Using simple, explainable neural networks to predict the Madden-Julian oscillation

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

[1]Colorado State University

November 30, 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 ˜17 days in October-March and ˜10 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 is 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.

Using simple, explainable neural networks to predict the Madden-Julian oscillation

Zane K. Martin[1]

Elizabeth A. Barnes[1]

Eric Maloney[1]

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

Corresponding author: Zane Martin, zkmartin@colostate.edu

Draft to be submitted to

Journal of Advances in Modeling Earth Systems

2021

Key points

1. Simple machine learning models are an efficient, flexible tool to predict and study the Madden-Julian oscillation (MJO)
2. Shallow neural networks skillfully predict an MJO index out to ~17 days in winter and ~10 days in summer, outperforming linear models
3. Varying ANN input and using explainable artificial intelligence methods offer insights into the MJO and key regions for prediction skill

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 ~17 days in October-March and ~10 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 is 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 important 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 and Julian 1971, 1972; Zhang 2005), is a key source of subseasonal-to-seasonal (S2S) predictability (Vitart et al. 2017; Kim et al. 2018). Skillful MJO prediction has important societal implications (Meehl et al. 2021; Vitart et al. 2017; Kim et al. 2018), and extensive research has explored using both statistical models and initialized dynamical forecast models to predict the MJO (e.g. Waliser 2012; Vitart et al. 2017; Kim et al. 2018; Meehl et al. 2021; and references therein). Before the late 2000s, statistical models showed superior MJO prediction skill (~2 weeks; Waliser 2012; Kang and Kim 2010) compared to dynamical models, but S2S forecast models have continually improved and several now skillfully predict the MJO beyond one month (Vitart 2014; Vitart 2017; Kim et al. 2018).

In contrast, statistical MJO modeling has stagnated in recent years. Compared to dynamical models, statistical MJO models have the advantage of being computationally and are often much simpler to formulate and in some cases understand. To date, the most common statistical MJO models use linear methods (e.g. Maharaj and Wheeler 2005; Jiang et al. 2008; Seo et al. 2009; Kang and Kim 2010; Marshall et al. 2016; Kim et al. 2018), 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; Lagerquist et al.2017; McGovern et al. 2017; Weyn et al. 2019; Rasp et al. 2020; Ham et al. 2019; Mayer and Barnes 2021), 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 and Barnes 2021).

Studies using machine learning 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 (Kim et al. 2021), but only one study to our knowledge has examined MJO prediction solely using ML (Love and 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. This work further helps 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, the exploration of both deterministic and probabilistic ML model architectures, and the application of tools from the field of explainable AI (XAI; McGovern et al. 2019; Toms et al. 2020; Mamalakis et al. 2021).

This paper thus addresses three aspects of using machine learning to study the MJO: (1) developing ML frameworks, (2) analyzing ML model performance, and (3) 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 machine learning 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-20°S. The predictand is the observed Real-time Multivariate MJO index ("RMM"; Wheeler and 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 ($\sqrt{RMM1^2 + RMM2^2}$) measures MJO strength. RMM has known limitations (Roundy et al. 2009; Straub 2013) and other MJO indices exist (e.g. Kikuchi et al. 2012; Ventrice et al. 2013; Kiladis et al. 2014), 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 dataset (Liebmann and Smith 1996), sea-surface temperature (SST) is from the NOAA OI SST V2 High Resolution dataset (Reynolds et al. 2007), and all other variables are from ERA-5 reanalysis (Hersbach et al. 2020). Additional data from the ERA-20C dataset (Poli et al. 2016) is used in the Supplemental Material, as described therein. We use daily mean data from January 1, 1979 (1982 for SST) to December 31, 2019 that are interpolated onto a common 2.5° x 2.5° grid.

ANN input data are pre-processed in a similar way to that of the RMM input variables (Wheeler and 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 by the ANNs (sensitivity tests exploring latitudinal averaging are discussed in Supplemental Material). We also 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).

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 June 1, 1979 to December 31, 2009; the validation data is from January 1, 2010 to December 31, 2015; and the testing is from January 1, 2016 to November 30, 2019. Results from the validation and testing period are shown together in the manuscript.

In Section 4, where sensitivity of the model to the phase of the stratospheric quasi-biennial oscillation (QBO; Ebdon 1960; Reed et al. 1961; Baldwin et al. 2001) is shown, 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 phases, and months greater than half a standard deviation from the mean are QBO westerly phases (e.g. Yoo and Son 2016; Son et al. 2017).

3. Machine Learning and Linear Statistical MJO Models

Here we first discuss the two types of artificial neural networks (ANNs) and an ANN explainability technique used in this study. We then describe three conventional statistical MJO models used in prior studies (Maharaj and Wheeler 2005; Jiang et al. 2008; Kang and Kim 2010; 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 the 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 and Kim 2010; Waliser 2012). Using the ANNs in this 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 (**Fig. 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). An example regression ANN output is 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, 3): either active (RMM amplitude ≥ 1) in one of the eight canonical RMM phases (Wheeler and Hendon 2004) or weak (“phase 0”; RMM amplitude < 1). The predicted class is the highest probability. 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. 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 period. Values of key hyperparameters used in both architectures and additional model details are listed in Table 1. Sensitivity tests varying ANN parameters and input 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 Supplemental Material.

ANN performance is slightly improved if the models are trained separately on different seasons (Figure 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 and Salby 1994; Hendon et al. 1999; Zhang and Dong 2004), and we found splitting the data into two six-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 seasons and lead times, 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 ensures convergence of our results and quantifies sensitivity to ANN initialization.

3.1.2. Layer-wise Relevance Propagation (LRP)

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 (Bach et al. 2015; Samek et al. 2016; Montavon et al. 2019). LRP has been used in Earth science as a tool for understanding the decision-making process of ANNs (Toms et al. 2019; Toms et al. 2020; Barnes et al. 2020; Mayer and Barnes 2021; Mamalakis et al. 2021; Madakumbura et al. 2021), 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), 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; Samek et al. 2016; Montavon et al. 2019; Mamalakis et al. 2021). Based on results in Mamalakis et al. (2021) assessing various implementations of LRP in a synthetic dataset, 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:

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

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

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

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 and 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 one day at a time out to a lead time of 20 days.

Our third simple model, the MLR model (Jiang et al. 2008; Kang and 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:

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

$L_{MLR,\tau}$ 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 one 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. Vitart et al. 2017; Kim et al. 2018), 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:

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

Here N is the number of forecast, 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:

$$RPSS = 1 - \frac{RPS_{model}}{RPS_{ref}}$$

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 (Fig. 1) for forecasts initialized daily over the validation and testing period.

Overall, the winter and summer regression ANNs show prediction skill, respectively, of ~17 days and ~11 days (Fig. 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 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. This indicates that simple ANNs are at 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 ANN. The improved performance of the ANN relative to the MLR and VAR model further demonstrates that the ANNs learn not only to identify the MJO and propagate it east, but also capture more nuanced MJO behavior. 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.

The regression ANN skill shows relatively small sensitivity to initial MJO phase (Fig. 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 (Fig. 5c). This is consistent with findings in other statistical and dynamical models (Kim et al. 2018). ANNs also capture more mysterious aspects of MJO predictability, such as the sensitivity to the phase of the stratospheric quasi-biennial oscillation (Marshall et al. 2017; Martin et al. 2021). Studies in both dynamical and statistical models have found improved MJO prediction skill in QBO easterly months compared to QBO westerly months during December-February (DJF; Marshall et al. 2017; Lim et al. 2019; Kim et al. 2019; Wang et al. 2019). Defining the QBO using the U50 index, the wintertime regression ANN skill during QBO easterly DJF periods is nearly 20 days, whereas during QBO westerly DJF skill is only 15 days (**Fig. 5c**). This modulation is quantitatively consistent with findings in dynamical models (Lim et al. 2019; Kim et al. 2019), though we note here the number of QBO cycles is limited since only winters from 2010-2019 are considered.

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 past a few days, especially in phases 4-7 (Fig. 5b). Amplitude biases are 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 on 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 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% **(Fig. 6)**, an accuracy comparable to (Toms et al. 2019), despite differences in our input variables, data pre-processing, MJO index, and ANN complexity. Most 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 (e.g. Figure 3 at lead 10 and 15).

While classification ANN skill is substantially better at predicting active MJO events, it struggles to predict weak MJO days, with an accuracy at short leads of only ~40%, which falls to near random chance after ~10 days (Figure 6). This is in part due to the strategy used to train 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 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 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. This aspect of the classification ANN is distinct from the deterministic output provided by linear models or even 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 15 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 7b). 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 (Fig. 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 and 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. 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 (**Fig. S2**). 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 two weeks in winter, and the ANNs outperform existing linear statistical models. A range of sensitivity tests (**Supple. Text and Figs. S3, S4, S5**), including increasing the amount of training data using 20th-century reanalysis, 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 ANNs we explored are not yet competitive with most S2S dynamical forecast models (e.g. Vitart 2017; Kim et al. 2018). State-of-the-art dynamic model skill predicting the MJO generally falls between 25-35 days when assessed via the BCC (Vitart 2017; Kim et al. 2018), and probabilistic MJO forecasts formed by running ensembles of dynamical models showed 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 explore the 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 explainable AI 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 (Fig. 8). For example, across 1-variable ANNs (Fig 8a) 850 hPa meridional wind and sea-surface temperature (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 and captured by RMM.

