Subseasonal Forecasts of Opportunity Identified by an Interpretable Neural Network

Kirsten J Mayer[1,1] and Elizabeth A. Barnes[2,2]

[1]Colorado State University - Department of Atmospheric Science
[2]Colorado State University

November 30, 2022

Abstract

Midlatitude prediction on subseasonal timescales is difficult due to the chaotic nature of the atmosphere and often requires the identification of favorable atmospheric conditions that may lead to enhanced skill ("forecasts of opportunity"). Here, we demonstrate that an artificial neural network can identify such opportunities for tropical-extratropical teleconnections to the North Atlantic circulation at a lead of 22 days using the network's confidence in a given prediction. Furthermore, layer-wise relevance propagation, an ANN interpretability technique, pinpoints the relevant tropical features the ANN uses to make accurate predictions. We find that layer-wise relevance propagation identifies tropical hot spots that correspond to known favorable regions for midlatitude teleconnections and reveals a potential new pattern for prediction over the North Atlantic on subseasonal timescales.

Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network

Kirsten J. Mayer [1] **and Elizabeth A. Barnes** [1]

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

Key Points:

- Neural networks can be used to identify forecasts of opportunity for subseasonal prediction
- Neural network explainability techniques pinpoint relevant tropical regions for predictions in the North Atlantic
- Clustering of neural network relevance heat maps reveals a potential new forecast of opportunity for the North Atlantic

Corresponding author: Kirsten J. Mayer, kjmayer@rams.colostate.edu

Abstract

Midlatitude prediction on subseasonal timescales is difficult due to the chaotic nature of the atmosphere and often requires the identification of favorable atmospheric conditions that may lead to enhanced skill ("forecasts of opportunity"). Here, we demonstrate that an artificial neural network can identify such opportunities for tropical-extratropical circulation teleconnections within the North Atlantic (40°N, 325°E) at a lead of 22 days using the network's confidence in a given prediction. Furthermore, layer-wise relevance propagation, an ANN explainability technique, pinpoints the relevant tropical features the ANN uses to make accurate predictions. We find that layer-wise relevance propagation identifies tropical hot spots that correspond to known favorable regions for midlatitude teleconnections and reveals a potential new pattern for prediction in the North Atlantic on subseasonal timescales.

Plain Language Summary

Weather forecasting on 2 week to 2 month timescales is known for its lack of predictability due to the chaotic nature of the atmosphere. One way to improve prediction skill on these timescales involves the identification of periods of atmospheric conditions that lead to enhanced predictability ("forecasts of opportunities"). Here, we show that a neural network can accurately identify these opportunities when trying to predict the atmospheric circulation over the North Atlantic Ocean 4 weeks in advance. A neural network explainability technique is then used to uncover what the network has "learned" to make these accurate predictions. We show that the network identifies known patterns of storminess ideal for midlatitude prediction and uncovers a possible new favorable region for enhanced prediction.

1 Introduction

Subseasonal timescales (2 weeks - 2 months) are known for their lack of predictability (Mariotti et al., 2018), yet reliable and actionable information on these timescales are required for decision making in many sectors such as public health and water management (e.g. Vitart et al., 2012; White et al., 2017). Over the past decade, there has been a substantial research effort to improve prediction on these timescales (e.g. Vitart et al., 2012; Robertson et al., 2015; Vitart et al., 2017; Pegion et al., 2019). One area of subseasonal prediction research focuses on forecasts of opportunity, the idea that certain earth system conditions provide opportunities for enhanced subseasonal prediction skill (Mariotti et al. 2020). When these opportunities arise, the information provided by the earth system's state can then be leveraged to improve forecast skill. For example, when the Madden-Julian Oscillation (MJO; Madden and Julian (1971, 1972)), a propagating tropical convective phenomenon, is active, its convective heating can lead to the excitation of quasi-stationary Rossby waves (Hoskins and Ambrizzi 1993) that subsequently modulate the midlatitude circulation over the first few weeks following MJO activity (e.g., Hoskins and Karoly, 1981; Sardeshmukh and Hoskins, 1988; Henderson et al., 2016; Baggett et al., 2017; Zheng et al., 2018). When opposing convective anomalies are located over the Indian Ocean and western Pacific (defined as phases 2, 3, 6, and 7), the MJO has been shown to lead to more coherent and consistent modulations of midlatitude weather on subseasonal timescales and consequently, enhanced prediction skill (Tseng et al., 2018). Using the strength and location of tropical convective activity of the MJO to identify periods of enhanced midlatitude prediction skill is, therefore, an example of forecast of opportunity identification. Mundhenk et al. (2018) also show that an empirical model, which solely uses information about the state of the MJO and the Quasi-Biennial Oscillation, outperforms a state-of-the-art numerical prediction model for prediction of atmospheric river activity on subseasonal timescales. This highlights the importance of statistical models for enhancing subseasonal prediction.

Albers and Newman (2019) demonstrate a technique for forecast of opportunity identification through the utilization of expected skill from a linear inverse model. The study demonstrates the ability of the linear statistical model to identify forecasts of opportunity, and raises the question of whether other statistical models, such as artificial neural networks (ANNs), can identify forecasts of opportunity for subseasonal prediction. ANNs are very good at nonlinear function estimation (Chen & Chen, 1995), and thus, may be able to identify both linear and nonlinear relationships that lend predictability. Recently, ANNs have been successfully applied to seasonal prediction of meteorological variables such as monthly rainfall (Abbot & Marohasy, 2014) and surface temperature (Toms et al., 2020) as well as yearly prediction of the El Nino Southern Oscillation (Ham et al., 2019), suggesting ANNs may be useful for identifying subseasonal forecasts of opportunity as well.

In this paper, we test whether an ANN can be used for subseasonal forecast of opportunity identification. To do so, we input tropical outgoing longwave radiation (OLR) anomalies into an ANN and task the network to predict the sign of 500 hPa geopotential height (z500) anomalies in the North Atlantic (40°N, 325°E) 22 days later (e.g. Week 4). Tropical OLR is used to explore the ability of an ANN to identify known relationships between the MJO and the North Atlantic via tropical-extratropical teleconnections (e.g. Cassou, 2008; Henderson et al., 2016). We demonstrate that an ANN can identify subseasonal forecasts of opportunity related to tropical OLR, and through an ANN explainability technique, demonstrate that the ANN identifies these known MJO-like OLR patterns. In addition, we find a possible new tropical OLR pattern associated with predictable behavior of the North Atlantic circulation on subseasonal timescales.

2 Data and Methods

2.1 Data

We use daily mean OLR (1979-2019) from the National Center for Atmospheric Research/National Oceanic and Atmospheric Administration (NCAR/NOAA; Liebmann and Smith (1996)) and daily mean z500 (1979-2019) from the European Centre for Medium-Range Weather Forecasts (ECMWF) Interim reanalysis (ERA-I; Dee et al. (2011)). MJO teleconnections tend to be stronger during boreal winter (Madden, 1986), and therefore, the extended boreal winter months (November-February) are used for the OLR fields. Since we task the network to predict the sign of the z500 anomaly 22 days following a given OLR field, March is also included in the z500 analysis (see Text S1 for reasoning behind the choice of lead).

