hemisphere moisture signals may be due to the tendency of the winter-time MJO to detour south of the Maritime Continent (D. Kim et al., 2017). Upper-level temperature signals at lead 10 show highest relevance over the Maritime Continent, and focus mainly on near-equatorial warm anomalies in that region. It is noteworthy that while the composite (Figure 10i) shows equally strong temperature signals on the equator and in the subtropics to the west, the LRP map (Figure 10j) indicates the model focuses on the strong equatorial signals.

LRP thus provides information about how the ANN identifies the MJO and what signals across variables are most associated with future MJO behavior. The unique information LRP outputs may be useful to continue to explore sources of MJO prediction skill in simple ANNs, for example, under different large-scale states or for case studies.

5. Discussion and Conclusions

Motivated by the ability of ML methods to skillfully predict other climate and weather phenomena, as well as a lack of recent progress in statistical MJO modeling, here, we demonstrate how simple ML frameworks can be used to predict the MJO. We establish two straightforward neural network architectures (a regression and

classification approach) that use shallow ANNs to predict an MJO index. The regression ANN shows prediction skill out to ∼18 days in winter and ∼11 days in summer, which is high skill for a statistical approach. The classification ANN shows probabilistic skill better than climatology out to similar leads of ∼16 days in winter. Both ANN architectures perform better than traditional statistical models and set benchmarks for continued ML modeling of the MJO. ANN prediction skill is not comparable to dynamical models, though continued work may improve ML prediction skill of the MJO, perhaps via other ML modeling frameworks, more advanced input processing, or leveraging larger data sets from climate model simulations. We also emphasize in this work that simple ANNs are efficiently able to reproduce aspects of MJO predictability found in more complex, computationally-expensive dynamical models, such as sensitivity to MJO initial amplitude and phase of the stratospheric QBO. This makes these models affordable tools to continue to study the MJO and MJO predictability. XAI tools can also help illuminate sources and regions of ANN model skill.

This work illustrates how simple ANNs can be used not only for prediction, but also as tools for hypothesis testing and experimentation that might drive new discoveries or scientific insights. While our focus here is on the MJO, the framework we establish is widely applicable to a range of different climate phenomena, especially oscillations that can be represented as simple indices. The performance, affordability, accessibility, and explainability of simple ANNs thus recommends their continued adoption by the climate community.

Data Availability Statement

All data sets used in this study are publicly available. The RMM index is available at http://www.bom.gov.au/climate/mjo/graphics/rmm.74toRealtime.txt. For reanalysis and observed data, NOAA Interpolated OLR (Liebmann & Smith, 1996) is available at https://psl.noaa.gov/data/gridded/data.interp_OLR.html; NOAA OI SST V2 High Resolution (Reynolds et al., 2007) is available at https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.highres.html; ERA-5 reanalysis (Hersbach et al., 2020) is available at https://cds.climate.copernicus.eu/#!/search?text=ERA5&type=dataset; and ERA-20C data (Poli et al., 2016) is available at https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-20c.

Acknowledgments
Z. K. Martin acknowledges support from the National Science Foundation under Award No. 2020305. E. A. Barnes and E. D. Maloney are supported, in part, by the NOAA Climate Test Bed Grant NA18OAR4310296 and NOAA WPO Grant NA19OAR4590151. E. D. Maloney also acknowledges support from the NSF Climate and Large-Scale grant AGS-1841754, and NOAA CVP Grant NA18OAR4310299.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Klaus-Robert Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Baldwin, M. P., Grey, L. J., Dunkerton, T. J., Hamilton, K., Hayne, P. H., Randel, W. J., et al. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, *39*(2), 179–229. https://doi.org/10.1029/1999rg000073

Barnes, E. A., Mayer, K., Toms, B., Martin, Z., & Gordon, E. (2020). Identifying opportunities for skillful weather prediction with interpretable neural networks. *arXiv [physics.ao-Ph]*. arXiv. Retrieved from http://arxiv.org/abs/2012.07830

Dasgupta, P., Metya, A., Naidu, C. V., Singh, M., & Roxy, M. K. (2020). Exploring the long-term changes in the Madden Julian oscillation using machine learning. *Scientific Reports*, *10*(1), 18567. https://doi.org/10.1038/s41598-020-75508-5