The most accurate models at short leads are those that input 850 hPa and/or 200 hPa zonal winds (Fig. 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 850hPa and 200 hPa zonal wind model has the clear highest accuracy among 2-variable models (Fig. 8b), but at lead 5 and 10 its accuracy overlaps with other configurations. Best performing models at longer leads are 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. Further, RMM input variables are not always clearly superior at leads 5 and 10: 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 (Fig. 8c).

Finally, while more input variables tend to improve model performance (Fig. 8), tests showed no substantial improvement using 4 or more inputs (Fig. S5), 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.

A second 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 (Fig. 9a,b), whereas wind fields focus on low-level westerly anomalies around the Maritime Continent (Fig. 9c,d) and upper level signals in the central and east Pacific (Fig. 9e,f), 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 (Fig. 9g,h). The lead-10 model also focuses on a small dipole region of strong low-level winds near the equatorial Maritime Continent, and upper-level easterly anomalies in the western Indian Ocean (Figs. 9i-l).

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 (grey 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 Figs. 9b and 10b), and the 200 hPa winds also focus on similar very 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 (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 (**Fig. 10i**) shows equally strong temperature signals on the equator and in the subtropics to the west, the LRP map **(Fig. 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 of particular events.

5. Discussion & Conclusions

Motivated by a lack of recent progress in statistical MJO modeling and the ability of machine learning methods to skillfully predict other climate and weather phenomena, here we demonstrate how simple machine learning frameworks can be used to predict the MJO. We established 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 ~17 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 15 days in winter. Both ANN architectures perform better than traditional statistical models and set benchmarks for continued ML modeling of the MJO. Note however that ANN prediction skill is not yet comparable to dynamical models, though continued work may improve prediction skill perhaps via other ML modeling frameworks, more advanced input processing, or leveraging larger datasets from climate model simulations. We further emphasize 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, making them affordable tools to continue to study the MJO and MJO predictability. Explainable AI 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.

Acknowledgments

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

Data availability

All datasets 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 and 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.

References

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.” *PloS One* 10 (7): e0130140.

Baldwin, M. P., L. J. Grey, T. J. Dunkerton, K. Hamilton, Peter Haynes, William J. Randel, James R. Holton, et al. 2001. “The Quasi-Biennial Oscillation.” *Reviews of Geophysics* 39 (2): 179–229.

Barnes, Elizabeth A., Kirsten Mayer, Benjamin Toms, Zane Martin, and Emily Gordon. 2020. “Identifying Opportunities for Skillful Weather Prediction with Interpretable Neural Networks.” *arXiv [physics.ao-Ph]*. arXiv. http://arxiv.org/abs/2012.07830.

Dasgupta, Panini, Abirlal Metya, C. V. Naidu, Manmeet Singh, and M. K. Roxy. 2020. “Exploring the Long-Term Changes in the Madden Julian Oscillation Using Machine Learning.” *Scientific Reports* 10 (1): 18567.

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): 540–42.

Gagne, David John, Amy McGovern, and Ming Xue. 2014. “Machine Learning Enhancement of Storm-Scale Ensemble Probabilistic Quantitative Precipitation Forecasts.” *Weather and Forecasting* 29 (4): 1024–43.

Ham, Yoo-Geun, Jeong-Hwan Kim, and Jing-Jia Luo. 2019. “Deep Learning for Multi-Year ENSO Forecasts.” *Nature* 573 (7775): 568–72.

Hendon, Harry H., and Murry L. Salby. 1994. “The Life Cycle of the Madden–Julian Oscillation.” *Journal of the Atmospheric Sciences* 51 (15): 2225–37.

Hendon, Harry H., Chidong Zhang, and John D. Glick. 1999. “Interannual Variation of the Madden–Julian Oscillation during Austral Summer.” *Journal of Climate* 12 (8): 2538–50.

Hersbach, Hans, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, et al. 2020. “The ERA5 Global Reanalysis.” *Quarterly Journal of the Royal Meteorological Society* 146 (730): 1999–2049.

Jiang, Xianan, Duane E. Waliser, Matthew C. Wheeler, Charles Jones, Myong-In Lee, and Siegfried D. Schubert. 2008. “Assessing the Skill of an All-Season Statistical Forecast Model for the Madden–Julian Oscillation.” *Monthly Weather Review* 136 (6): 1940–56.

Kang, In-Sik, and Hye-Mi Kim. 2010. “Assessment of MJO Predictability for Boreal Winter with Various Statistical and Dynamical Models.” *Journal of Climate* 23 (9): 2368–78.

Kikuchi, Kazuyoshi, Bin Wang, and Yoshiyuki Kajikawa. 2012. “Bimodal Representation of the Tropical Intraseasonal Oscillation.” *Climate Dynamics* 38 (9-10): 1989–2000.

Kiladis, George N., Juliana Dias, Katherine H. Straub, Matthew C. Wheeler, Stefan N. Tulich, Kazuyoshi Kikuchi, Klaus M. Weickmann, and Michael J. Ventrice. 2014. “A Comparison of OLR and Circulation-Based Indices for Tracking the MJO.” *Monthly Weather Review* 142 (5): 1697–1715.

Kim, Daehyun, Hyerim Kim, and Myong-In Lee. 2017. “Why Does the MJO Detour the Maritime Continent during Austral Summer?” *Geophysical Research Letters* 44 (5): 2579–87.

Kim, Hyemi, Y. G. Ham, Y. S. Joo, and S. W. Son. 2021. “Deep Learning for Bias Correction of MJO Prediction.” *Nature Communications* 12 (1): 1–7.

Kim, Hyemi, J. H. Richter, and Z. Martin. 2019. “Insignificant QBO-MJO Prediction Skill Relationship in the SubX and S2S Subseasonal Reforecasts.” *Journal of Geophysical*

Research. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD031416.

Kim, Hyemi, Frédéric Vitart, and Duane E. Waliser. 2018. “Prediction of the Madden–Julian Oscillation: A Review.” *Journal of Climate* 31 (23): 9425–43.

Lagerquist, Ryan, Amy McGovern, and Travis Smith. 2017. “Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind.” *Weather and Forecasting* 32 (6): 2175–93.

Liebmann, Brant, and Catherine A. Smith. 1996. “Description of a Complete (Interpolated) Outgoing Longwave Radiation Dataset.” *Bulletin of the American Meteorological Society* 77 (6): 1275–77.

Lim, Yuna, Seok-Woo Son, Andrew G. Marshall, Harry H. Hendon, and Kyong-Hwan Seo. 2019. “Influence of the QBO on MJO Prediction Skill in the Subseasonal-to-Seasonal Prediction Models.” *Climate Dynamics*, March, 1–15.

Love, Barnaby S., and Adrian J. Matthews. 2009. “Real-Time Localised Forecasting of the Madden-Julian Oscillation Using Neural Network Models.” *Quarterly Journal of the Royal Meteorological Society* 135 (643): 1471–83.

Madakumbura, Gavin D., Chad W. Thackeray, Jesse Norris, Naomi Goldenson, and Alex Hall. 2021. “Anthropogenic Influence on Extreme Precipitation over Global Land Areas Seen in Multiple Observational Datasets.” *Research Square*, April. https://doi.org/10.21203/rs.3.rs-227967/v2.

Madden, Roland A., and Paul R. Julian. 1971. “Detection of a 40–50 Day Oscillation in the Zonal Wind in the Tropical Pacific.” *Journal of the Atmospheric Sciences* 28 (5): 702–8.

Madden, Roland A., and Paul R. Julian. 1972. “Description of Global-Scale Circulation Cells in the Tropics with a 40–50 Day Period.” *Journal of the Atmospheric Sciences* 29 (6): 1109–

23.

Maharaj, Elizabeth A., and Matthew C. Wheeler. 2005. “Forecasting an Index of the Madden-Oscillation.” *International Journal of Climatology* 25 (12): 1611–18.

Mamalakis, Antonios, Imme Ebert-Uphoff, and Elizabeth A. Barnes. 2021. “Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset.” *arXiv [physics.geo-Ph]*. arXiv. http://arxiv.org/abs/2103.10005.

Marshall, Andrew G., Harry H. Hendon, and Debra Hudson. 2016. “Visualizing and Verifying Probabilistic Forecasts of the Madden-Julian Oscillation.” *Geophysical Research Letters* 43 (23): 12,278–12,286.

Marshall, Andrew G., Harry H. Hendon, Seok Woo Son, and Yuna Lim. 2017. “Impact of the Quasi-Biennial Oscillation on Predictability of the Madden–Julian Oscillation.” *Climate Dynamics* 49 (4): 1365–77.

Martin, Zane, Seok-Woo Son, Amy Butler, Harry Hendon, Hyemi Kim, Adam Sobel, Shigeo Yoden, and Chidong Zhang. 2021. “The Influence of the Quasi-Biennial Oscillation on the Madden–Julian Oscillation.” *Nature Reviews Earth & Environment*, June, 1–13.

Mayer, Kirsten J., and Elizabeth A. Barnes. 2021. “Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network.” *Geophysical Research Letters*, May. https://doi.org/10.1029/2020gl092092.

McGovern, Amy, Kimberly L. Elmore, David John Gagne, Sue Ellen Haupt, Christopher D. Karstens, Ryan Lagerquist, Travis Smith, and John K. Williams. 2017. “Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather.” *Bulletin of the American Meteorological Society* 98 (10): 2073–90.