The annual cycle is removed from both the z500 and OLR data. For z500, the annual cycle is removed by subtracting the daily climatology over the record (1979-2019). A Fast Fourier Transform high-pass filter is then applied to the z500 anomalies to remove seasonal oscillations (frequencies smaller than $\frac{1}{120 days}$) to ensure the network focuses on subseasonal anomalies. The median of the z500 anomalies for the training data (see 2.2.1) is subtracted to obtain an equal number of positive and negative values. These anomalies are then converted into 0s and 1s depending on the sign (negative or positive, respectively). To filter the testing data, z500 anomalies from 2017-2019 are appended to the unfiltered z500 anomalies from 1979-2016 and another FFT high pass filter is applied to all years. The now filtered 2017-2019 data are then subset and used as testing data. The median of the z500 anomalies for the training data (see 2.2.1) is then subtracted and the anomalies are converted into 0s and 1s. For OLR, the annual cycle is removed by subtracting the first 3 harmonics of the daily climatology from the raw field. The first 3 harmonics are used instead of the daily mean because OLR is a noisier field than z500.

2.2 Methods

2.2.1 Artificial Neural Network Architecture

A two layer ANN (Figure 1) is tasked to ingest tropical OLR and predict the *sign* of the z500 anomaly over the North Atlantic (40°N, 325°E; red dot in Figure 1) 22 days later. The North Atlantic is chosen for this analysis since the MJO is known to force circulation anomalies over this region on subseasonal timescales and thus allows us to explore the utility of an ANN in the context of a well known problem (e.g. Cassou, 2008; Roundy et al., 2010; Henderson et al., 2016). In addition, we find that this grid point is representative of a larger area within the North Atlantic (see supplemental Figure S1).

Each input sample to the ANN consists of vectorized daily anomalous OLR from 30°N to 20°S and 45 to 210°E, where the number of input nodes is equal to the number of OLR grid points (N = 1407). The ANN then outputs two values that describe the categorical prediction, positive or negative sign of z500, given the initial OLR input image. The softmax activation function is applied to this final layer and transforms the two output values such that they sum to 1. The output then represents an estimation of the likelihood that an input belongs to a particular category. We refer to this estimation of likelihood as "model confidence". A more confident prediction will, therefore, have a predicted category value closer to 1. We define forecasts of opportunities as the top 10% most confident predictions by the network, although we explore alternative percentages as well.

The ANN architecture consists of two hidden layers of 128 and 8 nodes, respectively, and both use the rectified linear activation function. The final layer includes 2 nodes and uses the softmax activation function. Categorical cross entropy is used for the loss function. This architecture is chosen because it was found to consistently lead to reasonably high accuracies across many combinations of training/validation sets, but our ANN approach should be equally applicable to both shallow and deep networks. The batch size is set to 256 samples (i.e. OLR vectorized images) and the ANN is trained for 50 epochs unless the validation loss increases for two epochs in a row. If this occurs, the ANN stops training early and restores the model's best weights to reduce overfitting. It is found that 50 epochs is sufficient for training as the ANN rarely completes all 50 epochs. A more detailed explanation of ANNs is provided in the supplemental material for reference along with a comparison of this ANN approach to multinomial logistic regression.

The data used to train and test the ANN is composed of three groups: training, validation, and testing. Training and validation data are used during training, where training data is used to update the weights and biases of the ANN and the validation data is used to evaluate the model. The testing data is data that has never been "seen" by the ANN to evaluate the ability of the ANN to generalize to new data. To create the testing data, we assume that the years 2017-2019 have not yet occurred when training the model. In this way, these years act as true testing data for the ANN. While the specific accuracies likely would change with different testing data, the main point of this paper is to introduce a method to identify forecasts of opportunity and then to further identify the associated relevant regions for the enhanced prediction skill, not to provide the most accurate model for this scenario.

For this analysis, the ANN validation data is from November 2007 through February 2011 (N = 481) and the testing data is from November 2017 through February 2019 (N = 240). The remaining extended boreal winter (NDJF) data are used for training (November 1979 - February 2007 and November 2011 - February 2016; N = 4450; see supplemental Figure S2). All data is standardized for each grid point by the years used for training and validation. To choose a model for the following analysis, ANN training is repeated for a variety of validation years. Different consecutive four-year chunks are removed from the training data and set aside to use as validation. For each of the nine four-year chunks, the ANN was trained 20 times with random initialized weights. We find that our conclusions are robust to our choice in training period and do not change with variations

in random initialization weights. We present one model with reasonably high accuracy here and using the training, validation, and testing groups outlined above.

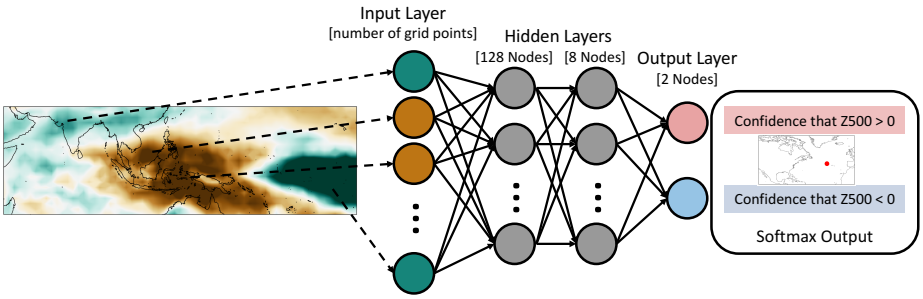

Figure 1. Artificial neural network architecture for prediction of the sign of z500 anomalies over the North Atlantic 22 days following tropical OLR anomalies. The neural network consists of two hidden layers of 128 and 8 nodes, respectively, and an output layer of two nodes (one node for each sign). The output layer uses the softmax activation function.

2.2.2 Layer-Wise Relevance Propagation (LRP)

While ANNs are a useful tool for making predictions, in doing so, they are learning *how* to make accurate predictions. Therefore, understanding the inner workings of a trained ANN can provide valuable information for improving prediction skill and understanding, as well as increasing user confidence in the results. Here, we utilize a relatively new neural network explainability technique to the geosciences called layer-wise relevance propagation (LRP; Bach et al. (2015); Montavon et al. (2019)) to extract and visualize the features the trained ANN employs to make accurate predictions. While Toms et al. (2020) describes the use of LRP for geoscience applications in detail, we briefly provide a high-level description here (see supplemental material for a more detailed explanation). After network training is completed, a single sample is passed through the network and a prediction is made (in our case, two output values are predicted). Our implementation of LRP then takes the highest of these values (i.e. the winning category) and back-propagates this value through the network via a series of predefined rules, ultimately distributing it across the input nodes (i.e. input gridpoints). What results is a heat map of "relevance" across the input space, where input nodes that are more relevant for the network's specific prediction for that sample are given higher relevance. This process is then repeated for every sample of interest, resulting in a unique relevance heat map for each sample. In our study, since the input layer consists of maps of OLR anomalies, the LRP heat maps are maps of the relevant tropical OLR patterns for each prediction of the circulation in the North Atlantic (40°N, 325°E). These maps are discussed in detail in Section 3.2.

