DISSERTATION

APPLICATION OF NEURAL NETWORKS TO SUBSEASONAL TO SEASONAL PREDICTABILITY IN PRESENT AND FUTURE CLIMATES

Submitted by

Kirsten J. Mayer

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2022

Doctoral Committee:

Advisor: Elizabeth A. Barnes

James W. Hurrell
Eric D. Maloney
Charles Anderson

ABSTRACT

APPLICATION OF NEURAL NETWORKS TO SUBSEASONAL TO SEASONAL PREDICTABILITY IN PRESENT AND FUTURE CLIMATES

The Earth system is known for its lack of predictability on subseasonal to seasonal timescales (S2S; 2 weeks to a season). Yet accurate predictions on these timescales provide crucial, actionable lead times for agriculture, energy, and water management sectors. Fortunately, specific Earth system states – deemed *forecasts of opportunity* – can be leveraged to improve prediction skill. Our current understanding of these opportunities are rooted in our knowledge of the historical climate. Depending on societal actions, the future climate could vary drastically, and these possible futures could lead to varying changes to S2S predictability. In recent years, neural networks have been successfully applied to weather and climate prediction. With the rapid development of neural network explainability techniques, the application of neural networks now provides an opportunity to further understand our climate system as well. The research presented here demonstrates the utility of explainable neural networks for S2S prediction and predictability changes under future climates.

The first study presents a novel approach for identifying forecasts of opportunity in observations using neural network confidence. It further demonstrates that neural networks can be used to gain physical insight into predictability, through neural network explainability techniques. We then employ this methodology to explore S2S predictability differences in two future scenarios: under anthropogenic climate change and stratospheric aerosol injection (SAI). In particular, we explore subseasonal predictability and forecasts of opportunity changes under anthropogenic warming compared to a historical climate in the CESM2-LE. We then investigate how future seasonal predictability may differ under SAI compared to a future without SAI deployment in the ARISE-SAI simulations. We find differences in predictability between the historical and future climates and the two future scenarios, respectively, where the largest differences in skill generally occur during forecasts of opportunity. This demonstrates that the forecast of opportunity approach, presented in the first study, is useful for identifying differences in future S2S predictability that may not have been identified if examining predictability across all predictions. Overall, these results demonstrate that neural networks are useful tools for exploring subseasonal to seasonal predictability, its sources, and future changes.

ACKNOWLEDGEMENTS

The National Science Foundation Graduate Research Fellowship (grant 006784) funded Kirsten J. Mayer for the entirety of the research presented in this dissertation. This work was also funded by the National Science Foundation Harnessing the Data Revolution (grant 1934668; Chapter 2), the Regional and Global Model Analysis program area of the Department of Energy's Office of Biological and Environmental Research as part of the Program for Climate Model Diagnosis and Intercomparison Project (Chapter 3), and by the Defense Advanced Research Projects Agency HR00112290071 (Chapter 4).

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: Introduction

In the past few decades, there has been a community effort to bridge the gap between weather and climate prediction [1–4], particularly focusing on subseasonal to seasonal timescales (S2S; 2 weeks to a season). These timescales are often referred to as a "predictability desert", as generally neither the atmospheric initial conditions nor the slower varying oceanic states provide ample information for skillful predictions [2,5,6]. However, accurate predictions on these timescales can provide crucial anticipatory information for public and private sectors including the agriculture, energy, and water management sectors [1,4].

One approach to improve S2S prediction skill is to leverage earth system states that are known to provide enhanced predictability, referred to as "forecasts of opportunity" [2]. These opportunities include the Madden-Julian Oscillation (MJO) [7, 8], the El Niño Southern Oscillation (ENSO) [9–12], stratospheric phenomenon (e.g. sudden stratospheric warming events and the Quasi-Biennial Oscillation) [13, 14], the East Asian summer Monsoon [15, 16], soil moisture [17, 18], and others. Two sources of S2S predictability in particular (MJO and ENSO) are located in the tropics and can influence midlatitude variability and predictability on S2S timescales through tropical-extratropical teleconnections [19–25].

The MJO is composed of an east-west oriented dipole of enhanced and suppressed convection that propagates from the Indian Ocean into the central tropical Pacific over about 20-90 days [26–28]. Through convective heating, the MJO excites quasi-stationary Rossby waves that are then steered by the Pacific subtropical jet [29], and these waves can modulate midlatitude circulation, on S2S timescales [19,20,22,30,31]. Certain phases (i.e. location) of the MJO have been shown to provide a more consistent modulation of midlatitude circulation and subsequently, lead to enhanced midlatitude S2S prediction skill [7].

Another source of S2S predictability comes from ENSO [2], an interannual coupled ocean-atmosphere mode (3-8 years) in the tropical Pacific Ocean. It is typically defined by two phases: La Niña and El Niño, characterized by anomalously cold and warm sea surface temperatures in the eastern tropical Pacific, respectively [32]. ENSO has been shown to alter the MJO and MJO teleconnections through its impact on the tropical basic state and the location and strength of the subtropical jet [23,33–37]. These impacts have been shown to influence atmospheric blocking frequency [23] and teleconnection consistency [35], as well as enhance S2S prediction under certain MJO-ENSO conditions [11]. Additionally, ENSO has its own teleconnections that can influence variability on longer timescales (e.g. Pacific North American pattern) [38–40], ultimately impacting seasonal predictability [9,10].

Previous work has demonstrated the utility of empirical models for S2S prediction [8,41–43] and shown that statistical methods, such as linear inverse models, can identify forecasts of opportunity [44,45]. Recently, neural networks have also been successfully applied to weather and climate prediction [46–50]. Neural networks are useful statistical methods for extracting nonlinear relationships [51]; however, previously the decision making process of the network had been relatively enigmatic. With the development of explainability techniques, or explainable artificial intelligence (XAI) [52], we are now able to extract and visualize what the neural network uses to make its predictions. As a result, XAI provides a means to gauge trust in the network's predictions as well as an opportunity for scientists to further improve our understanding of the climate system [53–58].

Given the success of statistical models for S2S prediction and forecast of opportunity identification, the rapidly growing successful applications of neural networks to the atmospheric sciences, and the recent advances in XAI, it raises the question as to whether neural networks could be used to identify physically meaningful S2S forecasts of opportunity. The first research chapter of this dissertation (Chapter 2) aims to address this question by using neural networks to examine a known opportunistic relationship between the MJO and circulation over the North Atlantic. Through this application, we demonstrate that neural networks can identify subseasonal forecasts of opportunity in observations by using the network's confidence in a prediction, and through an explainability technique, confirm the network is identifying physically relevant regions for enhanced prediction in the North Atlantic. These findings demonstrate the utility of neural networks for forecast of opportunity identification, and therefore, provide a framework for future applications of explainable neural networks to S2S prediction.

Our knowledge of the utility of phenomena like the MJO and ENSO for S2S prediction is rooted in our current understanding of the climate system. However, without extensive mitigation, the climate is projected to continue warming [59], and this can subsequently impact the MJO [60] and ENSO [61] as well as their teleconnections [62–69]. This suggests that the role of phenomena like the MJO and ENSO in S2S predictability may also change in the future. Chapter 3 of this dissertation explores possible subseasonal predictability changes under climate change using a global climate model large ensemble simulation (CESM2-LE). To do so, we use neural networks and the approach presented in Chapter 2 to quantify prediction skill across all predictions and during forecasts of opportunity. Overall, we demonstrate that neural networks are useful for evaluating predictability changes under future climate scenarios. Furthermore, we find that largest changes in future subseasonal predictability occur during forecasts of opportunity and minimal differences in skill are seen when comparing across all predictions. These results further demonstrate the value of the network-based forecast of opportunity approach for subseasonal predictability analyses.

In recent years, various forms of solar radiation modification have been proposed to reduce the impact of anthropogenic climate change [70]. One of the most well studied ways to reflect sunlight back into space, and thereby cool the planet, is through stratospheric aerosol injection (SAI): the injection of sub-micron sized reflective particles into the stratosphere. This method has shown promise for reaching global mean temperature targets in climate models, but it may also have climate impacts beyond surface temperature, such as on tropical precipitation [71, 72]. Given the global importance of tropical precipitation [29, 30], midlatitude S2S variability and predictability could be impacted. Chapter 4 of this dissertation investigates how future seasonal variability and predictability may differ under SAI implementation compared to a climate scenario without SAI, using the Assessing Responses and Impacts of Solar climate intervention on the Earth system under SAI (ARISE-SAI) simulations [72]. We find higher seasonal variability throughout the Northern Hemisphere and differences in ENSO teleconnections to the northwest coast of North America.

Motivated by ENSO teleconnection differences, we apply the framework presented in Chapter 2 to explore predictability changes across network confidence. We find that seasonal predictability over the northwest coast of North America is higher under SAI, again largest at higher confidence values.

Overall, this dissertation demonstrates the utility of neural networks for S2S prediction through a forecast of opportunity lens. Specifically, we present a neural network-based approach for identifying forecasts of opportunity, and then further demonstrate how this approach can be used to identify S2S predictability changes under future climates.

The following chapters of this dissertation are organized into three research chapters and a conclusion chapter. Chapter 2 and 3 of this work are published in Geophysical Research Letters [73, 74] and therefore, have been included in this dissertation without changes. Chapter 4 is in preparation to be submitted for publication soon after the submission of this dissertation. The last chapter provides a summary of the three research chapters outlined above and future directions for research at the intersection of machine learning and S2S prediction.

Chapter 2: Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network

2.1 Introduction

Subseasonal timescales (2 weeks - 2 months) are known for their lack of predictability [75], yet reliable and actionable information on these timescales are required for decision making in many sectors such as public health and water management [1,5]. Over the past decade, there has been a substantial research effort to improve prediction on these timescales [5,6,76,77]. 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 [2]. 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) [26,27], a propagating tropical convective phenomenon, is active, its convective heating can lead to the excitation of quasi-stationary Rossby waves [29] that subsequently modulate the midlatitude circulation over the first few weeks following MJO activity [19,20,22,30,31]. 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 [7]. 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 [51], 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 [78] and surface temperature [79] as well as yearly prediction of the El Niño Southern Oscillation [47], 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 [19,42]. 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.

2.2 Data and Methods

2.2.1 Data

We use daily mean OLR (1979-2019) from the National Center for Atmospheric Research/National Oceanic and Atmospheric Administration (NCAR/NOAA) [80] and daily mean z500 (1979-2019) from the European Centre for Medium-Range Weather Forecasts (ECMWF) Interim reanalysis (ERA-I) [81]. MJO teleconnections tend to be stronger during boreal winter [82], 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 A.2 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 (FFT) high-pass filter is then applied to the z500 anomalies to remove seasonal oscillations (frequencies smaller than $\frac{1}{120days}$) 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 (see 2.2.1), 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. Initially, the FFT analysis is not applied to the full dataset to ensure the network has no information about the testing data during training. 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.

2.2.2 Methods

Artificial Neural Network Architecture

A two layer ANN (Figure 2.1) 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 2.1) 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 [19, 37, 42]. In addition, we find that this grid point is representative of a larger area within the North Atlantic (see supplemental Figure A.1).

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, where the predicted category is defined by a likelihood greater than 0.5. 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 rectified linear and 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 A.2). 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.