McGovern, Amy, Ryan Lagerquist, David John Gagne, G. Eli Jergensen, Kimberly L. Elmore, Cameron R. Homeyer, and Travis Smith. 2019. “Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning.” *Bulletin of the American Meteorological Society* 100 (11): 2175–99.

Meehl, Gerald A., Jadwiga H. Richter, Haiyan Teng, Antonietta Capotondi, Kim Cobb, Francisco Doblas-Reyes, Markus G. Donat, et al. 2021. “Initialized Earth System Prediction from Subseasonal to Decadal Timescales.” *Nature Reviews Earth & Environment*, April, 1–18.

Montavon, Grégoire, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. “Layer-Wise Relevance Propagation: An Overview.” In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, edited by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, 193–209. Cham: Springer International Publishing.

Newman, Matthew, Prashant D. Sardeshmukh, and Cécile Penland. 2009. “How Important Is Air–Sea Coupling in ENSO and MJO Evolution?” *Journal of Climate* 22 (11): 2958–77.

Poli, Paul, Hans Hersbach, Dick P. Dee, Paul Berrisford, Adrian J. Simmons, Frédéric Vitart, Patrick Laloyaux, et al. 2016. “ERA-20C: An Atmospheric Reanalysis of the Twentieth Century.” *Journal of Climate* 29 (11): 4083–97.

Rasp, Stephan, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. 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, Richard J., William J. Campbell, Lowell A. Rasmussen, and Dale G. Rogers. 1961. “Evidence of a Downward-Propagating, Annual Wind Reversal in the Equatorial

Stratosphere.” *Journal of Geophysical Research* 66 (3): 813–18.

Reynolds, Richard W., Thomas M. Smith, Chunying Liu, Dudley B. Chelton, Kenneth S. Casey, and Michael G. Schlax. 2007. “Daily High-Resolution-Blended Analyses for Sea Surface Temperature.” *Journal of Climate* 20 (22): 5473–96.

Roundy, Paul E., Carl J. Schreck, and Matthew A. Janiga. 2009. “Contributions of Convectively Coupled Equatorial Rossby Waves and Kelvin Waves to the Real-Time Multivariate MJO Indices.” *Monthly Weather Review* 137 (1): 469–78.

Samek, Wojciech, Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. “Interpreting the Predictions of Complex ML Models by Layer-Wise Relevance Propagation.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1611.08191.

Seo, Kyong-Hwan, Wanqiu Wang, Jon Gottschalck, Qin Zhang, Jae-Kyung E. Schemm, Wayne R. Higgins, and Arun Kumar. 2009. “Evaluation of MJO Forecast Skill from Several Statistical and Dynamical Forecast Models.” *Journal of Climate* 22 (9): 2372–88.

Son, Seok Woo, Yuna Lim, Changhyun Yoo, Harry H. Hendon, and Joowan Kim. 2017. “Stratospheric Control of the Madden-Julian Oscillation.” *Journal of Climate* 30 (6): 1909–22.

Straub, Katherine H. 2013. “MJO Initiation in the Real-Time Multivariate MJO Index.” *Journal of Climate* 26 (4): 1130–51.

Toms, Benjamin A., Elizabeth A. Barnes, and Imme Ebert-Uphoff. 2020. “Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability.” *Journal of Advances in Modeling Earth Systems* 12 (9): 1.

Toms, Benjamin A., Karthik Kashinath, Prabhat, and Da Yang. 2019. “Testing the Reliability of Interpretable Neural Networks in Geoscience Using the Madden-Julian Oscillation.” *arXiv* *[physics.ao-Ph]*. arXiv. http://arxiv.org/abs/1902.04621.

Ventrice, Michael J., Matthew C. Wheeler, Harry H. Hendon, Carl J. Schreck, Chris D. Thorncroft, and George N. Kiladis. 2013. "A Modified Multivariate Madden–Julian Oscillation Index Using Velocity Potential." *Monthly Weather Review* 141 (12): 4197–4210.

Vitart, Frédéric, C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, et al. 2017. "The Subseasonal to Seasonal (S2S) Prediction Project Database." *Bulletin of the American Meteorological Society* 98 (1): 163–73.

Vitart, Frédéric. 2014. "Evolution of ECMWF Sub-Seasonal Forecast Skill Scores." *Quarterly Journal of the Royal Meteorological Society* 140 (683): 1889–99.

Vitart, Frédéric. 2017. "Madden-Julian Oscillation Prediction and Teleconnections in the S2S Database." *Quarterly Journal of the Royal Meteorological Society* 143 (706): 2210–20.

Waliser, Duane 2012. "Predictability and Forecasting." In *Intraseasonal Variability in the Atmosphere-Ocean Climate System*, edited by William K-M Lau and Duane E. Waliser, 433–76. Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, Shuguang, Michael K. Tippett, Adam H. Sobel, Zane K. Martin, and Frederic Vitart. 2019. "Impact of the QBO on Prediction and Predictability of the MJO Convection." *Journal of Geophysical Research: Atmospheres* 124 (22): 11766–82.

Weyn, Jonathan A., Dale R. Durran, and Rich Caruana. 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): 2680–93.

Wheeler, Matthew C., and Harry H. Hendon. 2004. "An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction." *Monthly Weather*

Review 132 (8): 1917–32.

Yoo, Changhyun, and Seok Woo Son. 2016. “Modulation of the Boreal Wintertime Madden-Julian Oscillation by the Stratospheric Quasi-Biennial Oscillation.” *Geophysical Research Letters* 43 (3): 1392–98.

Zhang, Chidong. 2005. “Madden-Julian Oscillation.” *Reviews of Geophysics* 43 (2). https://doi.org/10.1029/2004rg000158.

Zhang, Chidong, and Min Dong. 2004. “Seasonality in the Madden–Julian Oscillation.” *Journal of Climate* 17 (16): 3169–80.

Tables

ANN Model Details & Hyperparameters		
Name	***Regression ANN value***	***Classification ANN value***
Winter/summer training samples	5,560/5,612	3,990/3,726
Winter/summer validation & 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

Table 1. Regression and classification neural network model architecture details and key hyperparameters used in this study. Sensitivity tests to various aspects of these and other aspects of the ANN models are discussed in the Supplemental Material.

Figures

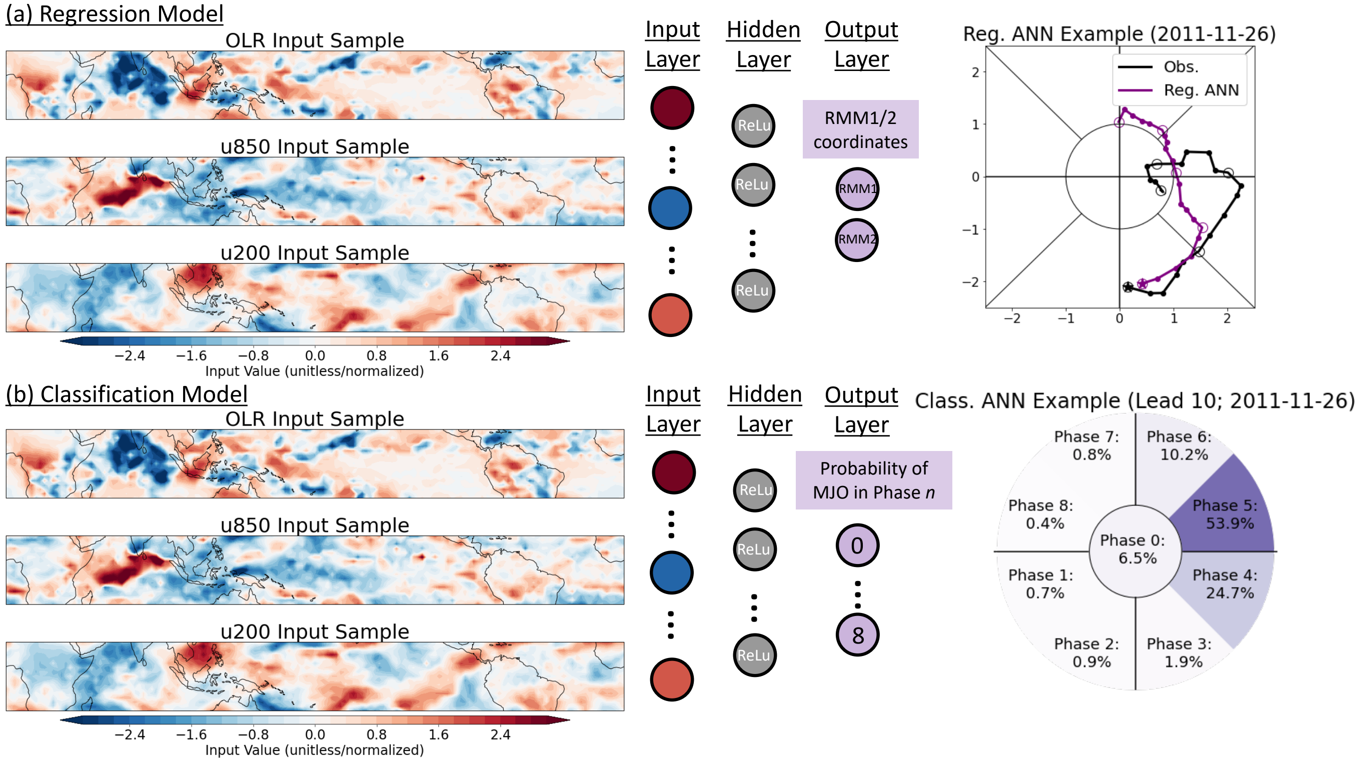

Figure 1. ANN model schematics. (a) The regression ANN; leftmost panels show a sample input of OLR and zonal wind at 850 hPa (u850) and 200 hPa (u200) from November 26, 2011. The input is passed through a 16-node hidden layer with a rectified linear unit ("ReLU") activation function. The regression ANN outputs values of RMM1 and RMM2 at a single lead time, and separate ANNs are trained for leads from 0-20 days. An example 20-day ANN forecast (purple) versus observations (black) is shown in the rightmost panel; dots denote days with open circles every five days. (b) The classification ANN; input is identical to the regression ANN, but the output is the probability the MJO is active in RMM phase 1-8 or is inactive ("phase 0"). An example forecast at a 10-day lead from November 26, 2011 is shown on the right. The model correctly identifies the MJO as in phase 5.