3 Results

3.1 Identifying Forecasts of Opportunity

ANNs with the architecture shown in Figure 1 are trained 100 times with random initialized weights to predict the sign of the z500 anomalies 22 days following the tropical OLR anomalies. Figure 2a shows the distribution of the testing prediction accuracy for all 100 models, where dark teal represents the distribution of all predictions and light

teal represents the distribution of the 10% most confident predictions. The corresponding colored vertical dashed lines indicate a threshold for what is expected by random chance. To calculate the random chance accuracy threshold, 100,000 randomly generated groups (N=240 for all and N=24 for 10% most confident predictions) of zeros and ones are used to create a distribution of accuracies, and the 90^{th} percentile of this distribution is used as the random chance threshold. In Figure 2a, the top 10% most confident prediction accuracies (light teal) are shifted towards higher accuracies compared to the distribution with all predictions (dark teal). This shift in the distributions demonstrates that in general, higher model confidence leads to substantially enhanced prediction accuracy.

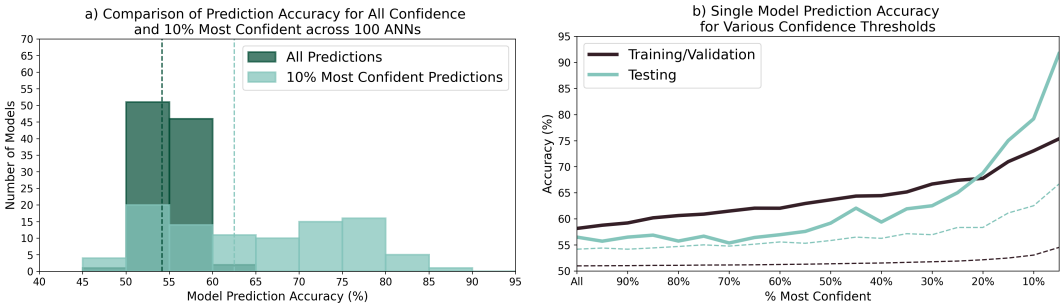

Figure 2. (a) Histograms of testing prediction accuracy for 100 trained ANNs. The dark teal represents the histogram of all prediction accuracies and the light teal represents the histogram for the 10% most confident prediction accuracies. The dark teal and light teal dashed lines in (a) are the maximum accuracies expected by random chance at the 90% confidence level for the corresponding colored histogram (see text for details). (b) Accuracy of one particular model as a function of the percent most confident predictions for training and validation (black) and testing (light teal) data. The dashed lines indicate the maximum accuracies expected by random chance at the 90% confidence level for the corresponding colored lines (see text for details).

We chose one model from Figure 2a to further understand how accuracy varies when a different percent model confidence is used (Figure 2b). The solid lines represent the accuracy across various model confidence values for training and validation (black) and testing (light teal) data sets. Figure 2b shows that the testing accuracy (light teal line) barely outperforms the random chance 90% confidence bound (light teal dashed line) for all predictions ("all") while the skill is substantially larger than random chance for the top 10% of predictions. Accuracy increasing with increasing model confidence is also apparent in the training and validation data. Together, Figure 2a and b illustrate that model confidence and prediction accuracy generally increase together and therefore, can be used to identify forecasts of opportunities, or periods of enhanced prediction skill. From this analysis, the 10% most confident predictions are chosen to define forecasts of opportunity since this threshold has one of the largest accuracy differences from random chance while still retaining 10% of the samples.

When evaluating the network with the training and validation data, the prediction accuracy for all predictions is 58% and for the top 10% most confident predictions is 73%. For the testing data, the prediction accuracy for all predictions is 56% and for the top 10% most confident predictions is 79%. The ANN predictions as a function of time are detailed in Figure S2, and additional skill metrics are provided in Figure S3 and Table S1.

3.2 Tropical Sources of Predictability

We have shown that ANNs can identify forecasts of opportunity using model confidence; however, understanding where this enhanced skill originates is critical for improving physical understanding as well as gaining trust in the network's predictions. To do so, layer-wise relevance propagation is used to identify the OLR patterns that lead the ANN to make correct predictions (see Section 2.2.2). The correct 10% most confident predictions from the training, validation and testing data sets are combined for this LRP analysis. All three sets of data are used instead of only testing data because all data sets have similar accuracies and LRP values (not shown). Thus, including all the data increases the sample sizes for the analysis. The shading in Figure 3c-h shows the regions the network found relevant, on average, to make confident and correct positive (Figure 3c,e,g) and negative (Figure 3d,f,h) z500 predictions. The contours correspond to the average OLR anomalies for these confident and correct predictions.

The average LRP heat map for the correct forecasts of opportunity of positive sign predictions (Figure 3c) indicates four hot spots, one over the southern Indian Ocean into the southern Maritime Continent (20-0°S, 70-130°E), one over the western Pacific (20-0°S, 155°E-170°E), another northwest of Hawaii (25°N, 170°W), and the fourth over Saudi Arabia (30°N, 40-60°E). The average LRP heat map for the correct forecasts of opportunity of negative sign predictions (Figure 3d) indicates four hot spots, one over the Maritime Continent (20-0°S, 110-150°E), one in the western and central Pacific Ocean (20-0°S, 155°E-170°W), another to the west of Hawaii (20°N, 170°W), and the fourth over Saudi Arabia (30°N, 40-60°E).

For both sign predictions, the hot spots over the Maritime Continent and the western Pacific have opposing signed OLR anomalies (contours) that straddle 150°E. These dipoles of convection over the Indian Ocean into the Maritime Continent and over the western Pacific have similar structures to phase 4-5 and phase 1,7-8 of the MJO (Wheeler & Hendon, 2004). This structure of OLR is consistent with previous research of MJO teleconnections over the North Atlantic for average lead times of 10-14 and 15-19 days (e.g. Cassou, 2008; Henderson et al., 2016; Henderson & Maloney, 2018; Tseng et al., 2018). In addition, this dipole structure is known to lead to higher pattern consistency of teleconnections in the midlatitudes (Tseng et al., 2019), which has been shown to lead to enhanced prediction skill (Tseng et al., 2018). Rossby waves initiated by the MJO tend to be quasi-stationary, which suggests that these OLR anomalies may also correspond to 22 day leads as well. This Maritime Continent and western Pacific Ocean dipole highlighted in part by LRP is therefore consistent with previous research and demonstrates that the ANN has learned physically relevant structures.

To test the robustness of these average LRP results for this particular ANN, we calculated the frequency of occurrence of average relevance hotspots greater than 0.5 for models with testing accuracies greater than 70% (Figure 3a,b, n = 42 models). We find that all of the hotspots (i.e. the MJO-like structure, the hot spot over Saudi Arabia and the hot spot west of Hawaii) are robust features for enhanced subseasonal prediction throughout these 42 models. In the next section, we hypothesize that the hot spot over Saudi Arabia is associated with the two-way relationship between the North Atlantic Oscillation (NAO) and the MJO (Lin et al., 2009). On the other hand, the hot spot west of Hawaii in both sign predictions is discussed as a possible new region relevant for enhanced subseasonal prediction.