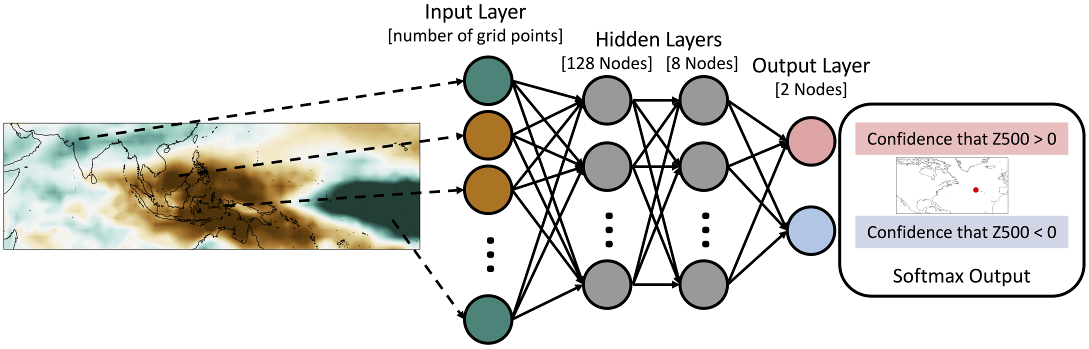


Figure 2.1: Artificial neural network architecture for prediction of the sign of z500 anomalies over the North Atlantic 22 days following tropical OLR anomalies. The neural network consists of two hidden layers of 128 and 8 nodes, respectively, and an output layer of two nodes (one node for each sign). The output layer uses the softmax activation function.

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) [83,84]) 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 2.3.2.

2.3 Results

2.3.1 Identifying Forecasts of Opportunity

ANNs with the architecture shown in Figure 2.1 are trained 100 times with random initialized weights to predict the sign of the z500 anomalies 22 days following the tropical OLR anomalies. Figure 2.2a 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 2.2a, 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.

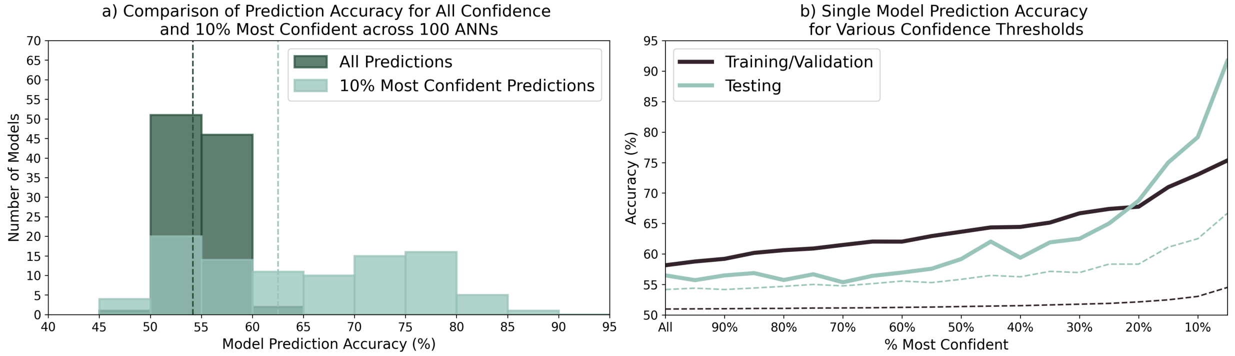


Figure 2.2: (a) Histograms of testing prediction accuracy for 100 trained ANNs. The dark teal represents the histogram of all prediction accuracies and the light teal represents the histogram for the 10% most confident prediction accuracies. The dark teal and light teal dashed lines in (a) are the maximum 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 2.2a to further understand how accuracy varies when a different percent model confidence is used (Figure 2.2b). The solid lines represent the accuracy across various model confidence values for training and validation (black) and testing (light teal) data sets. Figure 2.2b 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 2.2a 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 A.2, and additional skill metrics are provided in Figure A.3 and Table A.4.

2.3.2 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 2.3c-h shows the regions the network found relevant, on average, to make confident and correct positive (Figure 2.3c,e,g) and negative (Figure 2.3d,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 2.3c) 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 2.3d) 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 [85]. 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 [7, 19, 23, 42]. In addition, this dipole structure is known to lead to higher pattern consistency of teleconnections in the midlatitudes [86], which has been shown to lead to enhanced prediction skill [7]. 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 2.3a,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 [87]. 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.

K-means Clustering of LRP Maps

To further distinguish the relevant regions for the ANN's predictions, k-means clustering [88] (see supplemental material for more information) is applied to the LRP maps (Figure 2.3e-h).

This analysis reveals that the composite LRP maps for each sign (Figure 2.3c,d) actually consist of multiple distinct patterns used by the ANN. For positive sign predictions (Figure 2.3e,g), both clusters have a hot spot located between the central Indian ocean and the maritime continent, which are associated with negative OLR anomalies. While not highlighted by LRP in cluster 2 (Figure 2.3g), 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 2.3e) highlights both the positive and negative OLR anomalies. As previously mentioned, these regions lead to more consistent midlatitude teleconnections [7] and have been shown to be associated with a positive NAO anomaly [42], 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 2.3g) 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 [89]. This relationship is nicely illustrated in Figure 2.3a 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 2.3f), 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 2.3b 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 2.3a 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 [87]. This modulation has been hypothesized to play a role in MJO initialization [87, 90]. 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 2.3h) 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 [31]. 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 [29, 30]. 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 2.3e), 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.

2.4 Conclusions

Improving subseasonal prediction accuracy and understanding requires identifying opportunities that can lead to enhanced predictability [2]. 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.

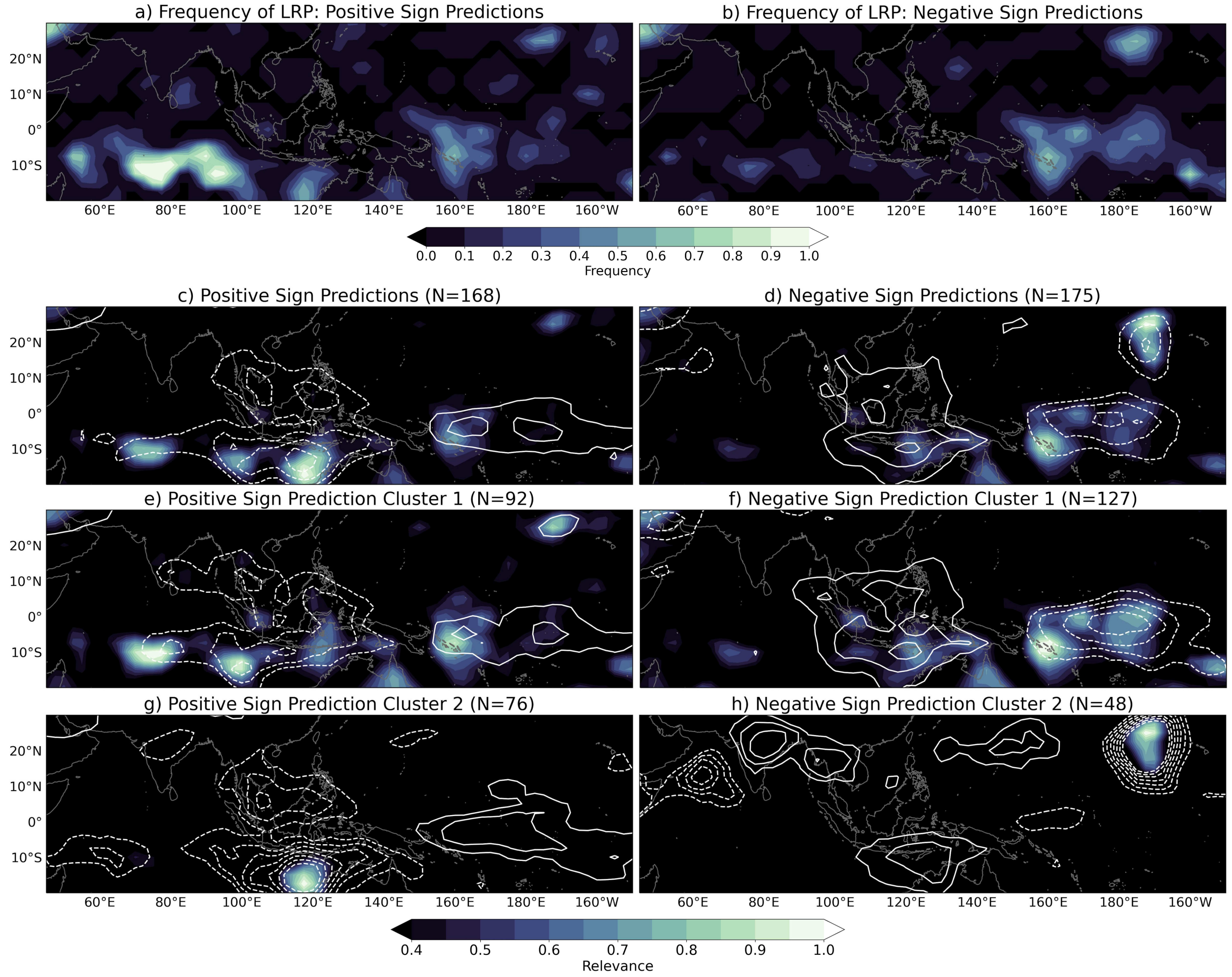

a) Frequency of LRP: Positive Sign Predictions
b) Frequency of LRP: Negative Sign Predictions
20°N
10°N
0°
10°S
60°E
80°E
100°E
120°E
140°E
160°E
180°E
160°W
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Frequency
c) Positive Sign Predictions (N=168)
d) Negative Sign Predictions (N=175)
e) Positive Sign Prediction Cluster 1 (N=92)
f) Negative Sign Prediction Cluster 1 (N=127)
g) Positive Sign Prediction Cluster 2 (N=76)
h) Negative Sign Prediction Cluster 2 (N=48)
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Relevance

Figure 2.3: (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$\frac{W}{m^2}$.

Chapter 3: Quantifying the Effect of Climate Change on Midlatitude Subseasonal Prediction Skill provided by the Tropics

3.1 Introduction

Accurate predictions on subseasonal timescales (2 weeks - 2 months) are important for many public and private sectors such as water management and agriculture [4]. This is because prediction on these timescales provides pivotal lead times for saving lives and property in these sectors [4]. The tropics is of particular importance for this timescale because of intraseasonal phenomena like the Madden-Julian Oscillation (MJO) [26,27]. Quasi-stationary Rossby waves generated by upper level divergence associated with MJO convection [29] can modulate midlatitude circulation in the following weeks [19, 20, 22, 30, 31] and these tropical-extratropical teleconnections are known to lead to enhanced midlatitude prediction skill on subseasonal lead times [7]. Phenomena like the El Niño Southern Oscillation (ENSO), an interannual oceanic mode in the tropical Pacific Ocean [32], can also impact subseasonal prediction. It can do so through modulation of the MJO [33, 95, 96] or modulation of the large-scale background state [36, 97, 98], and both can ultimately impact teleconnection propagation [21, 23, 24, 99] and subseasonal prediction skill [11, 12]. Therefore, when phenomena like the MJO and ENSO are present, they can provide a predictable signal above climate noise and be used to enhance subseasonal prediction skill, known as forecasts of opportunity [2].