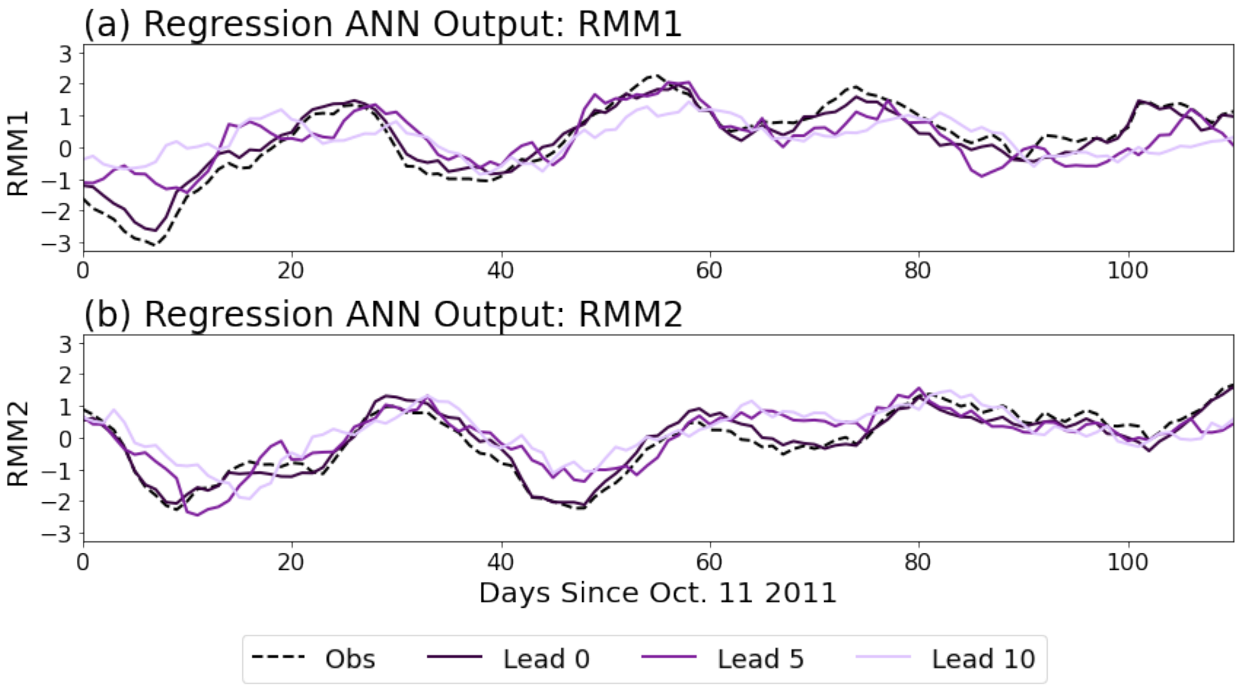

Figure 2 Regression ANN example. Example output from the regression ANN during one extended winter season. The observed RMM1 and RMM2 values are shown in black dashed. The regression ANN prediction for each day at a lead of 0, 5, and 10 days are shown in shades of purple.

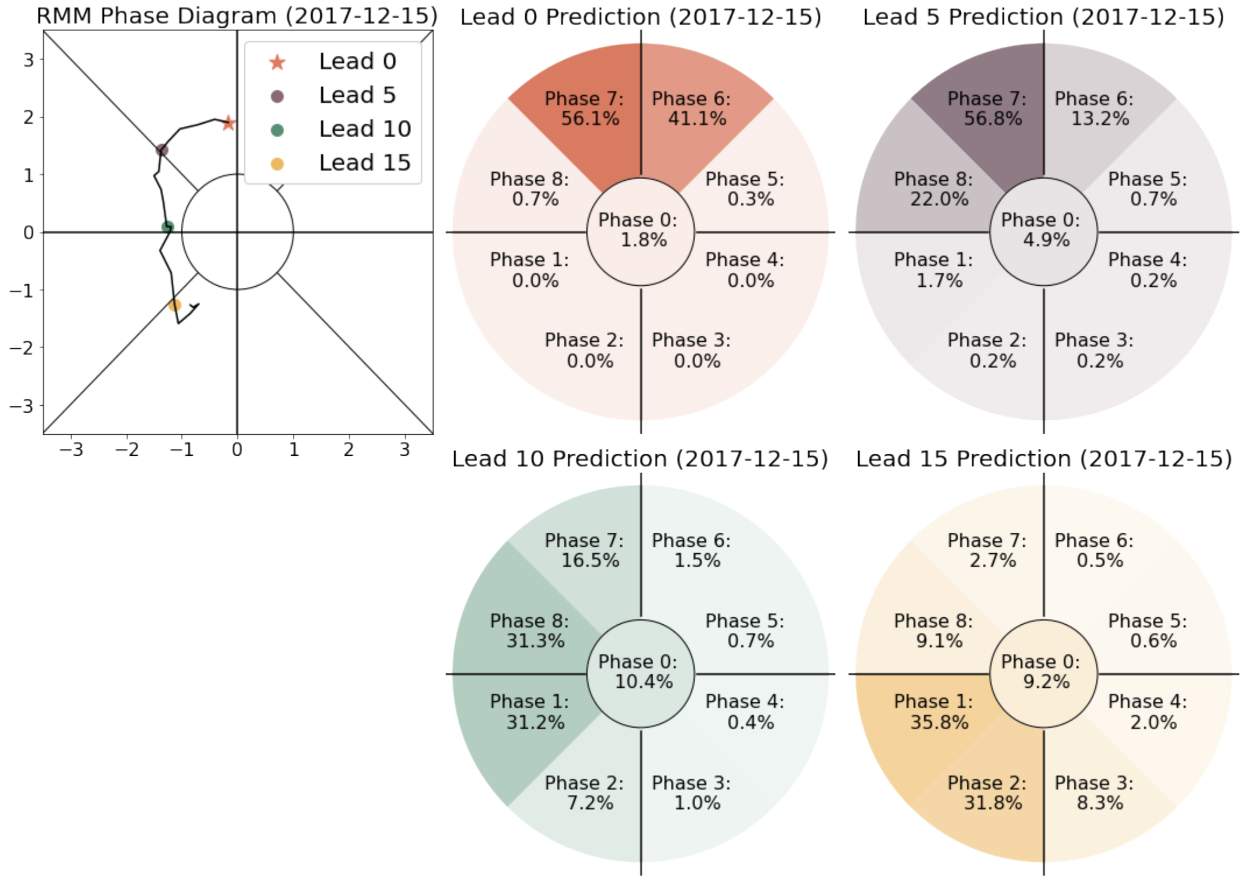

Figure 3. Classification ANN example forecast. Example output from the classification ANN for lead times of 0, 5, 10, and 15 days. The left panel shows the observed RMM index for 20 days beginning December 15, 2017. The right four panels show the classification ANN confidence for each of the 9 MJO phases at the indicated lead time. The predicted class is the one with the highest probability; in this example predictions are phase 7 (lead 0; correct), phase 7 (lead 5; correct), phase 8 (lead 10; correct), and phase 1 (lead 15; incorrect).