3.2.1 K-means Clustering of LRP Maps

To further distinguish the relevant regions for the ANN's predictions, k-means clustering (Hartigan and Wong (1979), see supplemental material for more information) is applied to the LRP maps (Figure 3e-h). This analysis reveals that the composite LRP maps for each sign (Figure 3c,d) actually consist of multiple distinct patterns used by the ANN. For positive sign predictions (Figure 3e,g), both clusters have a hot spot located between the central Indian ocean and the maritime continent, which are associ-

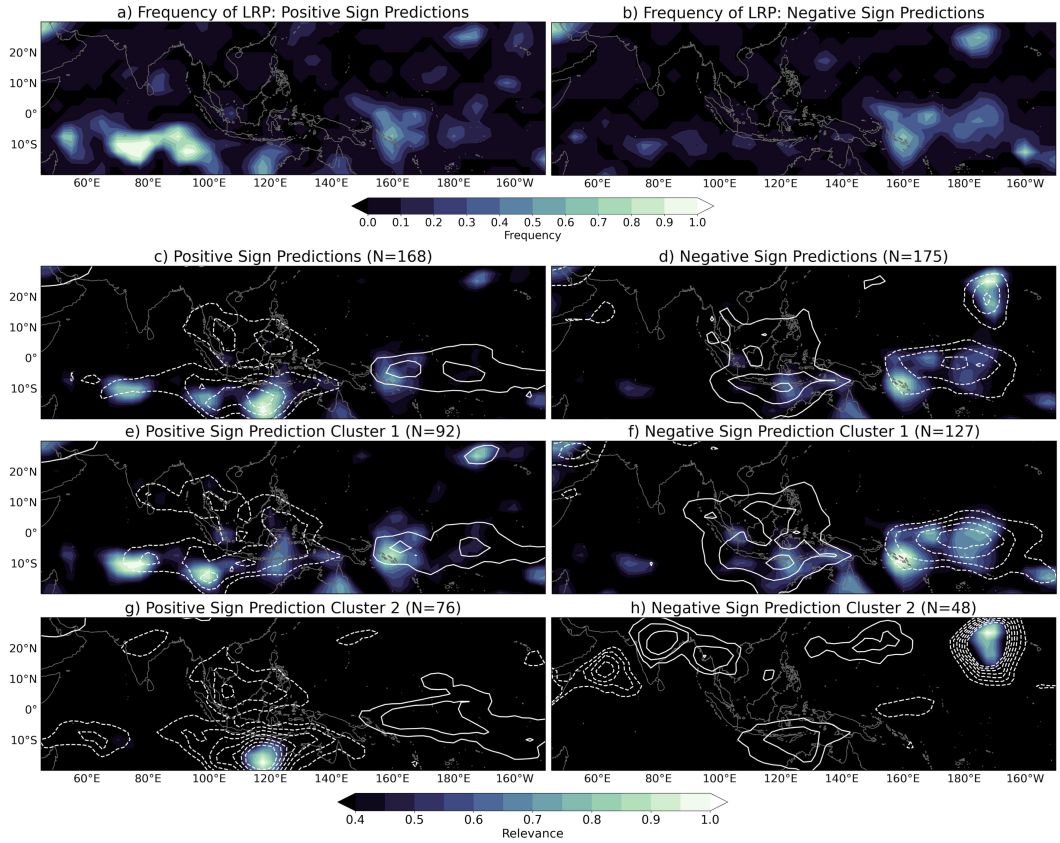

Figure 3. (a,b) LRP frequency of occurrence maps for average relevance values greater than 0.5. Both (a) and (b) consist of models from every 4-year validation chunk. Of these models, only average LRP maps of confident and correct predictions (training, validation, and testing) from models with testing accuracies greater than 70% are included. Maps (c-h) are the LRP maps associated with the ANN from Figure 2b where the shading denotes smoothed composites of LRP fields for all correct forecasts of opportunity for (c) positive sign and (d) negative sign predictions across training, validation and testing periods. The associated two k-means clusters of LRP for (e,g) positive sign predictions and (f,h) negative sign predictions are also shown. Contours represent the corresponding smoothed OLR anomalies where solid lines are positive values and dashed lines are negative values. (a) and (b) contours range from $0.4-1.0\frac{W}{m^2}$ and -1.0− $-0.4\frac{W}{m^2}$ and (e-h) contours range from 0.4-1.6$\frac{W}{m^2}$ and -1.6−$-0.4\frac{W}{m^2}$, both with a contour interval of $0.2\frac{W}{m^2}$.

ated with negative OLR anomalies. While not highlighted by LRP in cluster 2 (Figure 3g), each negative OLR anomaly region is accompanied by a region of positive OLR anomalies over the western Pacific. This suggests the model is identifying an MJO-like pattern. More specifically, the clustering has identified two types of relevance for this MJO-like pattern. The LRP map for cluster 1 (Figure 3e) highlights both the positive and negative OLR anomalies. As previously mentioned, these regions lead to more consistent midlatitude teleconnections (Tseng et al., 2018) and have been shown to be associated with a positive NAO anomaly (Cassou, 2008), which corresponds to a positive z500 anomaly at the predicted location. Cluster 1, therefore, supports previously identified tropical OLR regions and patterns ideal for enhanced prediction skill on subseasonal timescales in the North Atlantic. On the other hand, the LRP map for cluster 2 (Figure 3g) focuses exclusively on the south-central Maritime Continent, which is associated with enhanced

convection from the Indian Ocean to the Maritime Continent. This is more consistent with recent research that suggests that convection over the Indian Ocean dominates the formation of a positive NAO anomaly (Shao et al., 2020). This relationship is nicely illustrated in Figure 3a as the Indian Ocean is highlighted by the LRP analysis more often than the western Pacific.

For cluster 1 of the negative sign predictions (Figure 3f), there are two hot spots, one over the Maritime Continent and the other over the Pacific Ocean. As with the positive sign predictions, each hot spot is associated with opposing sign OLR anomalies, however, unlike cluster 1 of the positive sign predictions, the LRP analysis more strongly highlights the western Pacific region, and suggests that the network finds the region of enhanced convection more relevant. This is similar to cluster 2 of the positive sign predictions and is consistent with Figure 3b which shows that the region over the western Pacific is more often highlighted by the LRP analysis compared to the Maritime Continent. This suggests that the network often focuses on the region of enhanced convection for both sign predictions.

Unexpectedly, there is also a hot spot located over Saudi Arabia in cluster 1 for both positive and negative predictions. As seen in Figure 3a and b, this region is frequently highlighted by LRP in many ANNs. This hot spot appears to only be important when an MJO-like dipole structure is present. To the authors' knowledge, this region has not been shown to be important for tropical-extratropical teleconnections to the North Atlantic. However, previous research has shown that there is a two-way relationship between the MJO and NAO. Following the NAO, there tends to be a significant modulation of the tropical upper troposphere zonal wind over the Atlantic-Africa region (Lin et al., 2009). This modulation has been hypothesized to play a role in MJO initialization (Lin et al., 2009; Lin & Brunet, 2011). Since the NAO can persist over many weeks, the network may be identifying an influence of the NAO on the MJO and back to the NAO. We leave a deeper exploration of this possible mechanism to future work.