Ebdon, R. A. (1960). Notes on the wind flow at 50 Mb in tropical and sub-tropical regions in January 1957 and January 1958. *Quarterly Journal of the Royal Meteorological Society*, *86*(370), 540–542. https://doi.org/10.1002/qj.49708637011

Gagne, D. J., McGovern, A., & Xue, M. (2014). Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Weather and Forecasting*, *29*(4), 1024–1043. https://doi.org/10.1175/waf-d-13-00108.1

Hagos, S., Ruby Leung, L., Zhang, C., & Balaguru, K. (2021). An observationally trained Markov Model for MJO propagation. *Geophysical Research Letters*, *48*, e2021GL095663. https://doi.org/10.1029/2021GL095663

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572. https://doi.org/10.1038/s41586-019-1559-7

Hendon, H. H., & Salby, M. L. (1994). The life cycle of the Madden–Julian oscillation. *Journal of the Atmospheric Sciences*, *51*(15), 2225–2237. https://doi.org/10.1175/1520-0469(1994)051<2225:tlcotm>2.0.co;2

Hendon, H. H., Zhang, C., & Glick, J. D. (1999). Interannual variation of the Madden–Julian oscillation during Austral summer. *Journal of Climate*, *12*(8), 2538–2550. https://doi.org/10.1175/1520-0442(1999)012<2538:ivotmj>2.0.co;2

Hersbach, H., Bell, B., Paul, B., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Jiang, X., Waliser, D. E., Wheeler, M. C., Jones, C., Lee, M.-I., & Schubert, S. D. (2008). Assessing the skill of an all-season statistical forecast model for the Madden–Julian oscillation. *Monthly Weather Review*, *136*(6), 1940–1956.

Kang, I.-S., & Kim, H.-M. (2010). Assessment of MJO predictability for boreal winter with various statistical and dynamical models. *Journal of Climate*, *23*(9), 2368–2378. https://doi.org/10.1175/2010jcli3288.1

Kikuchi, K., Wang, B., & Kajikawa, Y. (2012). Bimodal representation of the tropical intraseasonal oscillation. *Climate Dynamics*, *38*(9–10), 1989–2000. https://doi.org/10.1007/s00382-011-1159-1

Kiladis, G. N., Dias, J., Straub, K. H., Wheeler, M. C., Tulich, S. N., Kikuchi, K., et al. (2014). A comparison of OLR and circulation-based indices for tracking the MJO. *Monthly Weather Review*, *142*(5), 1697–1715. https://doi.org/10.1175/mwr-d-13-00301.1

Kim, D., Kim, H., & Lee, M.-I. (2017). Why does the MJO detour the Maritime continent during Austral summer? *Geophysical Research Letters*, *44*(5), 2579–2587. https://doi.org/10.1002/2017gl072643

Kim, H., Ham, Y. G., Joo, Y. S., & Son, S. W. (2021). Deep learning for bias correction of MJO prediction. *Nature Communications*, *12*(1), 1–7. https://doi.org/10.1038/s41467-021-23406-3

Kim, H., Richter, J. H., & Martin, Z. (2019). Insignificant QBO-MJO prediction skill relationship in the SubX and S2S subseasonal reforecasts. *Journal of Geophysical Research*. https://doi.org/10.1029/2019jd031416

Kim, H., Vitart, F., & Waliser, D. E. (2018). Prediction of the Madden–Julian oscillation: A review. *Journal of Climate*, *31*(23), 9425–9443. https://doi.org/10.1175/jcli-d-18-0210.1

Lagerquist, R., McGovern, A., & Smith, T. (2017). Machine learning for real-time prediction of damaging straight-line convective wind. *Weather and Forecasting*, *32*(6), 2175–2193. https://doi.org/10.1175/waf-d-17-0038.1

Liebmann, B., & Smith, C. A. (1996). Description of a complete (interpolated) outgoing longwave radiation dataset. *Bulletin of the American Meteorological Society*, *77*(6), 1275–1277.