The current understanding of the importance of the tropics on midlatitude subseasonal predictability is rooted in our knowledge of the historical climate. However, with the climate continuously warming, it is unclear how transferable this knowledge will be to a future, warmer climate. Therefore, research on subseasonal timescales has examined how the MJO [60] and ENSO [61] will change in the future as well as the subsequent changes to their teleconnections [62–69]. It stands to reason that these changes will likely impact subseasonal predictability across the Northern Hemisphere, but little work has been done in this area [100]. Here, we utilize the Community

Earth System Model Version 2 - Large Ensemble (CESM2-LE) [101] and simple artificial neural networks to identify changes in subseasonal predictability provided by the tropics under future warming.

In recent years, neural networks have been successfully applied to weather and climate prediction [46–49, 55, 56, 58] due to their ability to extract nonlinear relationships from large amounts of data. This makes them advantageous for learning nonlinear relationships in the climate system. In addition, recent advances in explainability techniques and their application to climate sciences demonstrate than neural networks can identify physical relationships in the Earth system [57, 73, 79, 91]. For example, Mayer and Barnes (2021) demonstrate that neural networks can be used to identify subseasonal forecasts of opportunity through the neural network's confidence in a given prediction. They further show that the network identifies physically meaningful sources of subseasonal predictability for the North Atlantic.

Here we use artificial neural networks to quantify how subseasonal prediction skill provided by the tropics may change under future climate warming. Given the importance of forecasts of opportunity for subseasonal prediction in the current climate, we examine both total changes to overall prediction skill as well as changes to skill during forecasts of opportunity, in particular. The artificial neural networks identify subseasonal prediction skill changes across the Northern Hemisphere in the CESM2-LE. In particular, there is an increase in prediction skill over the North Atlantic and western North America as well as a decrease over the North Pacific. In addition, this approach shows that the greatest changes in skill occur during forecasts of opportunity and that these changes appear linked to changes in seasonal variability in the CESM2-LE.

3.2 Data and Methods

3.2.1 Data

Here, we examine midlatitude subseasonal prediction skill changes using the first 10 members from the coupled Community Earth System Model Version 2 - Large Ensemble (CESM2-LE) [101]. CESM2 has both a well represented MJO [102] and MJO teleconnections [103] and thus, is ideal for this analysis. We note that the results presented are specifically for the CESM2-LE, and as with all model-based results, are dependent on specific model biases (e.g. SST biases) [104]. We use the years 1970-2015 as our 'historical period' to represent a climate similar to today and compare it to the latter half of the century (2055-2100; 'future period') under the SSP3-7.0 climate change scenario. We find that 10 members are sufficient for this analysis as the network skill plateaus when at least 5 ensemble members are used for training, depending on location and time period (Figure B.1; Text B.2). While additional ensemble members could be used, we believe our conclusions would remain unaffected, as the sign of the change in prediction skill of the 20% most confident predictions remains consistent regardless of the number of members examined here.

The CESM2-LE members #1-10 are split into training (members #1-8), validation (member #9) and testing data (member #10). To simultaneously detrend and remove the seasonal cycle for each grid point, the 3rd order polynomial fit of the training and validation members' ensemble mean is subtracted from every ensemble member individually for each day of the year. We find the conclusions are insensitive to the specific members assigned to training, validation and testing (Figure B.2).

We utilize the CESM2-LE tropical precipitation (28.5°S-28.5°N) and geopotential height at 500 hPa (z500; 31.25°N-88.75°N) during the extended boreal winter (November-March) since this is when MJO teleconnections tend to be strongest (Madden 1986). Tropical precipitation anomalies are computed for each member and grid point by standardizing with the training data mean and standard deviation. For computational purposes, the z500 field is partitioned into non-overlapping 5° x 5° boxes, where the average of these values is assigned to the center grid point latitude and longitude. This decreases the z500 resolution from 2.5° x 2.5° to 7.5° x 7.5°, however, given the large scale structure of z500, we do not expect the resolution reduction to impact the conclusions. The sign of the z500 anomalies are defined by subtracting the training data median from the training, validation and testing data and converting the anomalies into 0s and 1s depending on the sign (negative and positive, respectively).

Sea surface temperatures (SST) from the first 10 members of the CESM2-LE are also used to calculate the Niño 3.4 index for each member, following the NCAR Climate Data Guide [105]. The trend and seasonal cycle is removed simultaneously as aforementioned, and a 5 month running mean is applied prior to standardizing the SSTs with each member's mean and standard deviation. An El Niño/La Niña event is therefore defined as a standardized Niño 3.4 index value of greater/less than +/- 1σ. We use this index to examine any possible role that ENSO may play in the identified changes to subseasonal predictability.

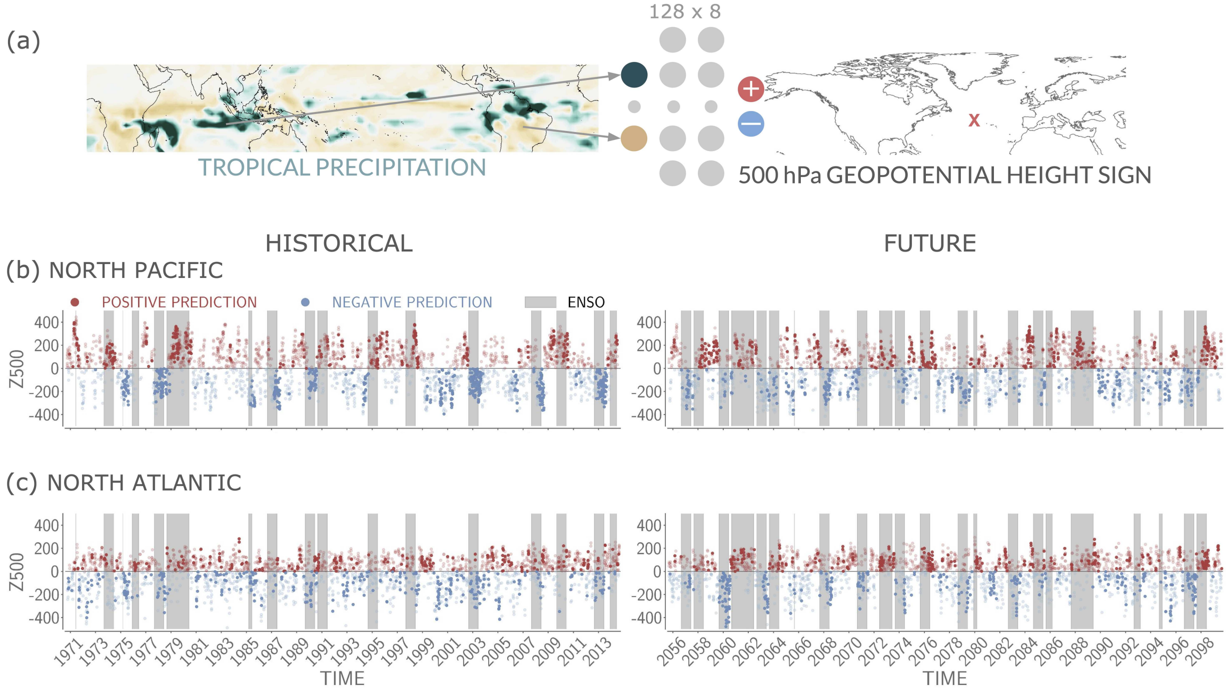


Figure 3.1: (a) The artificial neural network input (tropical precipitation), architecture (first hidden layer: 128 nodes, second hidden layer: 8 nodes) and output (sign of z500hPa at a location 'x'). (b,c) Timeseries of the *correct* sign predictions of z500 in ensemble member #10 for the historical (left column) and future (right column) for (b) the North Pacific and (c) the North Atlantic. Red (blue) dots indicate positive (negative) predictions. Darker dots denote the 20% most confident predictions, and the grey shading indicates when the standardized Niño 3.4 index exceeds +/-1σ.

3.2.2 Neural Network Architecture and Application

The neural network ingests daily tropical precipitation anomalies and makes a prediction of the sign of z500 at a given grid point at a lead of 21 days (Week 3; Figure 3.1a). Prediction of the sign of z500 at each grid point allows the network freedom to learn important patterns and relationships between tropical precipitation and z500. The number of input nodes is equal to the number of precipitation grid points (N=3456). The first and second layer of the network consist of 128 and 8 nodes, respectively. A softmax activation function is applied to the output layer of 2 nodes which transforms the network output into values which sum to one. These transformed values represent a network estimation of likelihood, which we refer to as 'model confidence', where the predicted category is defined as a value greater than 0.5. As shown in Mayer and Barnes (2021), when prediction skill increases with model confidence, higher model confidence can be used to identify subseasonal forecasts of opportunity.

We use this network architecture because it has some of the highest validation skill for both the historical and the future time periods in the North Atlantic and also performs well in the North Pacific (Figure B.3-B.4). We note that slight variations of the hyperparameters (i.e. network depth, nodes per layer, learning rate, ridge regression parameter) show similar skill. While one could optimize the architecture and hyperparameters for every gridpoint individually, we have not done this due to the considerable computational resources necessary and find it unlikely to lead to substantially different conclusions. For additional information on the network architecture and hyperparameters see Text B.3.

Example correct network predictions for the testing ensemble member #10 are shown in Figure 3.1(b-c) for the historical (left column) and the future (right column) periods in (b) the North Pacific and (c) the North Atlantic. The color denotes the sign of the prediction and the darker colors denote the (20% most) confident predictions. The vertical grey shading indicates periods of ENSO events. Figure 3.1(b-c) demonstrates that the networks can accurately and confidently predict both sign anomalies. In addition, it shows a possible relationship between confident subseasonal predictions and ENSO events, but the amount which confident predictions coincide with ENSO events depends on location and time period. This relationship will be addressed further in section 3.3.2.

3.3 Results

3.3.1 Changes in Subseasonal Prediction Skill

To examine how subseasonal prediction skill provided by tropical-extratropical teleconnections changes in a warmer climate, 100 networks are trained for the North Pacific (41.25°N, 205°E) and the North Atlantic (41.25°N, 325°E) for both the historical and future periods. These two locations are chosen because they encompass regions known to be significantly impacted by the MJO [42,87,106] and ENSO [107,108] teleconnections, which subsequently have North American and European impacts. The 100 networks are created by varying their random seed to test the sensitivity of the network to the random initialized weights.

Accuracies binned by various model confidence thresholds are shown in Figure 3.2. Accuracy increases with model confidence (moving from left to right), suggesting the network is identifying forecasts of opportunity for these regions. In addition, we find that all networks at almost every confidence level perform better than random chance (Figure 3.2, Text B.4). The North Pacific (Figure 3.2a) has higher accuracy compared to the North Atlantic (Figure 3.2c), likely due to the strong influence of tropical phenomena like the MJO and ENSO in modulating the circulation in the North Pacific [37, 106–109]. In the future, subseasonal prediction skill increases in the North Atlantic (Figure 3.2c) and decreases in the North Pacific (Figure 3.2a) in the CESM2-LE, and this is most evident at higher confidence values. If one examines the accuracy for all (100% most confident) predictions, the North Atlantic and North Pacific accuracies exhibit almost no difference between the two time periods. It is when we focus on the higher confidence predictions that a clear signal emerges. In other words, the changes in subseasonal prediction skill are most evident during forecasts of opportunity in these regions.