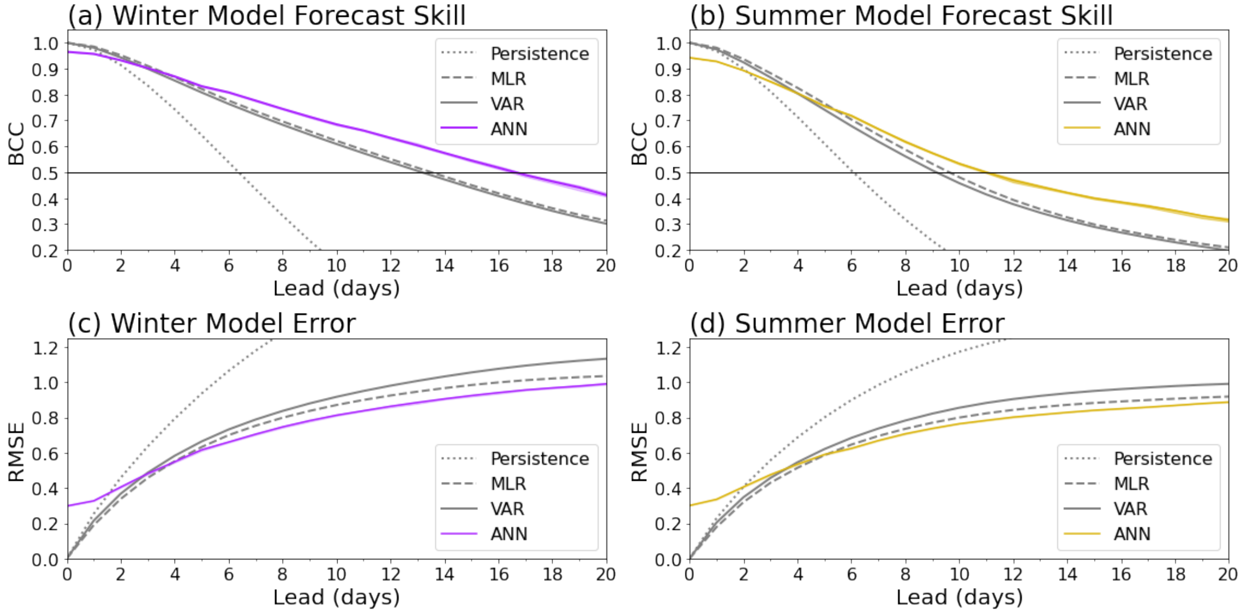

Figure 4. Regression ANN overall performance. RMM prediction skill (a/b) and root-mean-square error (c/d) for the regression ANN (purple/gold) and other simple statistical models (grey). Skill in the top panels is measured via the bivariate correlation coefficient (BCC); a threshold of 0.5 denotes skill.

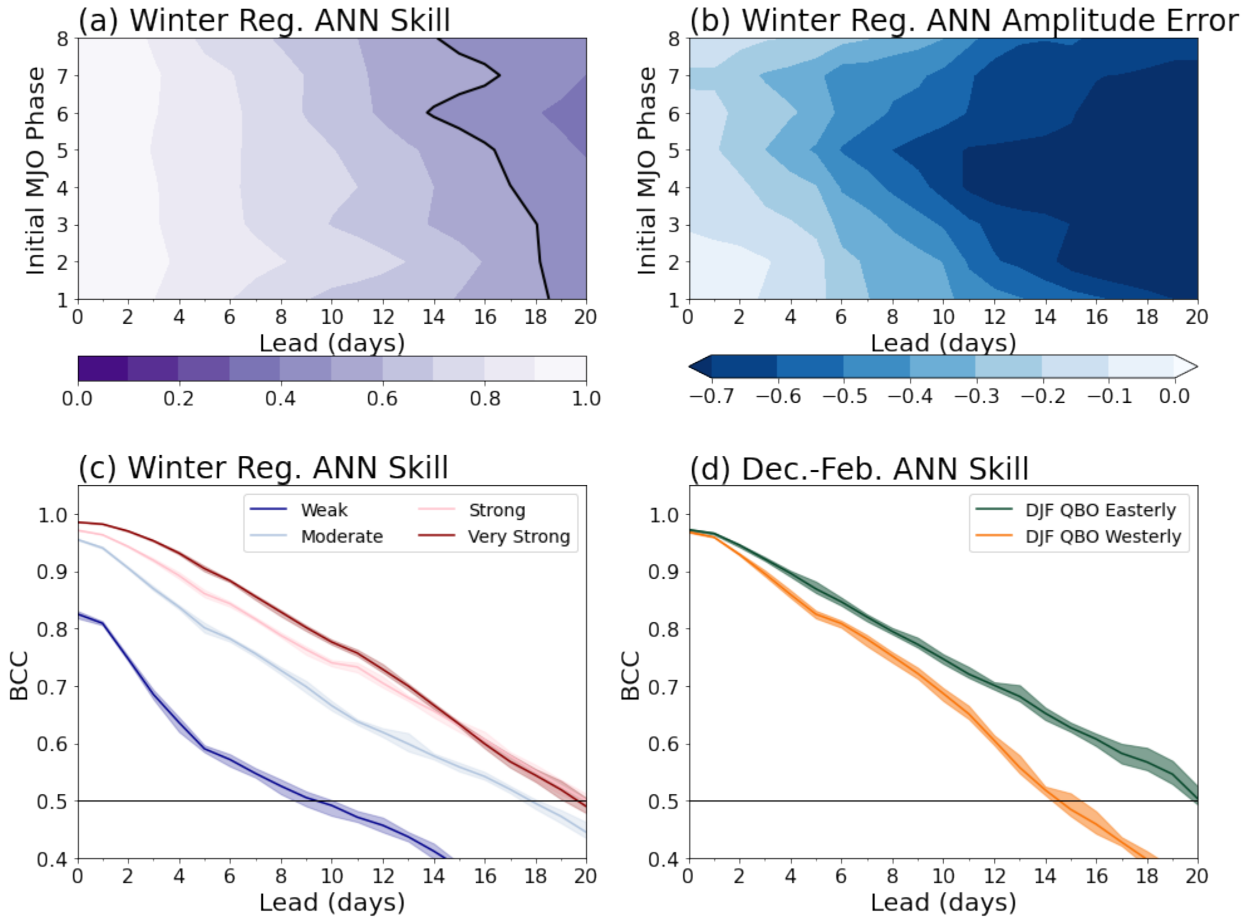

Figure 5. Regression ANN detailed performance. (a) The BCC as a function of initial MJO phase, without a threshold for MJO activity (i.e. all days are assigned a phase 1-8). Black line denotes a BCC of 0.5. (b) The average RMM amplitude difference between observations and ANN-forecasted events: negative 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 December-February separated by phase of the stratospheric quasi-biennial oscillation, defined using the U50 index. Shading in panels (c/d) denotes the spread across a 10-member ANN ensemble.

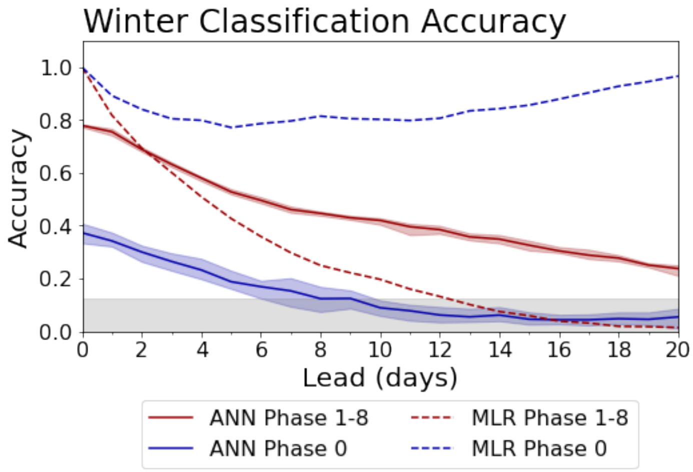

Figure 6. Classification model accuracy. Winter classification ANN accuracy forecasting active MJO days (phase 1-8; red) and accuracy for weak MJO days (phase 0; blue). Dashed line is the same but for the MLR model. Grey shading indicates random chance (1/9) assuming all classes are equally likely. Blue/red shading denotes the spread across a 10-member ANN ensemble.

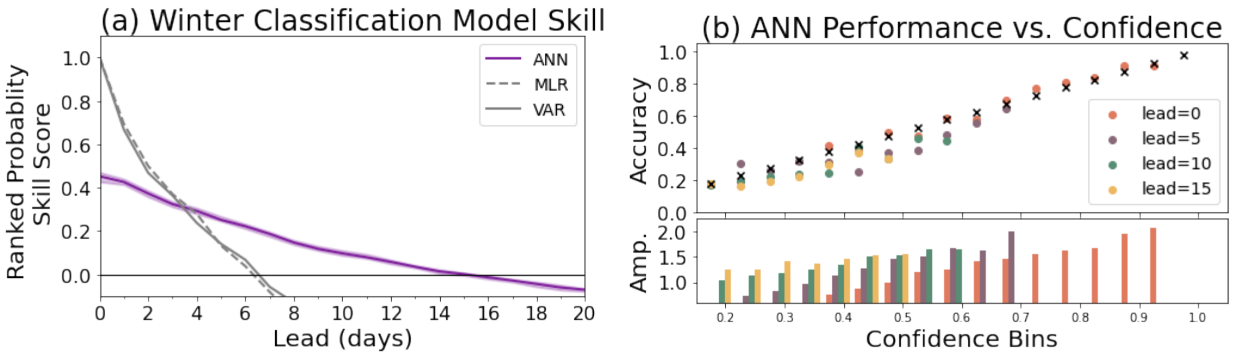

Figure 7. Classification model probabilistic forecasting. (a) The ranked probability skill score in winter for the ANN, MLR, and VAR model predictions relative to climatology; a score greater than zero denotes skill. (b) Winter classification ANN accuracy (top panel) and initial observed MJO amplitude (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.