Unlike the other clusters, cluster 2 of the negative sign predictions (Figure 3h) has only one hot spot west of Hawaii (25°N, 170°W) and no MJO-like OLR anomalies. We hypothesize that this region is physically important as it is located south of the subtropical jet exit region and is associated with a large OLR anomaly. Rossby waves can be generated through advection of vorticity by upper level divergence or convergence associated with OLR anomalies (Sardeshmukh & Hoskins, 1988). Since this hot spot region is close to the jet exit region, these waves can more easily propagate into the midlatitudes or become trapped within the North Atlantic jet and directed into the North Atlantic (Hoskins & Karoly, 1981; Hoskins & Ambrizzi, 1993). Based on these known tropical-extratropical teleconnection dynamics, it is likely that this hot spot west of Hawaii is a new pattern identified by the ANN. This hot spot is also weakly apparent in cluster 1 of the positive sign predictions (Figure 3e), but is associated with MJO-like OLR anomalies. Given the lack of MJO-like patterns in cluster 2 of the negative sign predictions for this region, we hypothesize that this hot spot in cluster 1 of the positive sign prediction may not actually be associated with the MJO, but instead acting as an additional source of predictability.

4 Conclusions

Improving subseasonal prediction accuracy and understanding requires identifying opportunities that can lead to enhanced predictability (e.g. Mariotti et al., 2020). Here, we show that an artificial neural network can identify forecasts of opportunity for subseasonal prediction using the network's confidence in its prediction. In addition, we demonstrate that layer-wise relevance propagation can extract knowledge gained by the ANN to identify relevant physical tropical features important for the predictions. K-means clustering of the LRP maps further provides insight into multiple distinct patterns used by the ANN for enhanced prediction and reveals a possible new forecast of opportunity for prediction over the North Atlantic.

The hot spots identified by the ANN provide a stepping stone to further our understanding of tropical-extratropical teleconnections. For example, lagged composite analysis or simplified models can be used to further explore the physical mechanisms behind enhanced midlatitude predictability associated with these regions. In addition, analysis of the incorrect predictions made by the ANN may also be useful for improving our understanding of ideal tropical patterns for enhanced subseasonal prediction. Finally, while our application is focused on subseasonal prediction, the approach outlined here should be applicable to predictions across timescales. Ultimately, this paper demonstrates that ANNs are not only a useful tool for prediction, but can also be used to gain physical insight into predictability and subsequently, improve prediction skill.

Acknowledgments

This research is partially funded by the National Science Foundation Graduate Research Fellowship under the grant number 006784 supporting Kirsten J. Mayer and partially funded by the National Science Foundation Harnessing the Data Revolution through supporting Elizabeth A. Barnes with grant 1934668.

The authors declare that they have no conflict of interest.

Data availability: ERA-I reanalysis data are provided by the European Centre for Medium-Range Forecasts (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim; Dee et al., 2011). The interpolated OLR data is provided by the NOAA/OAR/ESRL PSL, Boulder, CO, USA (https://psl.noaa.gov/data/gridded/data.interp_OLR.html; Liebmann and Smith, 1996).

References

Abbot, J., & Marohasy, J. (2014, March). Input selection and optimisation for monthly rainfall forecasting in queensland, australia, using artificial neural networks. *Atmos. Res.*, *138*, 166–178.

Albers, J. R., & Newman, M. (2019, November). A priori identification of skillful extratropical subseasonal forecasts. *Geophys. Res. Lett.*, *46*(21), 12527–12536.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015, July). On Pixel-Wise explanations for Non-Linear classifier decisions by Layer-Wise relevance propagation. *PLoS One*, *10*(7), e0130140.

Cassou, C. (2008, September). Intraseasonal interaction between the Madden-Julian oscillation and the north atlantic oscillation. *Nature*, *455*(7212), 523–527.

Chen, T., & Chen, H. (1995). Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Trans. Neural Netw.*, *6*(4), 911–917.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., ... Others (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, *137*(656), 553–597.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019, September). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.*, *28*(1), 100–108.

Henderson, S. A., & Maloney, E. D. (2018, July). The impact of the Madden–Julian oscillation on High-Latitude winter blocking during el niño–southern oscillation events. *J. Clim.*, *31*(13), 5293–5318.

Henderson, S. A., Maloney, E. D., & Barnes, E. A. (2016, June). The influence of the Madden–Julian oscillation on northern hemisphere winter blocking. *J. Clim.*, *29*(12), 4597–4616.

Hoskins, B. J., & Ambrizzi, T. (1993, June). Rossby wave propagation on a realistic longitudinally varying flow. *J. Atmos. Sci.*, *50*(12), 1661–1671.

Hoskins, B. J., & Karoly, D. J. (1981, June). The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.*, *38*(6), 1179–1196.

Kingma, D. P., & Ba, J. (2014, December). Adam: A method for stochastic optimization.

Liebmann, B., & Smith, C. (1996). Description of a complete (interpolated) outgoing longwave radiation dataset. *Bull. Am. Meteorol. Soc.*, *77*, 1275–1277.

Lin, H., & Brunet, G. (2011, January). Impact of the north atlantic oscillation on the forecast skill of the Madden-Julian oscillation: IMPACT OF NAO ON MJO FORECAST. *Geophys. Res. Lett.*, *38*(2).

Lin, H., Brunet, G., & Derome, J. (2009, January). An observed connection between the north atlantic oscillation and the Madden–Julian oscillation. *J. Clim.*, *22*(2), 364–380.

Madden, R. A. (1986). Seasonal variations of the 40-50 day oscillation in the tropics. *J. Atmos. Sci.*, *43*(24), 3138–3158.

Madden, R. A., & Julian, P. R. (1971, July). Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. *J. Atmos. Sci.*, *28*(5), 702–708.

Madden, R. A., & Julian, P. R. (1972, September). Description of Global-Scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.*, *29*(6), 1109–1123.

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., ... Albers, J. (2020, January). Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Am. Meteorol. Soc.*.

Mariotti, A., Ruti, P. M., & Rixen, M. (2018, March). Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *npj Climate and Atmospheric Science*, *1*(1), 1–4.

McGovern, A., Lagerquist, R., Gagne, D. J., Eli Jergensen, G., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019, November). Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Am. Meteorol. Soc.*, *100*(11), 2175–2199.

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). Cham: Springer International Publishing.

Mundhenk, B. D., Barnes, E. A., Maloney, E. D., & Baggett, C. F. (2018, February). Skillful empirical subseasonal prediction of landfalling atmospheric river activity using the Madden–Julian oscillation and quasi-biennial oscillation. *npj Climate and Atmospheric Science*, *1*(1), 20177.

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). Determination press San Francisco, CA.

Pegion, K., Kirtman, B. P., Becker, E., Collins, D. C., LaJoie, E., Burgman, R., ... Kim, H. (2019, October). The subseasonal experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Am. Meteorol. Soc.*, *100*(10), 2043–2060.

Robertson, A. W., Kumar, A., Peña, M., & Vitart, F. (2015, March). Improving and promoting subseasonal to seasonal prediction. *Bull. Am. Meteorol. Soc.*, *96*(3), ES49–ES53.