Histograms of the accuracies at the 20% most confident threshold (Figure 3.2 b,d) further show that the future period has substantially shifted away from the historical period in both regions. The majority of the future North Atlantic accuracies exceed the 95th percentile of the historical accuracies, and all of the future North Pacific accuracies lie below the 5th percentile of the historical accuracies.

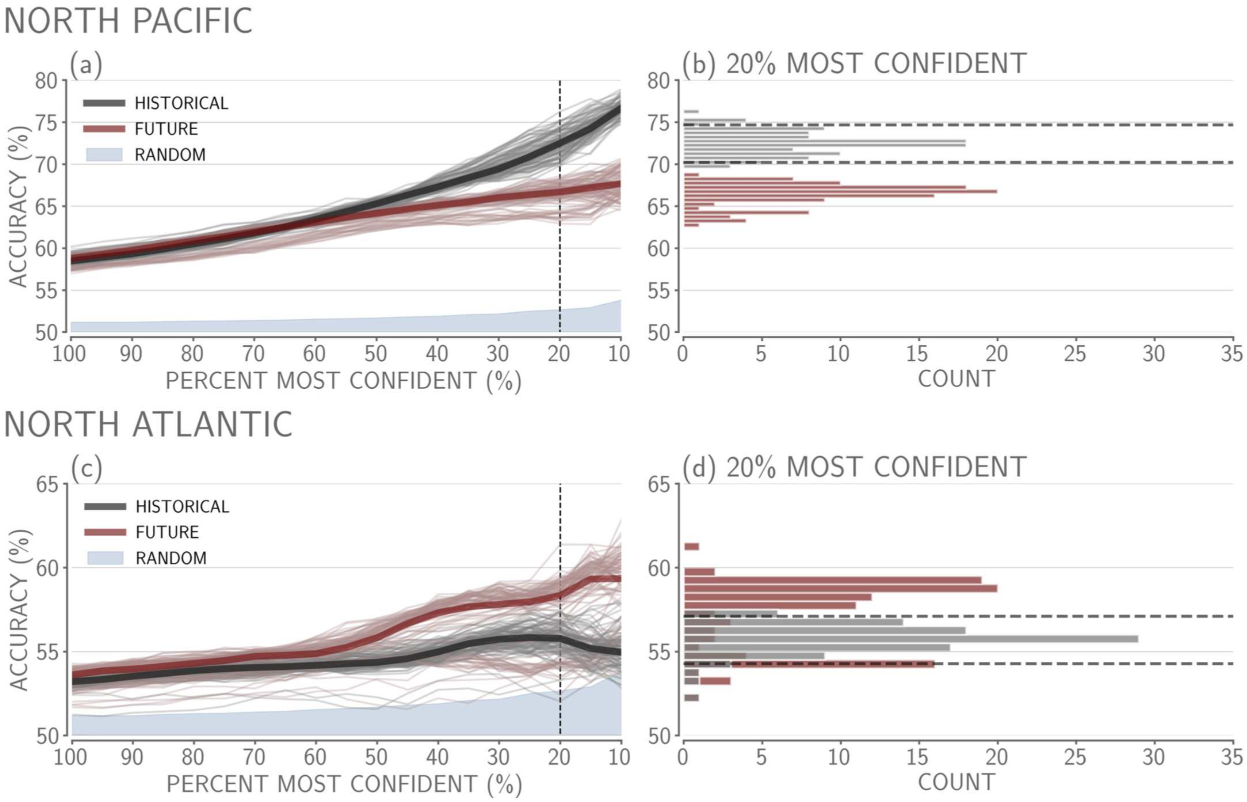


Figure 3.2: (a,c) Accuracy versus confidence for 100 trained networks in the North Pacific and the North Atlantic from testing member #10. Testing samples are subset so that random chance for all predictions is 50%. Thick grey and red lines denote the median accuracy across the 100 networks at each confidence threshold. Vertical black dashed lines indicate the 20% most confident predictions. (b,d) Histograms of the 100 accuracies at the 20% most confident threshold, using a bin size of 0.5%. Horizontal grey dashed lines indicate the 5th and 95th percentile bounds of the historical accuracies at the 20% most confident level.

To explore whether the results in Figure 3.2 hold for other regions, we train 10 neural networks for each grid point and time period across the Northern Hemisphere. We train 10 networks instead of 100 for computational efficiency. To test whether these changes in skill in the North Atlantic and North Pacific could be seen with only 10 networks, we conducted a bootstrapping analysis (Text B.5; Figure B.5) following the method used to create Figure 3.3, and find that 10 networks are sufficient for identifying these changes. Figure 3.3 shows the resulting mean testing accuracy of the top three of the 10 networks for each location. The top three networks are defined as the networks with the three highest 20% most confident validation accuracies. We use the top three networks so that the mean accuracies for each region are not as influenced by models that learn very little or not at all.

For all predictions (Figure 3.3a-b) and 20% most confident predictions (“confident predictions" from here on; Figure 3.3d-e), the locations of highest skill are in regions associated with the Pacific/North America pattern (PNA) [107]. The higher accuracies over PNA regions suggests the network is most likely identifying forecasts of opportunity associated with teleconnections from the MJO and/or ENSO [37, 106–109]. We also find that the spatial coherence in accuracies across networks corresponds to the networks correctly predicting many of the same days for neighboring grid points (not shown). In the future period (Figure 3.3b,e), there is an additional region of higher accuracies spread across Asia and the North Atlantic. Overall, the confident predictions have higher accuracies than all predictions, indicating that higher model confidence predictions exhibit greater skill.

In the future, spatially coherent increases in skill are seen across Asia, along the west coast of North America, across the southern United States and throughout the North Atlantic (Figure 3.3c,f) while decreases are seen over the North Pacific, Canada and western Europe. While the change in skill over East Asia is substantial, it appears that the overall skill in East Asia for both time periods does not harness any subseasonal variability, but rather comes about exclusively from seasonal variability or longer timescales (Figure B.8-B.9). As a result, these changes in skill are not addressed further here. The difference plots for both all and the confident predictions (Figure 3.3c,f) have similar spatial patterns of changes in accuracy, however, the confident predictions show the largest changes in skill. Specifically, the absolute maximum change in skill for all predictions is about 5% while the absolute maximum change in skill for confident predictions is about 10%. This further demonstrates that the greatest changes to subseasonal prediction skill provided by the tropics occur during forecasts of opportunity across the Northern Hemisphere, consistent with Figure 3.2.

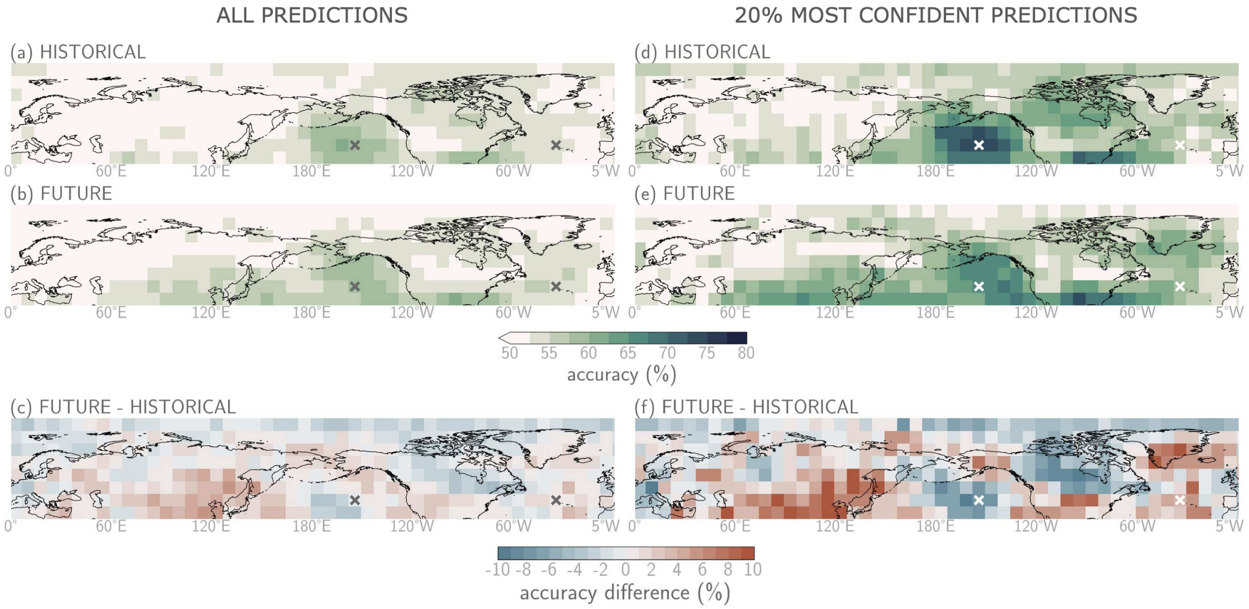


Figure 3.3: (a,b,d,e) Mean testing accuracy of the best 3 models for (a,b) all and (b,e) the 20% most confident predictions. (c,f) Difference in accuracy between the future and the historical time periods for (c) all and (f) the 20% most confident predictions, where red (blue) indicates an increase (decrease) in accuracy in the future. The grey and white ‘x’ indicate the North Pacific and North Atlantic regions (from left to right) used in Figures 3.1, 3.2.

3.3.2 Tropical Drivers of Changing Midlatitude Skill

Seasonal variability can have a large influence on subseasonal variability and prediction skill. In the tropics, ENSO can modulate the MJO [33, 95, 96] and the basic state [36, 97, 98], and ENSO teleconnections can (de)constructively interfere with MJO teleconnections [21, 23, 24, 99, 110]. Recent studies have identified possible changes to both MJO and ENSO variability [60, 61] as well as their teleconnections [62, 66, 67, 111] under future climate warming. Thus, the changes in midlatitude subseasonal prediction skill seen in Figures 3.2 and 3.3 could be a reflection of changes to subseasonal variability, seasonal variability, or through a combination of changes to both.

We find that the increase in skill along the west coast of North America and in the North Atlantic is supported by previous research on MJO and ENSO teleconnections in a warmer climate. In particular, the subseasonal skill increase along the west coast of North America (Figure 3.3f) appears to be associated with a north-eastward shift of higher accuracies over the North Pacific in the future (Figure 3.3d-e). This is consistent with research showing that PNA patterns initiated by ENSO [62, 68, 69, 111–114] and the MJO [67, 103, 115, 116] are projected to shift eastward in a warmer climate in a variety of climate models, including CESM2 [103, 111]. In the North Atlantic, increased skill is also consistent with research suggesting that the North Atlantic may become more sensitive to MJO teleconnections [66] and that the ENSO-NAO teleconnection may strengthen [64, 65] in the future. The decrease in skill over the North Pacific is also consistent with recent research using a variety of CMIP6 models that suggests the ENSO teleconnection amplitude over the North Pacific may weaken in a warmer climate [62, 111].