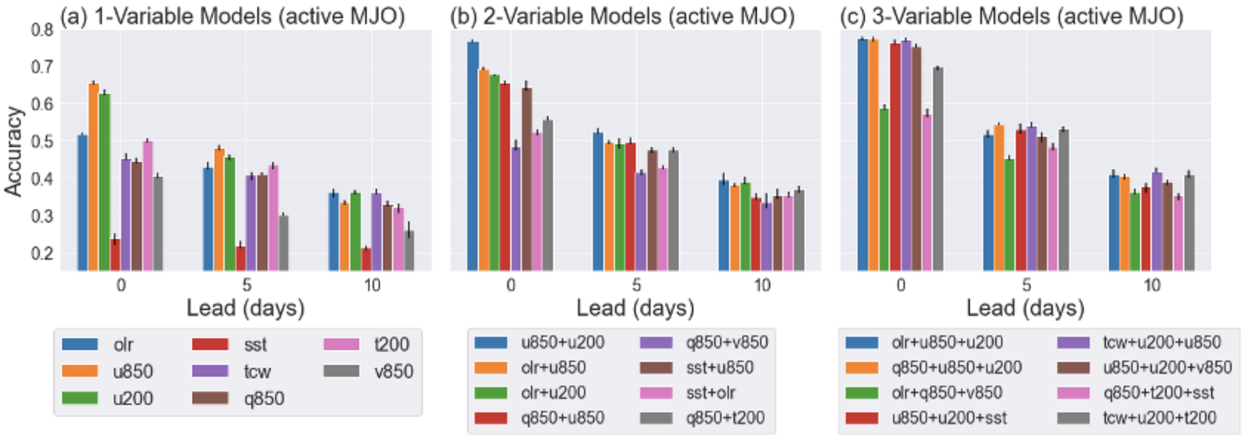

Figure 8 Sensitivity to input variables. Winter classification ANN accuracy predicting active MJO days at leads of 0, 5, and 10 days given different input variables. 1-variable (panel a), 2-variable (panel b), and 3-variable (panel c) models are shown. For each model, 5 ANNs are trained with different 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), and meridional wind (v), with numbers indicating the pressure level where relevant.

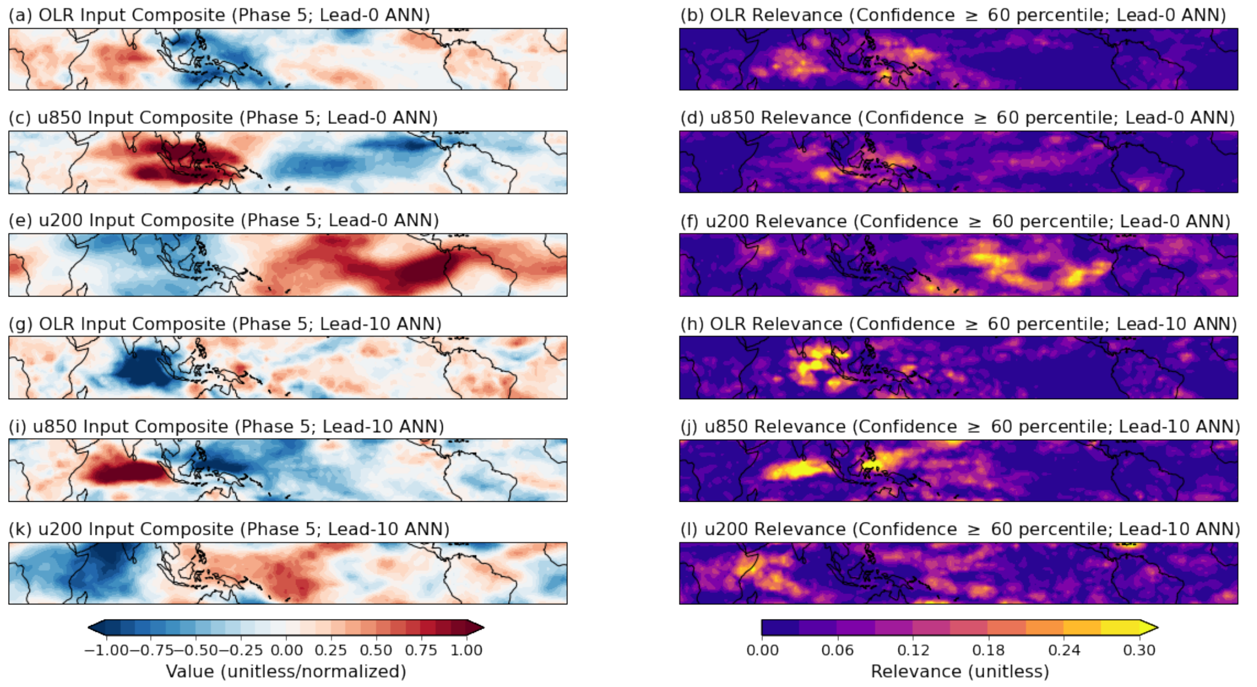
(a) OLR Input Composite (Phase 5; Lead-0 ANN)
(b) OLR Relevance (Confidence ≥ 60 percentile; Lead-0 ANN)
(c) u850 Input Composite (Phase 5; Lead-0 ANN)
(d) u850 Relevance (Confidence ≥ 60 percentile; Lead-0 ANN)
(e) u200 Input Composite (Phase 5; Lead-0 ANN)
(f) u200 Relevance (Confidence ≥ 60 percentile; Lead-0 ANN)
(g) OLR Input Composite (Phase 5; Lead-10 ANN)
(h) OLR Relevance (Confidence ≥ 60 percentile; Lead-10 ANN)
(i) u850 Input Composite (Phase 5; Lead-10 ANN)
(j) u850 Relevance (Confidence ≥ 60 percentile; Lead-10 ANN)
(k) u200 Input Composite (Phase 5; Lead-10 ANN)
(l) u200 Relevance (Confidence ≥ 60 percentile; Lead-10 ANN)
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
Value (unitless/normalized)
0.00 0.06 0.12 0.18 0.24 0.30
Relevance (unitless)

Figure 9. Layer-wise relevance propagation example. Composites of normalized input variables (left column) and LRP relevance (right column) for correct classification ANN predictions of MJO 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 3 variables: OLR, and 850 hPa zonal wind (u850) and 200 hPa zonal wind (u200).

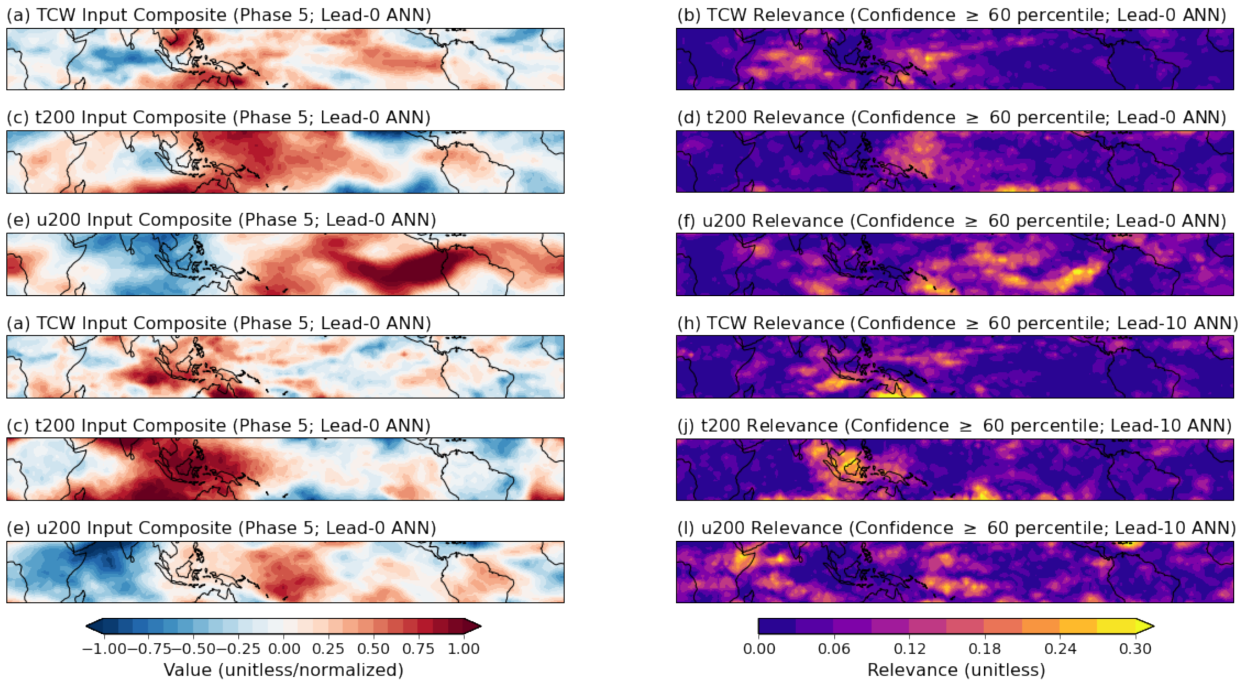
(a) TCW Input Composite (Phase 5; Lead-0 ANN)
(b) TCW Relevance (Confidence ≥ 60 percentile; Lead-0 ANN)
(c) t200 Input Composite (Phase 5; Lead-0 ANN)
(d) t200 Relevance (Confidence ≥ 60 percentile; Lead-0 ANN)
(e) u200 Input Composite (Phase 5; Lead-0 ANN)
(f) u200 Relevance (Confidence ≥ 60 percentile; Lead-0 ANN)
(a) TCW Input Composite (Phase 5; Lead-0 ANN)
(h) TCW Relevance (Confidence ≥ 60 percentile; Lead-10 ANN)
(c) t200 Input Composite (Phase 5; Lead-0 ANN)
(j) t200 Relevance (Confidence ≥ 60 percentile; Lead-10 ANN)
(e) u200 Input Composite (Phase 5; Lead-0 ANN)
(l) u200 Relevance (Confidence ≥ 60 percentile; Lead-10 ANN)
−1.00−0.75−0.50−0.25 0.00 0.25 0.50 0.75 1.00
Value (unitless/normalized)
0.00 0.06 0.12 0.18 0.24 0.30
Relevance (unitless)

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

Journal of Advances in Modeling Earth Systems

Supporting Information for

Using simple, explainable neural networks to predict the Madden-Julian oscillation

Zane K. Martin[1]

Elizabeth A. Barnes[1]

Eric Maloney[1]

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

Contents of this file:

1. Text S1 - Sensitivity Tests
2. Figures S1-S5

Text S1. Sensitivity Tests

In developing both the regression and classification ANN architectures, we conducted many tests exploring sensitivity to the processing of the input data, the ANN architectures, and the nature of model output. We show results from some of these sensitivity tests in Figures S1-S5 below: here we provide additional methodological detail regarding those tests. All tests below are shown for models during winter and, unless noted, for models that input OLR and zonal wind at 850 and 200 hPa.