Roundy, P. E., MacRitchie, K., Asuma, J., & Melino, T. (2010, August). Modulation of the global atmospheric circulation by combined activity in the Madden–Julian oscillation and the el niño–southern oscillation during boreal winter. *J. Clim.*, *23*(15), 4045–4059.

Sardeshmukh, P. D., & Hoskins, B. J. (1988, April). The generation of global ro-

tational flow by steady idealized tropical divergence. *J. Atmos. Sci.*, *45*(7), 1228–1251.

Shao, X., Straus, D. M., Li, S., Swenson, E., Yadav, P., & Song, J. (2020). Forcing of the MJO-related indian ocean heating on the intraseasonal lagged NAO.

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020, September). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *J. Adv. Model. Earth Syst.*, *12*(9).

Tseng, K.-C., Barnes, E. A., & Maloney, E. D. (2018, January). Prediction of the midlatitude response to strong Madden-Julian oscillation events on S2S time scales: PREDICTION OF Z500 AT S2S TIME SCALES. *Geophys. Res. Lett.*, *45*(1), 463–470.

Tseng, K.-C., Maloney, E., & Barnes, E. (2019, January). The consistency of MJO teleconnection patterns: An explanation using linear rossby wave theory. *J. Clim.*, *32*(2), 531–548.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., ... Zhang, L. (2017, January). The subseasonal to seasonal (S2S) prediction project database. *Bull. Am. Meteorol. Soc.*, *98*(1), 163–173.

Vitart, F., Robertson, A. W., & Anderson, D. L. T. (2012, January). Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *WMO Bull.*, *61*(61).

Wheeler, M. C., & Hendon, H. H. (2004, August). An All-Season Real-Time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Weather Rev.*, *132*(8), 1917–1932.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., ... Zebiak, S. E. (2017, July). Potential applications of subseasonal-to-seasonal (S2S) predictions. *Met. Apps*, *24*(3), 315–325.

Supporting Information for "Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network"

Kirsten J. Mayer [1] and Elizabeth A. Barnes [1]

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

Contents of this file

February 18, 2021, 6:08pm

Introduction Here we provide information about the choice of lead day for the predictand and a more detailed description of artificial neural networks (ANNs), layerwise relevance propagation (LRP), and k-means clustering. In addition, we include composite z500 figures for both positive and negative correct predictions, a timeseries of z500 anomaly predictions, as well as a confusion matrix and a table of additional skill metrics for all and the 10% most confident predictions.

Text S1: Reasoning behind Prediction of Lead Day 22

Previous research has shown that MJO impacts on the North Atlantic Oscillation occur approximately 5-15 days following phases 2-3 and 6-7 (Lin et al. 2009; Cassou 2008). Henderson et al. (2016) show that MJO impacts over the North Atlantic are statistically significant out to 20 days. In addition, Barnes et al. (2019) illustrate a causal connection between the MJO and NAO on the order of 15-20 days; however, they hypothesize that the MJO may still impact the NAO after the 20 days due to the autocorrelation of the NAO and MJO. Therefore, we evaluated the ANN on a variety of leads from 5-28 days. We found that the network performed well across leads within week 3 (days 15-21), but started to decrease in skill after lead day 22. A lead of 22 days is, therefore, used for our analysis, as it was one of the later leads with higher skill. While daily anomalies are used here, the ANN can also be used to predict a smoothed z500 anomaly (e.g. 7-day running mean anomalies). We find that the network performs similarly well for both weekly and daily anomalies, and therefore, use daily anomalies for this analysis.

Text S2: Artificial Neural Networks (ANNs)

In this analysis, we use an artificial neural network (ANN) as a tool for subseasonal forecast

of opportunity identification where Figure 1 shows the ANN architecture used for this analysis. The architecture includes an input layer (teal and brown nodes) and is followed by two hidden layers (grey nodes) and an output layer (red and blue nodes). The network is tasked to predict the sign of the geopotential height at 500hPa (z500) at a point in the North Atlantic (40°N, 325°E, white 'X' in Figure S1) given tropical OLR anomalies. The input layer receives vectorized OLR anomalies so that each input node represents an OLR anomaly from a single grid point. The output layer returns two values, one in each output node, where the nodes represent the sign of the z500 anomaly. The node with the larger value signifies the predicted sign of the z500 anomaly.

The network architecture is set up so that each node in a layer receives a value from the preceding layer. The value of a single node in a layer is calculated through a weighted sum of the incoming values in the preceding layer with an added bias (equation 1).

$$z_j = \sum_i w_{ij} x_i + b \tag{1}$$

In equation 1, j denotes the node for the value being calculated in a given layer and i denotes a node from the preceding layer. Therefore, w_{ij} signifies the weight connecting the ith and jth node and x_i represents the value of node i. b denotes the added bias term. A nonlinear transformation is then applied to z_j (equation 2). For this analysis, the Rectified Linear Unit (ReLU; equation 2) is used as the nonlinear activation function.

$$f(z_j) = max(0, z_j) \tag{2}$$

Both equation 1 and 2 are repeated for each node in the layer, which results in a single value ($f(z_j)$) for each node. These new calculated values are then be passed to the following layer and

the process continues. At the final layer, a softmax activation function is applied:

$$\tilde{y}_i = \frac{exp(x_i)}{\sum_j exp(x_j)} \tag{3}$$

where x_i represents the presoftmax value for output node i , the denomenator is the sum of the exponential of all the presoftmax output values, and $\tilde{y}_i$ represents the predicted output value for the ith output node. This function converts the raw values in the output layer into values that sum to one. By doing so, the output values then represent an estimation of likelihood that an input belongs to a particular category. We refer to this estimation of likelihood as "model confidence". A confident prediction will, therefore, have a value closer to one.

The architecture used here is often referred to as a fully-connected ANN since all the nodes from one layer are connected to all the nodes in the next layer. We have used the simplest ANN architecture that provided a relatively high accuracy since this set-up is sufficient for this application (two hidden layers). Additional information on ANNs can be found in Nielsen (2015) or Goodfellow et al. (2016).

In addition to the model architecture, there are also important parameters to specify for the training process. This includes the type of loss function, batch size, and number of epochs. The loss function estimates the accuracy of the predicted value to the actual value. For this example we use categorical cross entropy (equation 4) where $\tilde{y}_i$ is the predicted value of the ith node in the output layer and y_i is the actual value.

$$loss = -\sum_i y_i log(\tilde{y}_i) \tag{4}$$

This loss function assigns error to the ANN output so that larger errors are punished more than smaller errors due to the logarithmic transformation. The weights and biases of the neural network are updated using the gradient of the loss function through back propagation (a series

of chain-rule operations). An incremental step, defined here by the Adam method (Kingma & Ba, 2014), is then taken in the direction of greatest decrease along the loss function, in attempt to minimize the loss.

In addition, we also use ridge regression (L_2 norm penalty) to limit the magnitude of the coefficients. The penalty forces the model to combine values from many grid points for each prediction. We apply this additional penalty because individual grid points on the globe are spatially correlated with nearby points.