To gain insight into the neural network's identified sources of predictability, we apply explainable AI to create heatmaps of the relevant regions of the input tropical precipitation the network uses to make confident and correct predictions (see Text B.6) [83, 84]. In the North Pacific and North Atlantic, the network tends to focus on the tropical equatorial Pacific, typically associated with ENSO (Figure B.6). In the North Pacific, the future decrease in skill is associated with a decrease in relevance of the ENSO region (Figure B.6a-d). For the North Atlantic, the future increase in skill is associated with an increase in relevance of the ENSO region (Figure B.6e-h). These explainability results suggest that the changes in subseasonal prediction skill may be related to changes in the importance of the ENSO region (i.e. seasonal variability), even though both subseasonal and seasonal variability are contributing to the total skill (Figure B.8-B.9). This changing role of ENSO in both regions is also evident in the prediction timeseries in Figure 3.1. In the North Atlantic (Figure 3.1c), the confident predictions in the historical period are scattered throughout the years, whereas in the future period, the confident predictions correspond more frequently with ENSO events (darker dots mainly occur in the grey shading). The opposite is seen for the North Pacific (Figure 3.1b). Given the results of this analysis, we next examine if the changes in midlatitude subseasonal prediction skill are related to changes in ENSO teleconnections.

We analyze the relationship between ENSO teleconnections and subseasonal prediction skill changes across the Northern Hemisphere by calculating how often a positive z500 anomaly occurs 21 days following an El Niño/La Niña event (Figure 3.4). This metric quantifies the consistency of specific teleconnections following ENSO events and thus, demonstrates the downstream influence

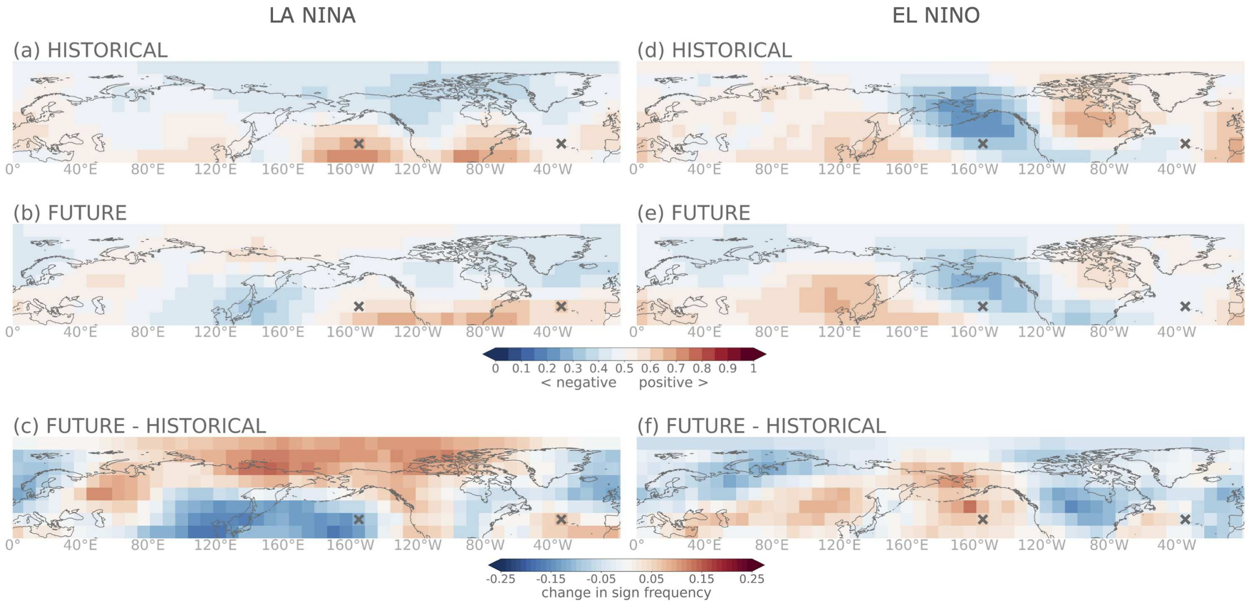


Figure 3.4: (a,b,d,e) Frequency of a positive sign anomaly 21 days following a standardized Niño 3.4 Index value of greater/less than +/- 1σ. Values greater (less) than 0.5 frequency indicate that positive (negative) sign anomalies are more frequent. (c,f) Difference in frequency between the future and historical time period. The left (right) column is for La Niña (El Niño). The grey 'x' indicate the North Pacific and North Atlantic regions (from left to right) used in Figures 3.1, 3.2.

of ENSO on specific regions. Therefore, any regional changes to the consistency between the two time periods implies changes to the impact of ENSO in that region. Over the North Pacific, the consistency of the z500 sign following both ENSO phases decreases (Figure 3.4c,f), suggestive of a reduction in the influence of ENSO teleconnections. Furthermore, the large decrease in skill over Canada (Figure 3.3f) aligns with the decrease in El Niño teleconnection consistency in the future (Figure 3.4f). Over the North Atlantic, there is a slight increase in ENSO teleconnection consistency which may be related to the projected strengthening of the ENSO-NAO teleconnection in the future [64, 65]. Lastly, the increase in skill along the west coast of North America (Figure 3.3f) aligns with an increase in consistency of La Niña teleconnections (Figure 3.4c). Thus, we hypothesize that the substantial changes in subseasonal prediction skill in regions across the Northern Hemisphere are connected to changes in ENSO teleconnections in the CESM2-LE.

We provide further evidence of the role of seasonal variability in changes to subseasonal prediction skill through an additional neural network analysis in the North Pacific and North Atlantic.

We filter out 60+ day variability from the z500 anomalies (Text B.7) to remove low-frequency signals such as those from ENSO teleconnections. With this filtering, there is almost no change in skill between the historical and future period in the North Pacific (Figure B.7c-d). This demonstrates that the decrease in skill in this region is mainly a result of changes to seasonal variability. In the North Atlantic, the increase in skill is still seen, but to a reduced degree when the lower frequencies are removed (Figure B.7e-f). This suggests that seasonal variability is playing a role in subseasonal prediction skill changes in this region, however, there is also likely a contribution from subseasonal variability to these changes. This is consistent with research that suggests the North Atlantic may become more sensitive to MJO teleconnections in the future [66].

The influence of seasonal variability on subseasonal prediction skill changes can be further examined in the North Pacific and North Atlantic by training the neural networks to instead predict the sign of unfiltered z500 anomalies on seasonal lead times. In the North Pacific, we find that *changes* in skill at 60 and 90 day leads are similar to that for a lead of 21 days. This again implies that the changes in subseasonal prediction skill seen in the North Pacific are due to changes in seasonal variability. In the North Atlantic, the change in skill for the seasonal lead time is larger than the 21 day lead time. This difference in the change suggests that the network is focusing on different sources of predictability for the 21 day prediction compared to the 60 or 90 day predictions, implying again that the change in skill in the North Atlantic is not purely due to seasonal variability changes in the future (Figure B.8-B.9).

3.4 Conclusions

While accurate subseasonal predictions are important for society [4], this timescale is known to exhibit limited predictability [6]. One method to improve prediction skill on subseasonal timescales is to utilize Earth system states which are known to provide enhanced subseasonal predictability when they are present (forecasts of opportunity) [2]. Previous research has examined how specific Earth system states important for subseasonal prediction (e.g. MJO and ENSO) and their teleconnections may change in a warmer climate [60, 61, 103]. To address whether these projected changes ultimately impact subseasonal predictability, we use the CESM2-LE and simple artificial neural networks to quantify and understand how subseasonal predictability provided by the tropics may change in a warmer climate. We find that there are changes to subseasonal prediction skill across the Northern Hemisphere and the largest differences in skill mainly occur during forecasts of opportunity.

Our results are supported by recent research on changes to MJO and ENSO teleconnections. In particular, the increase in skill along the west coast of North America is consistent with the projected eastward shift of MJO and ENSO teleconnections in the future [62, 103, 111, 116]. In addition, our results suggest there is a contribution from both subseasonal and seasonal variability changes to the increase in prediction skill in the North Atlantic. This is consistent with research suggesting the North Atlantic becomes more sensitive to the MJO [66] and ENSO [64, 65] in the future. We also identify a substantial decrease in skill over the North Pacific and from our analysis, hypothesize that this decrease is mainly driven by a reduced influence of ENSO teleconnections to this region in the future. Overall, while both MJO and ENSO teleconnections are projected to change in the future, our analysis demonstrates that changes to ENSO and its teleconnections (e.g. seasonal variability) at least partially explain substantial changes in subseasonal prediction skill across the North Hemisphere in the CESM2-LE. Changes to subseasonal variability may still play a role in changes to subseasonal prediction skill in certain locations (e.g. North Atlantic), but further work is needed to understand and quantify its contribution. In addition, we only explored the Niño 3.4 index as a metric for ENSO variability, and future work could further extend this to other metrics that capture ENSO dynamics and that may also account for possible changes in ENSO variability under climate change.

Using the CESM2-LE, we show that neural networks are a useful tool for identifying and understanding future changes in predictability. In addition, we find that changes in subseasonal prediction skill across the Northern Hemisphere are often largest during forecasts of opportunity, suggesting that future research on prediction skill changes should focus on periods of enhanced predictability. While this research addresses changes in boreal wintertime subseasonal predictability provided by the tropics, future research should also examine how other seasons and sources of predictability may be affected in a warmer climate. This could include identifying possible changes to the importance of the stratosphere for subseasonal prediction or changes to boreal summer subseasonal predictability due to changes to the importance of the boreal summer intraseasonal oscillation [117]. Furthermore, although this work examines subseasonal predictability changes by the end of the century, examining how quickly these changes may be detected is also worthy of study. Ultimately, this research demonstrates the utility of neural networks to quantify and gain physical insight into changes in subseasonal predictability in future climates.

Chapter 4: Investigating Northern Hemisphere Seasonal Variability and Predictability under ARISE Stratospheric Aerosol Injection

4.1 Introduction

Without extreme mitigation efforts, the climate is projected to continue warming over the coming decades [59]. With minimal progress in climate change mitigation efforts over the past few decades, various forms of solar radiation modification (SRM) have been proposed to cool the planet through the intentional reflection of sunlight back into space [70]. SRM could reduce the impacts of anthropogenic warming and provide society additional time to respond to climate change. The most well studied and feasible SRM method is stratospheric aerosol injection (SAI), which includes the hypothetical injection of sub-micron size reflective particles into the stratosphere [70]. These particles intentionally enhance Earth's albedo, reducing the amount of solar radiation at the surface, and subsequently, minimizing surface warming.

Previous SAI research has focused on topics ranging from long term cost of implementation [123,124] and intervention technology [125–127] to stratospheric chemical and dynamical impacts [128–130] and global temperature and precipitation responses [72,131–133]. Of particular interest here is research that suggests SAI may reduce future tropical precipitation compared to a future scenario without SAI [71,72]. Tropical precipitation is an important source of climate variability due to its ability to modulate the global circulation through atmospheric teleconnections [29,30]. As a result, differences in future tropical precipitation could have global impacts.