Lim, Y., Son, S.-W., Marshall, A. G., Hendon, H. H., & Seo, K.-H. (2019). Influence of the QBO on MJO prediction skill in the subseasonal-to-seasonal prediction models. *Climate Dynamics*, *53*(March), 1681–1695. https://doi.org/10.1007/s00382-019-04719-y

Ling, J., Li, C., Li, T., Jia, X., Khouider, B., Maloney, E., et al. (2017). Challenges and opportunities in MJO studies. *Bulletin of the American Meteorological Society*, *98*(2), ES53–ES56. https://doi.org/10.1175/bams-d-16-0283.1

Love, B. S., & Matthews, A. J. (2009). Real-time localised forecasting of the Madden-Julian oscillation using neural network models. *Quarterly Journal of the Royal Meteorological Society*, *135*(643), 1471–1483. https://doi.org/10.1002/qj.463

Madakumbura, G. D., Thackeray, C. W., Norris, J., Goldenson, N., & Hall, A. (2021). Anthropogenic influence on extreme precipitation over global land areas seen in multiple observational datasets. *Research Square*, *12*(April). https://doi.org/10.21203/rs.3.rs-227967/v2

Madden, R. A., & Julian, P. R. (1971). Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *Journal of the Atmospheric Sciences*, *28*(5), 702–708. https://doi.org/10.1175/1520-0469(1971)028<0702:doadoi>2.0.co;2

Madden, R. A., & Julian, P. R. (1972). Description of global-scale circulation cells in the tropics with a 40–50 Day period. *Journal of the Atmospheric Sciences*, *29*(6), 1109–1123. https://doi.org/10.1175/1520-0469(1972)029<1109:dogscc>2.0.co;2

Maharaj, E. A., & Wheeler, M. C. (2005). Forecasting an index of the Madden-oscillation. *International Journal of Climatology*, *25*(12), 1611–1618. https://doi.org/10.1002/joc.1206

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *arXiv [physics.geo-Ph]*. arXiv. Retrieved from http://arxiv.org/abs/2103.10005

Marshall, A. G., Hendon, H. H., & Hudson, D. (2016). Visualizing and verifying probabilistic forecasts of the Madden-Julian oscillation. *Geophysical Research Letters*, *43*(23), 12278–12286. https://doi.org/10.1002/2016gl071423

Marshall, A. G., Hendon, H. H., Son, S. W., & Lim, Y. (2017). Impact of the quasi-biennial oscillation on predictability of the Madden–Julian oscillation. *Climate Dynamics*, *49*(4), 1365–1377. https://doi.org/10.1007/s00382-016-3392-0

Martin, Z., Son, S.-W., Butler, A., Hendon, H., Kim, H., Adam, S., et al. (2021). The influence of the quasi-biennial oscillation on the Madden–Julian oscillation. *Nature Reviews Earth & Environment*, *2*(June), 477–489.

Mayer, K. J., & Barnes, E. A. (2021). Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophysical Research Letters*, *48*(May). https://doi.org/10.1029/2020gl092092

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., et al. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073–2090. https://doi.org/10.1175/bams-d-16-0123.1

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1

Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Kim, C., Doblas-Reyes, F., et al. (2021). Initialized Earth system prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment*, *2*(April), 340–357. https://doi.org/10.1038/s43017-021-00155-x

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: An overview. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 193–209). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_10

Newman, M., Sardeshmukh, P. D., & Penland, C. (2009). How important is air–sea coupling in ENSO and MJO evolution? *Journal of Climate*, *22*(11), 2958–2977. https://doi.org/10.1175/2008jcli2659.1

Poli, P., Hersbach, H., Dee, D. P., Paul, B., Simmons, A. J., Vitart, F., et al. (2016). ERA-20C: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, *29*(11), 4083–4097.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11). https://doi.org/10.1029/2020ms002203

Reed, R. J., Campbell, W. J., Rasmussen, L. A., & Rogers, D. G. (1961). Evidence of a downward-propagating, annual wind reversal in the equatorial stratosphere. *Journal of Geophysical Research*, *66*(3), 813–818. https://doi.org/10.1029/jz066i003p00813

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., & Schlax, M. G. (2007). Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, *20*(22), 5473–5496. https://doi.org/10.1175/2007jcli1824.1