The weights and biases are updated after each batch, a subset of the training data. A batch size of 256 is used. After the network iterates through the entire training dataset using a batch of 256 (an epoch), the process is repeated again for a defined number of epochs. In this analysis, we use 50 epochs, however, we apply early stopping (ending the training before 50 epochs) if the validation loss increases for 2 epochs in a row. This is done in order to reduce overfitting on the training data.

Text S3: Multinomial Logistic Regression

Multinomial logistic regression (MLR) is a form of logistic regression that can be used for a multi-class problem. Using ANN terminology, the MLR architecture can be described as an input layer and an output layer, where the output values are passed through the softmax activation function. The ANN architecture used for this analysis is similar, but also includes two hidden layers. These hidden layers in the ANN make the ANN more complex than MLR and able to account for additional nonlinearities. As ANN and MLR methods are similar to one another, we compare the accuracies between the two methods for reference. We find that the ANN and multinomial logistic regression models have similar accuracies for the validation data, but the

ANN performs much better (over 20% higher accuracy) on the testing data than MLR. However, regardless of accuracies, we use an ANN for this paper, instead of MLR, since an ANN makes the methods more generalizable to other more complex nonlinear systems.

Text S4: ANN Explainability - Layerwise Relevance Propagation

To understand how a trained network makes its prediction, explainability techniques can be used to extract and visualize what the network has learned. In this paper, we use an explainability technique known as layerwise relevance propagation (LRP; e.g. Bach et al. (2015); Montavon et al. (2019)). To apply LRP, a single sample of interest is initially passed through the trained network (with frozen weights) to obtain a prediction. Using the output values without the softmax activation, the output node with the highest value (the predicted category) is back-propagated through the network using the following rule

$$R_i = \sum_j \frac{a_i w_{ij}^+ + max(0, b_j)}{\sum_i a_i w_{ij}^+ + max(0, b_j)} R_j \tag{5}$$

where i denotes the node of the layer to which the relevance is being back-propagated to while j denotes the node of the layer in which the relevance is from. R_i is therefore, the relevance translated backward to the ith node and R_j is the relevance of the jth node. The weight connecting the ith and jth nodes is denoted as w_{ij}^+ where the $+$ signifies that only the positive weights are used for back propagation. Lastly, a_i signifies the value of the ith node (post activation function) and b_j signifies the bias term of the jth node. The above relevance equation is for the LRP-$\alpha\beta$ method where $\alpha = 1$ and $\beta = 0$. This type of LRP method only propagates information associated with positive weights. In other words, only the information that positively contributed to the prediction is propagated backward.

For relevance back-propagation from the first hidden layer to the input layer, the following equation is used:

$$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j \tag{6}$$

At the input layer, the relevance values for each node can then be used to create a heatmap of relevance where more relevant nodes have larger values. This process is then repeated for every prediction of interest, resulting in a unique relevance heat map for each prediction. These maps show the relevant regions from the input sample that positively contributed to the prediction.

For more information on LRP as well as other neural network explainability techniques, see Toms, Barnes, and Ebert-Uphoff (2020) and McGovern et al. (2019).

Text S5: K-Means Clustering

K-Means cluster analysis (Hartigan & Wong, 1979) is used to group the correct prediction LRP maps to further explore relevant regions for enhanced prediction skill. K-means clustering categorizes input data into a user specified number of groups. The method iteratively assigns the given data to centroids based on the minimum squared Euclidean distance, where each data point is assigned to the closest centroid. The centroids are moved to the center of their assigned data points after an iteration and then the process begins again, for a user specified number of iterations. The data points associated with each centroid are part of that centroid's cluster.

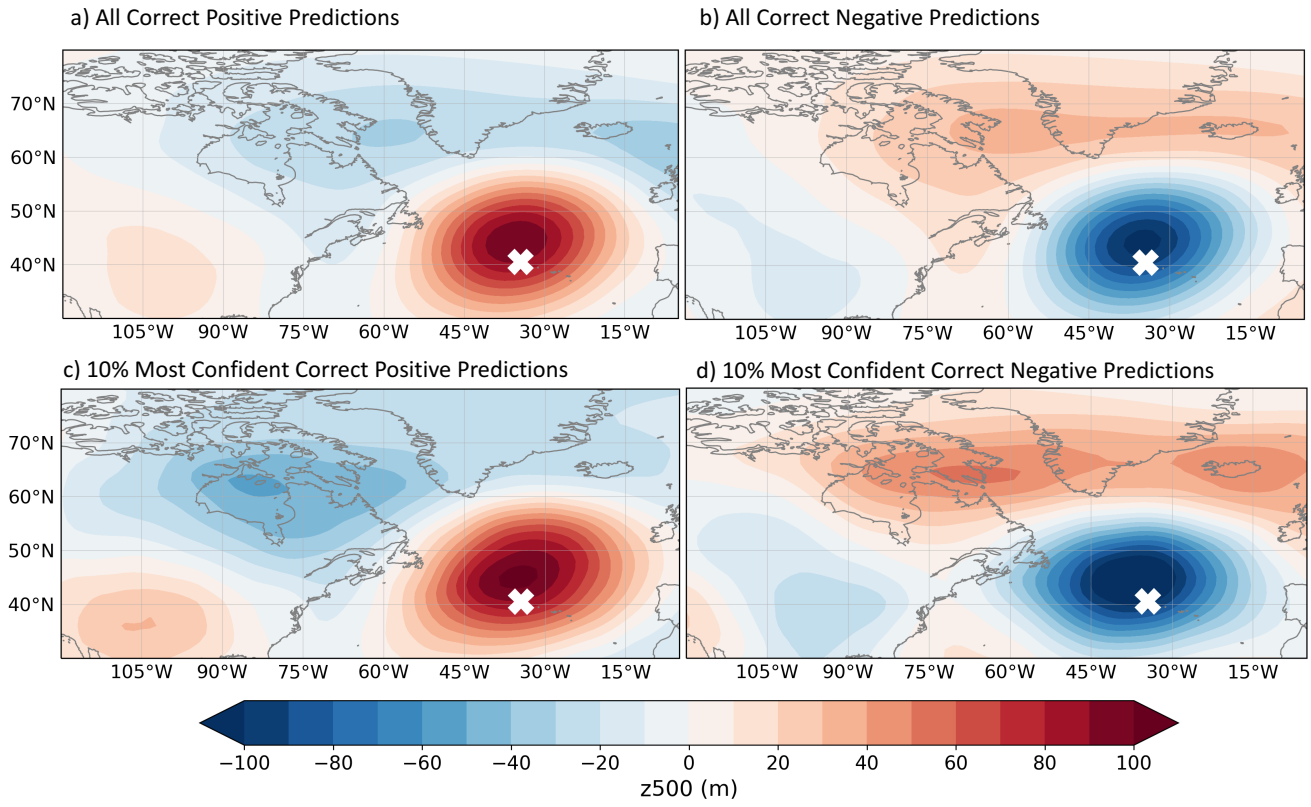

Figure S1. North Atlantic z500 composite: Composite of z500 anomalies for (a,b) all and the (c,d) 10% most confident predictions for correct (a,c) positive and (b,d) negative predictions. Shading represents the composite z500 anomalies and the white 'X' denotes the location of the ANN prediction over the North Atlantic (40°N, 325°E).