The El Niño Southern Oscillation (ENSO), an interannual coupled ocean-atmosphere mode in the east-central tropical Pacific Ocean [32], is a prominent driver of interannual changes to tropical precipitation [134]. Impacts of ENSO on tropical convection can lead to the formation of Rossby waves, which can propagate and impact the circulation world-wide [29,30,38,135]. As a result, ENSO acts as a prominent driver of boreal winter variability [39,40,136] and predictability [9,10,46] in the Northern Hemisphere.

Due to ENSO's global importance, climate change research has examined how ENSO [61] and its teleconnections [62, 64, 65, 68, 69] may change in the future under continued climate warming. In particular, many studies have shown that ENSO teleconnections may shift north-eastward in the North Pacific under climate change [62, 64, 65, 68, 69, 74], impacting seasonal variability along the west coast of North America, and subsequently predictability [74].

These previous results raise the question as to how northern hemisphere seasonal variability and predictability may differ under a future with SAI compared to one without it. Therefore, the purpose of this study is to explore future seasonal variability and predictability under SAI compared to a future scenario with no intervention. To do so, we use the Assessing Responses and Impacts of Solar climate intervention on the Earth system with Stratospheric Aerosol Injection (ARISE-SAI) [72] simulations. In particular, this study focuses on boreal winter mean state and variability differences in surface temperature across the Northern Hemisphere, and from these results, further concentrates on differences in ENSO teleconnections and seasonal predictability along the west coast of North America. We find significantly larger boreal winter variability over North America and more consistent La Niña teleconnections to northwest North America under SAI compared to a scenario with no intervention. We further show that under SAI this northwest region is associated with significantly higher seasonal predictability, possibly related to these differences in La Niña teleconnections.

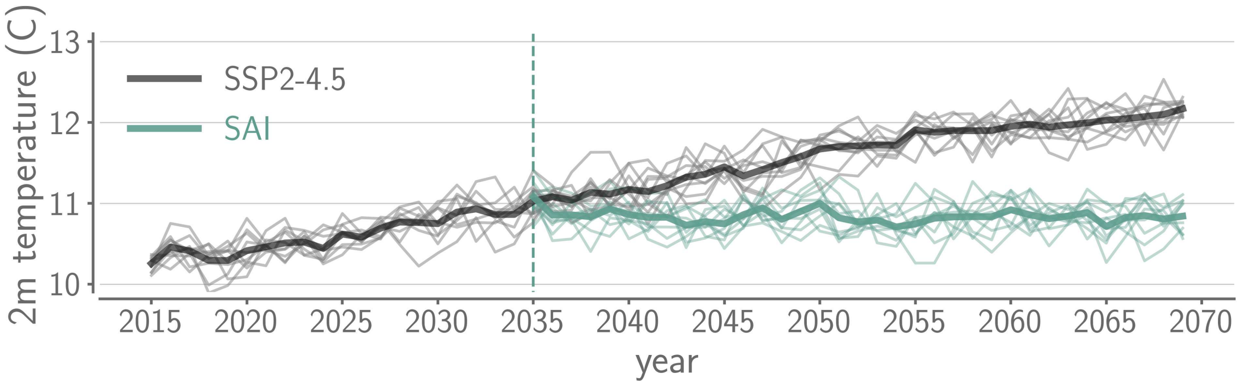


Figure 4.1: Annual global 2m temperature over land. The black (teal) line denotes the ensemble mean for the SSP2-4.5 (SAI) scenario. The grey and light teal lines show the 10 members for the SSP2-4.5 and SAI scenario, respectively. The vertical teal dashed line indicates the year 2035 when SAI is implemented.

4.2 Data and Methods

4.2.1 Data

To evaluate the influence of SAI on future boreal winter variability and predictability of 2 meter (m) temperature compared to a scenario with no intervention, we use the ARISE-SAI simulations [72]. These simulations are conducted with the Community Earth System Model, version 2 with the Whole Atmosphere Community Climate Model, version 6 (CESM2-WACCM6). ARISE-SAI is comprised of two scenarios, each with 10 corresponding ensemble members. One scenario follows the SSP2-4.5 emission scenario (2015-2069) while the other implements SAI beginning in 2035 (hereafter referred to as the SAI scenario). The primary goals of the SAI scenario are to maintain a global mean surface air temperature near 1.5°C above the pre-industrial value as well as maintain the pole-to-pole and pole-to-equator temperature gradients to their baseline values. As a result, the SAI scenario diverges from the SSP2-4.5 scenario in 2035, evident in global mean 2m land temperature (Figure 4.1). For this analysis, the ARISE-SAI simulations are examined over the final two decades (2050-2069) to maximize the difference between the two scenarios. Only the extended boreal winter months (November - March) are analyzed since this is when Northern Hemispheric variability is high and ENSO has a large impact [34, 40, 135, 137].

To remove the seasonal cycle and trend from monthly sea surface temperature (SST) and 2m temperature (over land) at each grid point, a 3rd order polynomial fit to the 2035-2069 ensemble mean of the SSP2-4.5 and SAI scenario is removed from each member. To define an 'ENSO event', we apply a 5 month running mean to the mean SST anomalies in the Nino 3.4 region (5°S - 5°N, 170° - 120°W) and then standardize these anomalies using the mean and standard deviation of the corresponding training members (see Section 4.2.2 for more details). An El Niño/La Niña 'event' is thus defined as a monthly value of greater/less than +/- 1 standard deviation. Further information on preprocessing of SST and 2m temperature relating to the neural network analysis is included in Section 4.2.2.

4.2.2 Quantifying Predictability

Neural networks have become increasingly popular for weather and climate prediction in recent years due to their ability to extract nonlinear [46–48] and physically meaningful relationships within climate data [57, 58, 73, 74, 79]. For the seasonal prediction analysis, maps of monthly tropical sea surface temperatures (SST; 20°S - 20°N) are input into a neural network to predict the sign of monthly 2m temperature averaged over Alaska/Western Canada (55-70°N, 190-250°E) at a 2 month lead. To calculate the impact of SAI on future seasonal predictability compared to the SSP2-4.5 scenario, neural networks are trained on both the SAI and SSP2-4.5 scenarios, separately. SST anomalies are used as a predictor since ENSO is a main source of seasonal variability for the midlatitudes [40, 136]. However, we do not isolate ENSO variability in the SST anomalies to allow the neural network freedom to identify other important tropical SST features that may provide predictability beyond that of ENSO.

Each network is trained with 8 members, validated on 1 member and tested on the remaining member (e.g. training members #1-8, validation member #9, and testing member #10), and cross-validation is implemented so every member is eventually used for testing. In addition, the softmax activation function is applied to the output layer of each network to transform the output into values which sum to one, where the predicted category is defined as a value greater than 0.5. This function is used because these values represent the network's estimation of likelihood or "model confidence". When prediction skill increases with model confidence, higher confidence predictions can be used to identify periods of enhanced predictability, or forecasts of opportunity [73]. Previous research has also shown that future changes in subseasonal predictability are most evident during forecasts of opportunity [74]. Therefore, to identify the impact of SAI on future seasonal prediction skill compared to under SSP2-4.5, we examine the accuracy of the network across network confidence values. The network architectures are determined through a hyperparameter sweep and the final hyperparameters along with additional information on the neural network are included in Text C.2.

To further preprocess the predictor, the training, validation and testing members' tropical SST anomalies are standardized at each grid point using the eight training members' mean and standard deviation. To define the predictand, the training members' median of the regionally averaged 2m temperature is removed from the training data to ensure balanced classes. This training median is also removed from the corresponding validation and testing member, and the validation and testing data is then randomly subset into balanced classes as well. Lastly, the anomalies are converted into labels of 0s and 1s depending on the sign (negative and positive, respectively).

4.3 Results

4.3.1 SAI impact on seasonal variability

The projected reduction of future tropical precipitation under SAI compared to a future without intervention [72, 132] has global implications due to tropical precipitation's ability to modulate the large-scale circulation [29, 30]. In particular, tropical precipitation associated with ENSO can influence Northern Hemisphere winter variability and predictability [40,46,137]. However, previous SAI research has mainly focused on *annual* mean changes, with no research examining changes to seasonal temperature variability in the ARISE-SAI simulations. Therefore, we provide an initial analysis of Northern Hemisphere boreal winter mean state and variability changes in 2m temperature under SAI compared to SSP2-4.5.

The 2m temperature change between 2050-2069 and 2015-2034 for both the SSP2-4.5 and SAI scenarios (Figure 4.2a,b) and their difference (Figure 4.2c) is computed to explore changes to mean boreal winter 2m temperature in the ARISE-SAI simulation. Under SSP2-4.5 (Figure 4.2b), the temperature increases across much of the Northern Hemisphere and largest temperature differences occur at higher latitudes ("Arctic amplification") [138]. On the other hand, under SAI, the temperature change is much smaller with no evident Arctic amplification signature (Figure 4.2a). Overall, 2m temperature increases more under SSP2-4.5 than SAI during boreal winter (Figure 4.2c), similar to the annual global mean response in Figure 4.1.

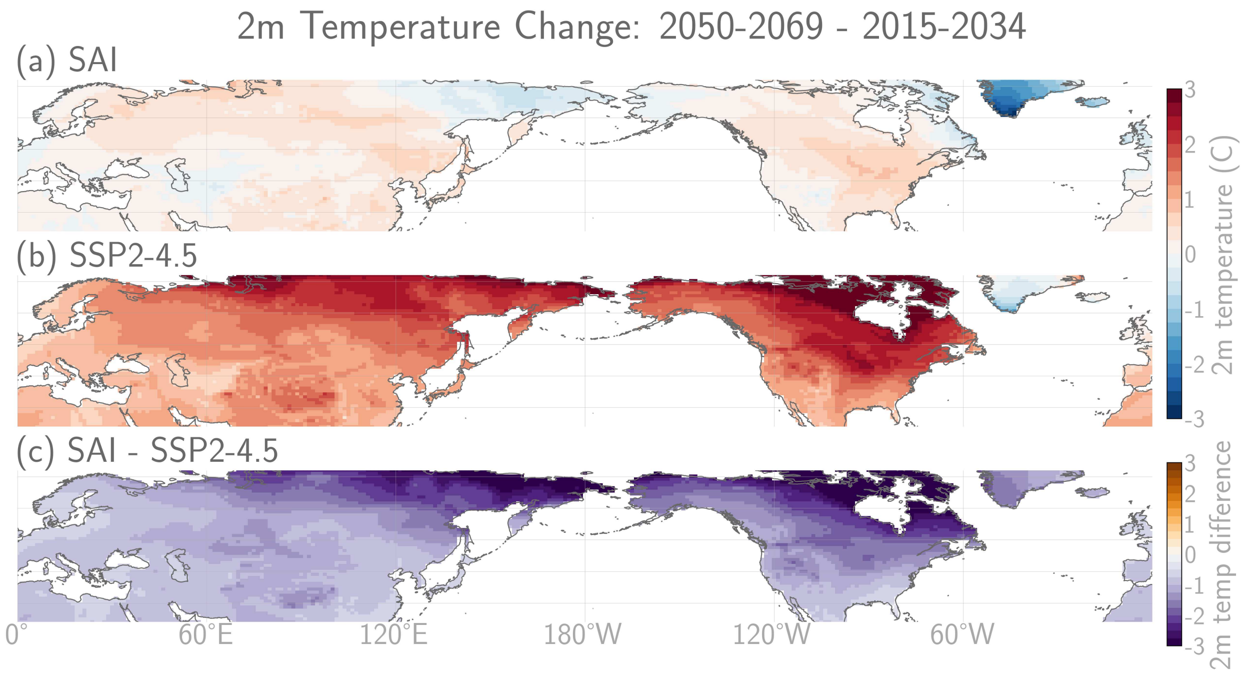


Figure 4.2: The 2m temperature difference (2050-2069 - 2015-2034) during extended boreal winter averaged across ensemble members under (a) SAI, (b) SSP2-4.5 and (c) the difference between the two.