Roundy, P. E., Schreck, C. J., & Janiga, M. A. (2009). Contributions of convectively coupled equatorial Rossby waves and Kelvin waves to the Real-time Multivariate MJO indices. *Monthly Weather Review*, *137*(1), 469–478. https://doi.org/10.1175/2008mwr2595.1

Samek, W., Montavon, G., Binder, A., Lapuschkin, S., & Müller, K.-R. (2016). Interpreting the predictions of complex ML models by layer-wise relevance propagation. *arXiv [stat.ML]*. arXiv. Retrieved from http://arxiv.org/abs/1611.08191

Seo, K.-H., Wang, W., Gottschalck, J., Zhang, Q., Schemm, J.-K. E., Higgins, W. R., & Kumar, A. (2009). Evaluation of MJO forecast skill from several statistical and dynamical forecast models. *Journal of Climate*, *22*(9), 2372–2388. https://doi.org/10.1175/2008jcli2421.1

Son, S. W., Lim, Y., Yoo, C., Hendon, H. H., & Kim, J. (2017). Stratospheric control of the Madden-Julian oscillation. *Journal of Climate*, *30*(6), 1909–1922. https://doi.org/10.1175/jcli-d-16-0620.1

Straub, K. H. (2013). MJO initiation in the Real-time Multivariate MJO index. *Journal of Climate*, *26*(4), 1130–1151. https://doi.org/10.1175/jcli-d-12-00074.1

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), 1. https://doi.org/10.1029/2019ms002002

Toms, B. A., Kashinath, K., & Yang, D. (2019). Testing the reliability of interpretable neural networks in geoscience using the Madden-Julian oscillation. *arXiv [physics.ao-Ph]*. arXiv. Retrieved from http://arxiv.org/abs/1902.04621

Ventrice, M. J., Wheeler, M. C., Hendon, H. H., Schreck, C. J., Thorncroft, C. D., & Kiladis, G. N. (2013). A modified multivariate Madden–Julian oscillation index using velocity potential. *Monthly Weather Review*, *141*(12), 4197–4210. https://doi.org/10.1175/mwr-d-12-00327.1

Vitart, F. (2014). Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1889–1899. https://doi.org/10.1002/qj.2256

Vitart, F. (2017). Madden-Julian oscillation prediction and teleconnections in the S2S database. *Quarterly Journal of the Royal Meteorological Society*, *143*(706), 2210–2220. https://doi.org/10.1002/qj.3079

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C. M. D., et al. (2017). The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/bams-d-16-0017.1

Waliser, D. (2012). Predictability and forecasting. In W. K.-M. Lau & D. E. Waliser (Eds.), *Intraseasonal variability in the atmosphere-ocean climate system* (pp. 433–476). Springer Berlin Heidelberg.

Wang, S., Tippett, M. K., Sobel, A. H., Martin, Z. K., & Vitart, F. (2019). Impact of the QBO on prediction and predictability of the MJO convection. *Journal of Geophysical Research: Atmospheres*, *124*(22), 11766–11782. https://doi.org/10.1029/2019jd030575

Weyn, J. A., Durran, D. R., & Rich, C. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2680–2693. https://doi.org/10.1029/2019ms001705

Wheeler, M. C., & Hendon, H. H. (2004). An all-season Real-time Multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, *132*(8), 1917–1932. https://doi.org/10.1175/1520-0493(2004)132<1917:aarmmi>2.0.co;2

Yoo, C., & Son, S. W. (2016). Modulation of the Boreal wintertime Madden-Julian oscillation by the stratospheric quasi-biennial oscillation. *Geophysical Research Letters*, *43*(3), 1392–1398. https://doi.org/10.1002/2016gl067762

Zhang, C. (2005). Madden-Julian oscillation. *Reviews of Geophysics*, *43*(2). https://doi.org/10.1029/2004rg000158

Zhang, C., & Dong, M. (2004). Seasonality in the Madden–Julian oscillation. *Journal of Climate*, *17*(16), 3169–3180. https://doi.org/10.1175/1520-0442(2004)017<3169:sitmo>2.0.co;2