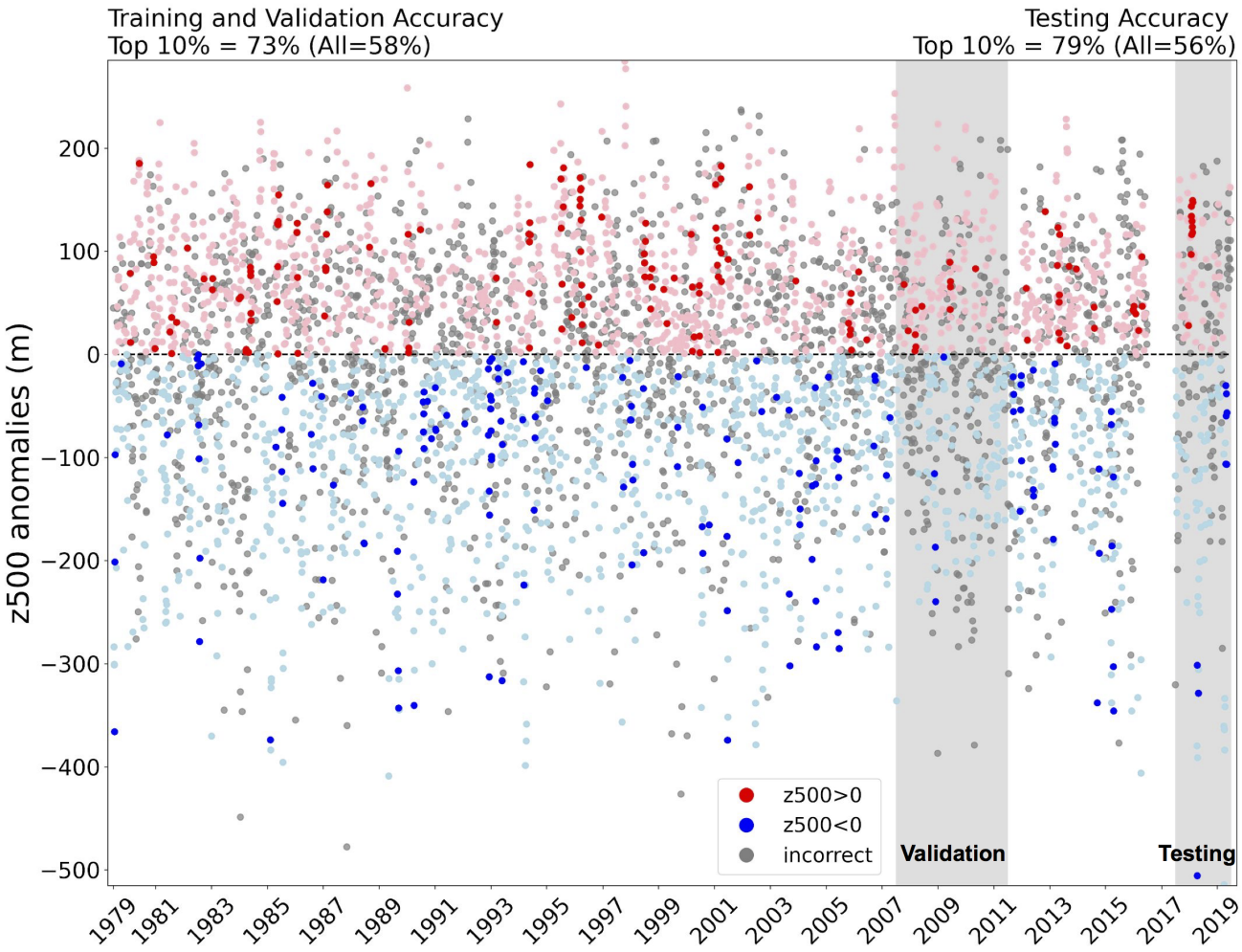

Figure S2. Timeseries of ANN z500 predictions: Timeseries of z500 anomalies shaded by the sign of the ANN predictions. Blue dots represent correct negative predictions, red dots represent correct positive predictions, and dark colored dots indicate forecasts of opportunities (i.e. 10% most confident predictions). Grey dots represent incorrect predictions. The vertical grey shading from 2007-2011 highlights the time period used for validation and the vertical grey shading from 2017-2019 highlights the time period used for testing. The accuracies for training and validation as well as testing data for forecasts of opportunities and all predictions are given in the top left and right, respectively.

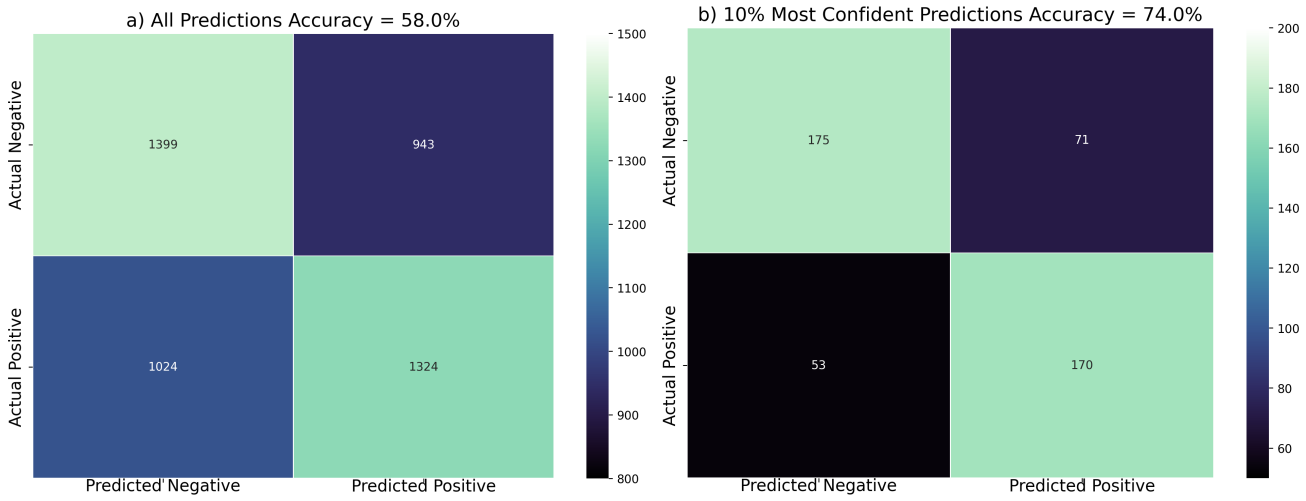

Figure S3. Confusion Matricies: Confusion matrix of training, validation, and testing data for (a) all predictions and (b) the 10% most confident predictions, where the accuracy is located at the top of each plot and the shading and the values inside each box represents the sample size for each category.

	(a) All Predictions			(b) 10% Most Confident Predictions		
	All	**Positive**	**Negative**	**All**	**Positive**	**Negative**
Accuracy	58%	-----	-----	74%	-----	-----
Precision	-----	58%	58%	-----	71%	77%
Recall	-----	56%	60%	-----	76%	71%

Table S1. Additional Skill Metrics: Table of accuracy, precision, and recall for (a) all predictions and (b) the 10% most confident predictions using training, validation, and testing data.