To analyze boreal winter variability, we examine the monthly November through March variance of 2m temperature during the last 2 decades of the ARISE-SAI simulations. While we find similar spatial distributions of variance between the scenarios (Figure 4.3a,b), the 2m temperature variance under SAI is larger across the Northern Hemisphere compared to SSP2-4.5. Specifically, there is significantly higher variance west of China and in northeastern Russia as well as across much of the north and west of North America (Figure 4.3c; Text C.3). Under climate change (without SAI), the effects of a reduced hemispheric meridional temperature gradient on meridional temperature advection is expected to lead to a decrease in midlatitude variability compared to the historical climate [139, 140]. Since a goal of the ARISE-SAI simulations is to maintain intra-hemispheric temperature gradients at their baseline values (Richter et al. 2022), we hypothesize that the higher variability at higher latitudes under SAI compared to SSP2-4.5 is likely a reflection of a reduction in variability under SSP2-4.5 compared to the historical climate. However if this was the only explanation, we may expect to see a more latitudinally uniform change between the two scenarios [139, supplemental Figure 8], when in fact we see more spatial heterogeneity,

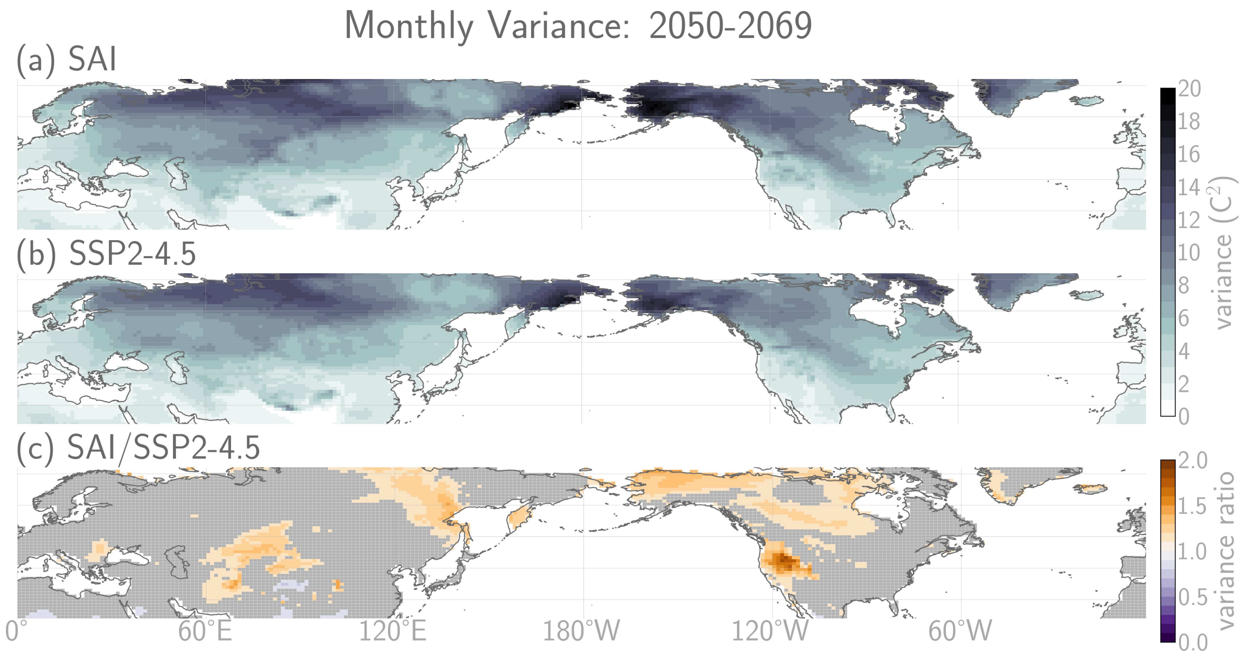


Figure 4.3: The 2m temperature monthly variance of extended boreal winter (2050-2069) averaged across ensemble members for the (a) SAI and (b) SSP2-4.5 scenario and (c) the ratio between the two scenarios. Orange/purple regions indicate locations of statistical significance at 95% confidence (Text C.3).

particularly over the north and west coast of North America. North America is also known to be significantly influenced by ENSO teleconnections on seasonal timescales [34, 39, 135, 137, 141], and therefore, this significant difference in boreal winter variability could be partially explained by differences in ENSO teleconnections. To address this question, we next examine differences to ENSO teleconnections along the west coast of North America.

ENSO teleconnections are quantified by calculating the frequency of a positive sign 2m temperature anomaly 2 months following an ENSO event. The frequency of a sign can be considered as a measure of response consistency. Values greater (less) than 0.5 imply that a positive (negative) temperature anomaly is more frequent in the two months following an ENSO event (Figure 4.4a-b, d-e). This method is used to identify ENSO teleconnections to provide a more direct comparison to the results discussed in Section 4.3.2.

The majority of the northwest coast of North America above (below) 40°N has higher (lower) temperature consistency following ENSO under the SAI scenario compared to the SSP2-4.5 scenario (Figure 4.4c,f). While there is spatial structure in the 2m temperature consistency differ-

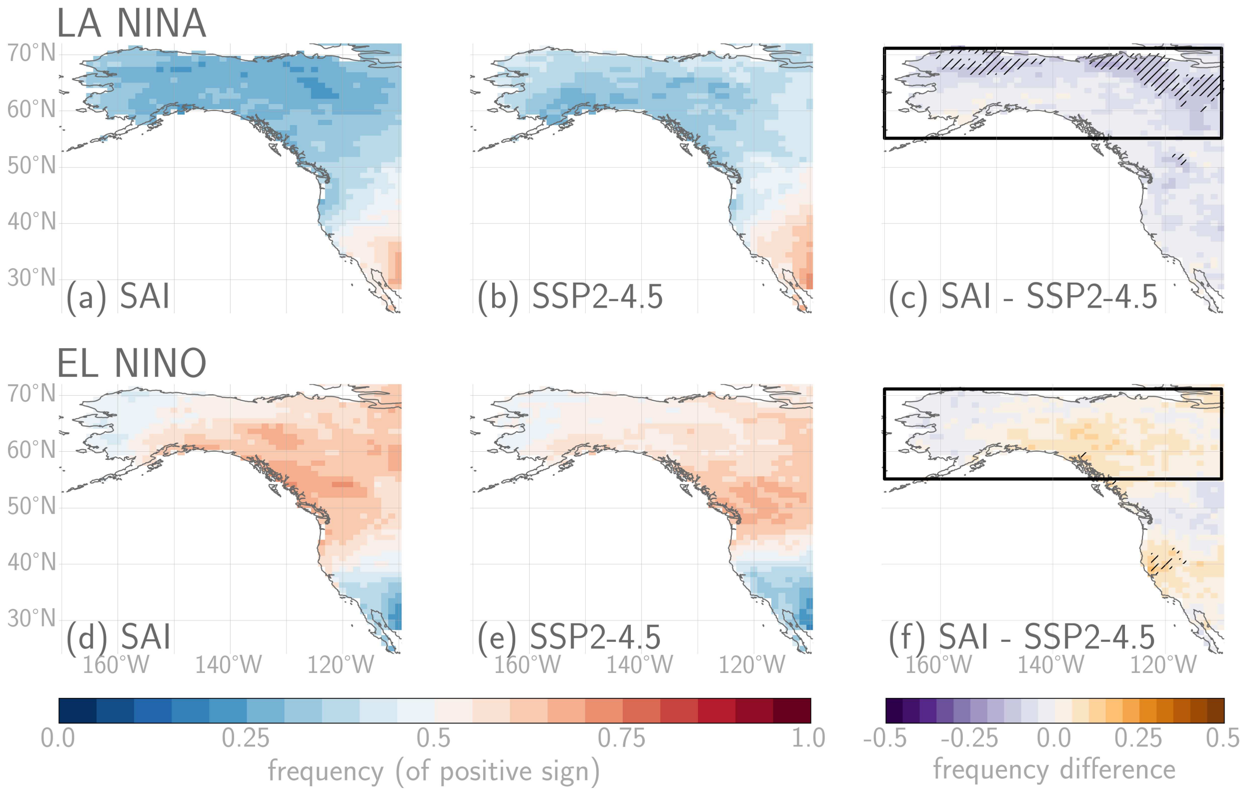


Figure 4.4: The ensemble member mean frequency of a positive sign 2m temperature anomaly 2 months following either (a,b,c) La Niña or (d,e,f) El Niño for (a,d) SAI, (b,e) SSP2-4.5 and (c,f) the difference between the two scenarios over boreal winter (2050-2069). The black box denotes the Alaska/Western Canada region (55-70°N, 190-250°E) used in the predictability analysis. Hatching indicates statistically significant differences at the 95% confidence level (Text C.4).

ences, there is only significantly higher consistency over Alaska/western Canada following La Niña (hatched regions in Figure 4.4c; see Text C.4). There is also a region of significant difference over the west coast of North America around 40°N following El Niño (Figure 4.4f). This region is a associated with a change in the most frequent sign, but the *consistency* of the associated sign is about the same between the two scenarios, in contrast to the difference seen in Alaska/western Canada (Figure 4.4c,f). The absence of significance in other regions may be due to low signal-to-noise driven by high internal variability of the system making it difficult to detect any differences 15-25 years into implementation within a 10-member ensemble [127].

While these two regions of significance align with regions of higher variability over North America (Figure 4.3c), these differences in consistency do not directly explain the higher boreal winter variability under SAI compared to SSP2-4.5. One possible explanation for this collocation over Alaska/western Canada may be due to higher ENSO variability in tandem with a more robust La Niña connection to this region, leading to higher ENSO teleconnection variability. In addition, the higher variability in both regions could also be related to the magnitude of ENSO teleconnections or from another source of seasonal variability. The connection between ENSO and higher seasonal variability along the west coast of North America is currently being investigated, and the results of this analysis will be included in the final publication.

Previous research has shown that changes to ENSO teleconnections can influence predictability [74], which raises the question as to whether the higher consistency over Alaska/western Canada under SAI may also influence future seasonal predictability. To examine if this difference in ENSO teleconnections is reflected in seasonal predictability differences, we use neural networks to quantify differences in seasonal prediction skill over Alaska/western Canada (denoted by the black box in Figure 4.4c,f) under the SAI scenario compared to the SSP2-4.5 scenario.