In Figure S1, we compare the regression and classification ANN skill in a model trained using all-year data evaluated over winter and summer periods, versus the models trained in winter and summer respectively. While changes are modest, we found season-specific training to be somewhat advantageous in improving skill.

In Figure S2, we show the sensitivity to a change in how the regression ANN is trained. Rather than training the regression ANN on all winter days, we instead train the model on all active MJO days and a random subset of inactive MJO days such that weak MJO days are 1/9 of the overall training datasets. This is analogous to how the classification model is trained (see Section 3.1.1), and provides the regression ANN with more strong MJO samples at all lead times. While it marginally improves the accuracy of the regression model when active MJO days are considered (Fig. S2), it does not have a large change on the overall accuracy or the BCC; the regression model still shows poor performance forecasting active MJO events at leads longer than a few days.

In Figure S3 we show accuracy over active MJO days from the classification ANN at lead times of 0, 2, 5, 10, 15, and 20 days from a range of sensitivity tests. For the “control” test, the model is the same as that discussed in Section 4.1, with the range across 10-member ANN ensemble shown to capture spread due to random initial ANN weights. For the sensitivity tests only one ANN was trained at each lead time.

The first set of sensitivity tests shown in Figure S3 are slight changes to the ANN architecture. For the "high_ridge" test the ridge regression penalty was increased from 0.25 to 1, and for the "low_ridge" test the penalty was decreased from 0.25 to 0.1. For the "wide_net" test the number of nodes in the hidden layer was increased from 16 to 64, and in the "deep_net" the single, 16 node single layer was replaced with 2 fully connected layers of 16 nodes each. Note across these tests, large changes relative to the control are not observed, and typically fall within the control spread (Fig. S3a).

The second set of tests explore changes to the model input. The "30NS" and "15NS" experiments alter the latitude bands over which the input data is retained. The "lat_avg" model takes the 15N-15S average of the input before feeding it into the neural network, such that the input is a function only of longitude (e.g. a vector of length 144 per variable). Further, in the "lat_avg" model the learning rate is increased to 0.001 from the 0.0005 value used in the control. The "prior_days" test includes not only the variables from forecast day 0 in the input, but also includes forecast day -5, doubling the size of the input vector.

Model performance in all of these tests lies within the range of the control, with the exception of the latitudinal averaging at lead times of less than 5 days, which shows notably higher accuracy. Because the RMM index takes 15N-15S averaged variables as input, this increase in accuracy at short leads is likely due to the fact that the input is more closely associated with the output (i.e. how the RMM is computed), making it easier for the ANN to learn the relationship between the latitudinally-averaged input and the RMM phase. The fact that this increase relative to the control fades at longer lead times suggests, consistent with the discussion in Section 4.2, that identifying the MJO at short leads is a different task than predicting MJO behavior. Because the improvement is only seen at short leads, and because we are interested in how the 2-D structure of input variables informs the ANN (e.g. for in the LRP plot in Figures 9 and 10), we prioritize the 2-D input approach. Many additional sensitivity tests were performed during model development, and similar tests were performed for the regression model, but for brevity are not shown here, as results are comparable to those discussed.

A third sensitivity test, shown in Figure S4, quantifies sensitivity to training the regression ANN using a longer training dataset than NOAA OLR and ERA5 data allow. For this, we use ERA 20th century reanalysis daily OLR and zonal wind at 850 and 200 hPa data (Poli et al. 2016), which we obtained over the full period of availability from January 1, 1901 to October 31, 2010. ERA-20C input data is processed identically to the input for ERA-5 described in Section 2.1. The RMM index is calculated from ERA-20C using the method described in Wheeler and Hendon (2004); over the period in which the ERA-20C data overlaps with the observed RMM index, we found the correlation between our calculated ERA-20C RMM1/2 and the observed RMM1/2 values to be approximately .89, indicating good agreement in how the RMM index is formed.

We train a regression ANN with an architecture identical to that discussed in Section 3.1.1 but using ERA-20C data instead of ERA-5 data. The validation period is January 1, 2001 to October 31, 2010. We explored varying the training dataset to see whether model performance improved if much more training data was included. For lead times of 0, 5, 10, and 15 days we trained separate models for 11 different training periods. All training periods end December 31, 1999, but start dates vary across June 1 of: 1994, 1989, 1984, 1979, 1974, 1969, 1959, 1949, 1929, 1909, and 1901. To facilitate comparison to ERA5, we trained an additional model on

NOAA/ERA5 data from June 1979 to December 1999, and validated on NOAA/ERA5 data from January 1, 2001 to October 31, 2010.

Results in Figure S4 show generally comparable performance between ERA5 and ERA20C when the same period is used for validation and training. For reasons that we did not explore in depth, the ERA20C model shows higher BCC values at 0 and 5 days than NOAA/ERA5, comparable performance at 10 days, and worse performance at 15 days. More importantly, Figure S4 indicates that training the simple ANN on ERA20C with significantly more data does not lead to substantial improvement in the BCC at any lead time after between 120 and 200 months. Further tests with wider or deeper ANNs using the full 1901-1999 period of training also did not show improved performance.

The final sensitivity test, shown in Figure S5, explores sensitivity to including four or more additional input variables, following the same procedure as described in the manuscript in Section 4.2. Legend conventions in Figure S5 follow Figure 8, except “d” denotes divergence. Overall, no substantial increase in skill is seen in models with four or more variables.

Supplemental Figures.

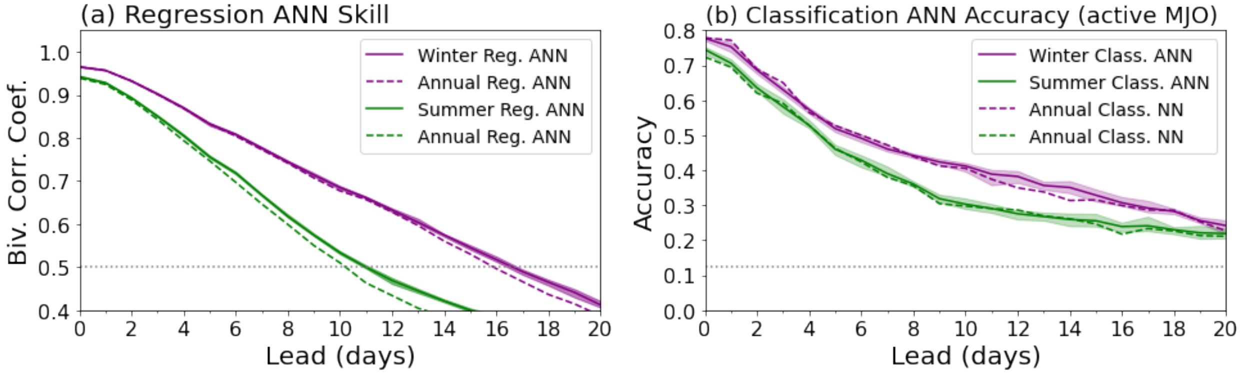

Figure S1 Regression ANN (panel a) and classification ANN (panel b) performance, similar to Figures 4 and 6, for ANNs trained specifically on winter and summer seasons (solid lines) versus a model trained on all seasons and evaluated separately in summer and winter (dashed line). The shading shows the seasonal model range across 10 ensemble models; for the annual model only one ensemble is considered.

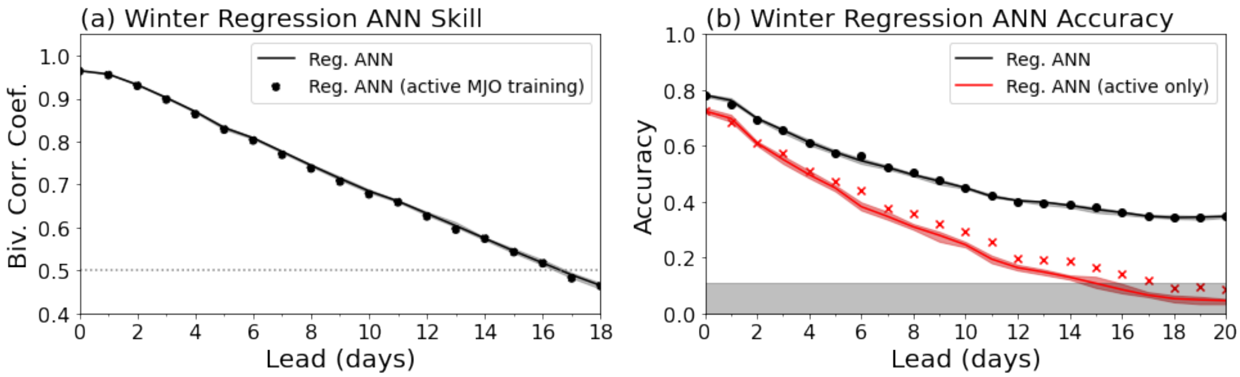

Figure S2 Winter regression ANN skill (panel a as in Figure 4) or accuracy of MJO phase (right; similar to Figure 6) for regression ANNs trained on all MJO days (lines) versus ANNs trained using fewer weak MJO days (as described in Supplemental Text S1; dots or x's). In the panel (b), black curves/dots are regression model accuracy evaluated over all MJO days, and red curves/x's are regression model accuracy evaluated only for active MJO days. Note the poor performance for active days, caused by the inability of the regression model to predict strong amplitude events.

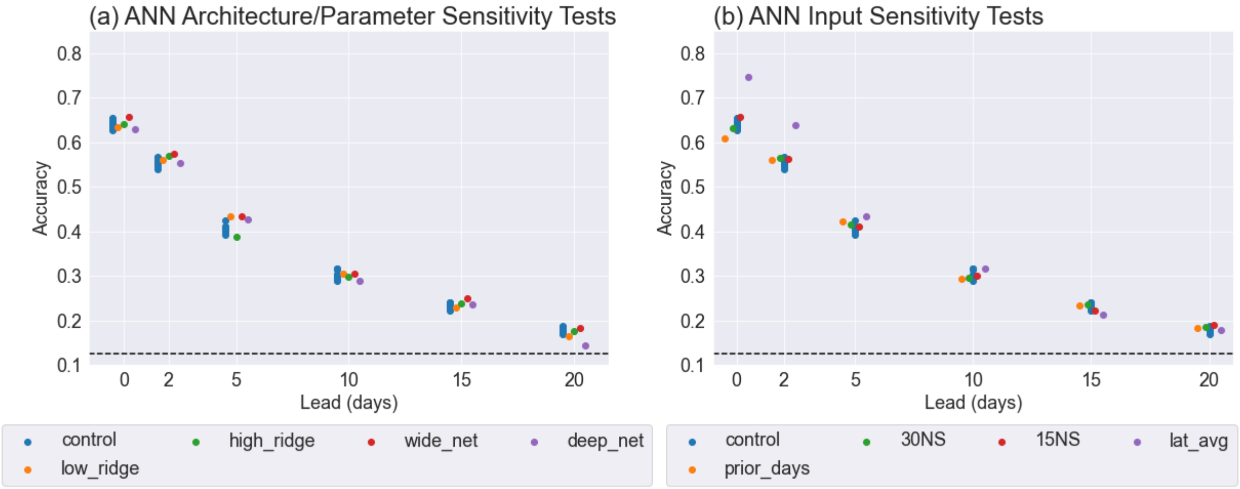

Figure S3 Model accuracy over active MJO days for sensitivity tests varying the architecture or hyperparameters of the ANN (panel a) and varying the model input (panel b). Tests are indicated by the legend as described in Supplemental Text S1. For the control, a 5 ANN ensemble was used.

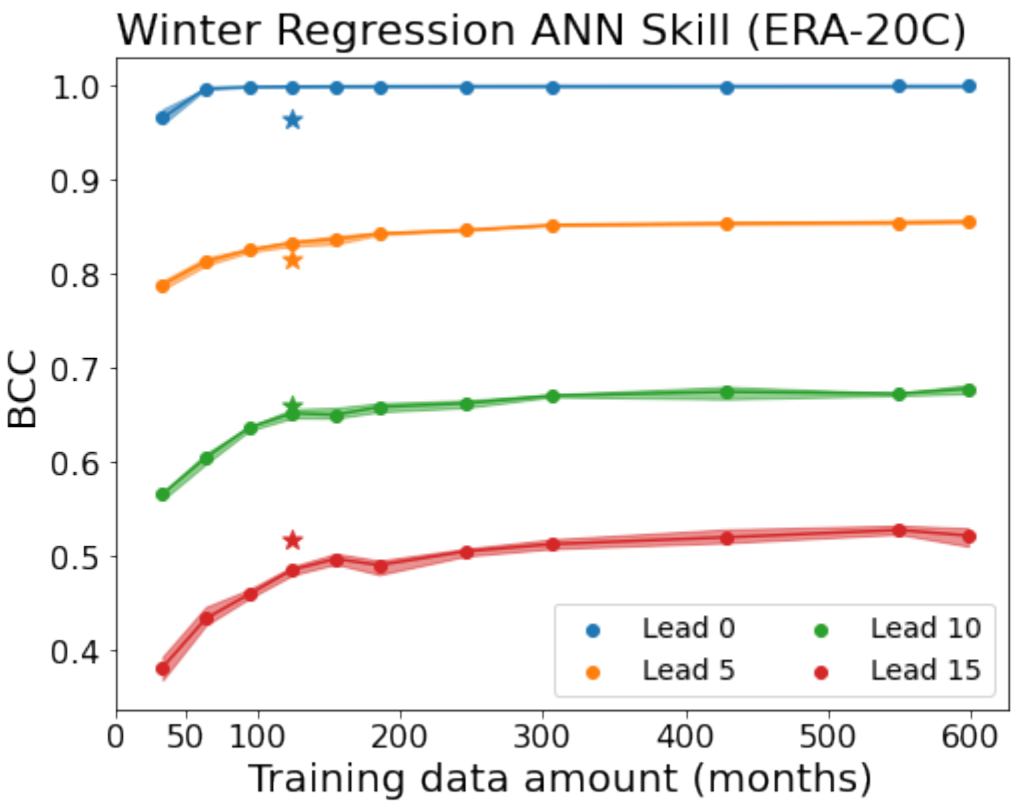

Figure S4 Winter regression ANN skill at lead times of 0, 5, 10, and 15 days (colors), trained using ERA-20C data. Models are trained using larger amounts of training data (dots; see Supplemental Text S1), and the *x*-axis shows the number of months in the training data period. Shading shows the range across 5 ANNs with different random starting weights. Stars show results using ERA5/NOAA winds and OLR data, as described in Supplemental Text S1.

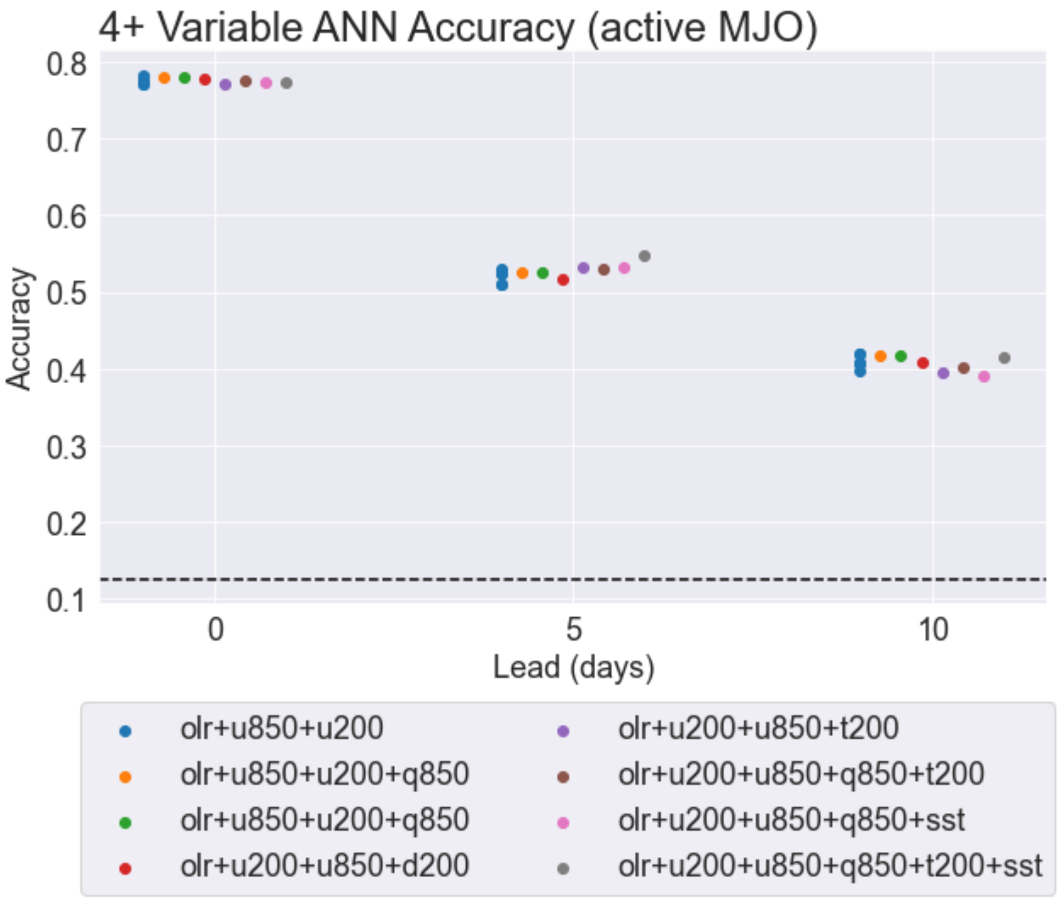

Figure S5 Similar to Figures 8 and S3, but for a series of tests with 4, 5, or 6 input variables. The blue "olr+u850+200" model is the same as in Figure 8c; other models have only 1 ensemble member.