4.3.2 SAI impact on seasonal predictability

With higher La Niña teleconnection consistency under SAI compared to SSP2-4.5 over Alaska/western Canada, one may expect higher seasonal predictability of 2m temperature under SAI. To test this hypothesis, we input tropical SSTs into a neural network to predict the sign of mean 2m temperature anomalies in this region 2 months later. Ten neural networks are trained for each testing member (100 networks in total) under the SAI and SSP2-4.5 scenarios, separately. These ten networks are created by varying the random seed to test the sensitivity of the network to the random initialized weights. The resulting spread of accuracy across confidence thresholds (Section 4.2.2) is plotted in Figure 4.5. The mean accuracy at each threshold is also included as the corresponding thick colored line (Figure 4.5a). The confident predictions, defined at the 20% confidence threshold following Mayer and Barnes (2022), are plotted separately as box and whisker plots for better visualization of the spread in accuracy (Figure 4.5b).

In general, mean accuracy increases as model confidence increases for both scenarios (Figure 4.5a), suggesting that the networks are identifying forecasts of opportunity for this region [73].

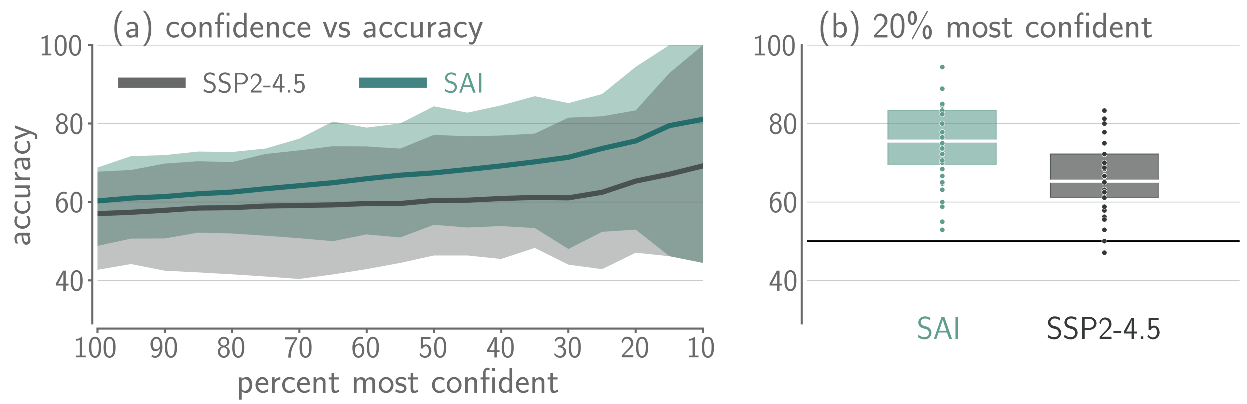


Figure 4.5: (a) Confidence versus accuracy for Alaska/Canada (55-70°N, 190-250°E). Teal and grey shading represents the spread in possible accuracies between network seed and testing member for the SAI and SSP2-4.5 scenario, respectively. The average accuracy at each confidence threshold is plotted as the correspondingly colored thick line. (b) Box and whisker plot of the accuracies at the 20% most confident level. The horizontal white line indicates the mean, the edges show the 25th and 75th percentile and the dots denote the individual accuracies for each network.

Specifically, it suggests that the neural network may be harnessing the state of ENSO to make its confident predictions, since the network uses monthly SST anomalies to make 2 month lead surface temperature predictions over northwest North America. In fact, we find that confident predictions do generally correspond to strong La Niña events under SAI and SSP2-4.5 (Figure C.1).

Even though both scenarios exhibit forecasts of opportunity, there is significantly higher skill under SAI compared to SSP2-4.5 at the higher confidence thresholds (see Text C.5 for significance test details). In other words, the largest differences in seasonal predictability between the two scenarios occur during forecasts of opportunity. However, we note that there is also a large spread in possible skill across members for each scenario that we hypothesize is a reflection of the large internal variability within ensemble members. Overall, we find that confident predictions generally corresponding to strong La Niña events and the largest differences in predictability occur at higher confidence values. This suggests that higher predictability under SAI compared to SSP2-4.5 may be related to the higher consistency of La Niña teleconnections over this region. However, confident predictions do not only occur during La Niña events (Figure C.1), and therefore, this difference could also be a reflection of differences in another source of predictability as well.

4.4 Discussion and Conclusion

Solar radiation modification (SRM) has been suggested as an approach to reduce the worst impacts of anthropogenic warming and provide society with additional time to deploy extensive mitigation efforts [70]. One plausible SRM strategy is stratospheric aerosol injection (SAI) [70], which would increase the reflectivity of the atmosphere through the introduction of small reflective particles into the stratosphere and thereby, reduce the surface temperature. Previous research has found that SAI implementation reduces future tropical precipitation compared to a future without intervention [72, 132], and this can ultimately have wide reaching impacts through atmospheric teleconnections initiated by tropical precipitation [29–31]. In particular, these changes can influence seasonal variability and predictability in the midlatitudes [40, 46, 135, 136]. Therefore, we investigate differences in Northern Hemisphere boreal winter variability and predictability under SAI compared to a future climate scenario with no SAI (SSP2-4.5) using output from the set of ARISE-SAI simulations [72].

We find that compared to a future under SSP2-4.5, there is higher boreal winter variance across much of the Northern Hemisphere under SAI, which is likely a reflection of a reduction in variability under SSP2-4.5 compared to the historical climate. In addition, we find significantly higher variance west of China, in northeast Russia, and across a large portion of north/west North America, where the North American regions align with locations known to be particularly influenced by ENSO [34, 39, 135, 137, 141]. Furthermore, we find significantly higher influence of La Niña on northwest North America under SAI than under SSP2-4.5, which corresponds with higher seasonal predictability under SAI as well. This suggests that ENSO plays a larger role in seasonal predictability of surface temperature over northwest North America under SAI compared to SSP2-4.5.

Overall, these results demonstrate that in the ARISE-SAI simulations, SAI can lead to higher seasonal variability and predictability along the northwest coast of North America compared to a future climate without SAI. However, future work is needed to examine how other seasons and sources of seasonal variability such as the North Atlantic Oscillation [142, 143] and sudden stratospheric warming events [144] may influence predictability in different climate states as well. To gain a more comprehensive understanding of the impact of SAI, we also need to examine how SAI may differ from the historical climate, and continue to explore the impact of various SAI implementation strategies and targets on the climate system. Ultimately, further research is required to determine whether SAI is the best course of action if climate change mitigation efforts are not enough.

Chapter 5: Conclusion and Discussion

5.1 Research Summary

The research presented in this dissertation demonstrates the utility of machine learning for subseasonal to seasonal (S2S; 2 weeks to a season) prediction. In particular, it demonstrates that neural networks, through network confidence and explainability techniques, can be used to identify physically meaningful S2S forecasts of opportunity. Further, it illustrates that this technique can be applied to assess S2S predictability under future climate scenarios, and demonstrates the use of explainability techniques to assist in identifying possible sources behind changes to predictability.

In the first study (Chapter 2) [73], we present a machine learning technique for identifying subseasonal forecasts of opportunity. We use a known connection between the MJO and the North Atlantic to construct a S2S prediction task for a neural network. Using this task, we then demonstrate that higher network confidence is associated with enhanced prediction skill, or forecasts of opportunity. To gauge the physical relevance of these network-identified opportunities, we employ an explainability technique (layer-wise relevance propagation) and find that the network is identifying known, physically relevant regions for enhanced S2S prediction skill in the North Atlantic. In addition, it also introduces a possible new forecast of opportunity. This study demonstrates that neural networks can identify physically meaningful forecasts of opportunity for S2S prediction.

In Chapter 3 [74], we use neural networks to quantify changes in prediction skill in a historical and future climate in the CESM2-LE. By applying the framework presented in the first study, we can identify both changes to predictability across all predictions and during forecasts of opportunity. In fact, we find that the largest changes in subseasonal predictability between the historical and future climate occur during forecasts of opportunity, whereas the accuracy for all predictions show minimal differences. These results demonstrate the value of this approach for identifying future changes in S2S predictability. We further show that these differences during forecasts of opportunity are mainly linked to changes in seasonal climate variability, and also find that the results have a physical basis in previous work on changes to MJO and ENSO teleconnections in a future climate.

For the final study (Chapter 4), we investigate the impact of SAI on seasonal variability and predictability compared to a climate change scenario without SAI deployed, using the ARISE-SAI simulations. Initially, we examine seasonal variability differences between the two future scenarios and find that under SAI, seasonal variability is higher across the northern hemisphere. Further, we also find that La Niña teleconnections to the northwest coast (Alaska/western Canada) are more consistent under SAI as well. We then investigate whether these differences in ENSO teleconnections correspond to differences in seasonal predictability, using the neural network approach presented in Chapter 2 and successfully applied in Chapter 3. We find differences in prediction skill between the two scenarios, where the largest differences in skill occur over higher confident predictions. Further, demonstrating the utility of this approach for identifying differences in future S2S predictability.

5.2 Future Avenues for Examining S2S Predictability with Neural Networks

This work provides a novel approach to the application of neural networks to S2S prediction and to assessing predictability in future climates. Moreover, it acts as a stepping stone for further development of the application of machine learning to S2S predictability, particularly through the lens of forecasts of opportunity.

The framework presented here for identifying subseasonal forecasts of opportunity has generated much interest in the academic and private sectors. However, the (artificial) neural networks used are simple, utilize only one input variable and are constructed to only predict the sign of an anomaly for either a single grid point or regional average. Here we provide a few future directions that could be used to either improve prediction skill or provide additional information to the user. First, convolutional neural networks may help improve accuracy, due to their ability to identify spatial structures from images (e.g. climate data maps) [145, 146]. They also require fewer trainable parameters than an artificial neural network, which may help minimize overfitting when applying this technique to observations [147]. Secondly, instead of a categorical prediction problem, regression could be used to provide additional magnitude information [58, 148–150]. This regression based approach could also be used to incorporate spatial patterns into the prediction. Specifically, a dimension reduction technique (i.e. empirical orthogonal function analysis) could be applied to the predictand, and therefore, the value predicted by the network would not only have a confidence value, but also a spatial pattern associated with it.

Another path forward for future work is the application of transfer learning to S2S prediction. Transfer learning applies information learned from one task, to a similar, data-limited task to help enhance the learning capabilities of a neural network [151]. With the breadth of data available through climate models compared to observational data, especially for longer time scale analyses, transfer learning could provide additional information for improving S2S prediction skill. Ham et al. (2019) present a successful example of this technique, and demonstrate that the larger sample sizes obtained through the application of transfer learning lead to an increase in ENSO prediction skill. Therefore, the application of transfer learning may also be beneficial to S2S prediction. While this approach can be used to improve prediction skill, it could also be used to learn more about the climate system. In particular, training neural networks for S2S prediction tasks initially on preindustrial (or historical) simulations, and then further training these networks on future climates such as under anthropogenic warming, could illuminate the usefulness of preindustrial runs for future climate analysis.

As a note for future applications, this research applies one explainability method to identify the relevant regions used to make confident predictions, but recent research has demonstrated that there is no optimal explainability technique. Therefore, multiple explainability methods (e.g. LRP, DeepSHAP and Integrated Gradients) should be applied for better transparency into the network and its decision making process [118, 152].

Beyond the application of neural networks, the thread of S2S predictability throughout this work is also woven through the focus on two tropical sources of predictability (MJO and ENSO) and their impact on the northern hemisphere boreal winter. However, S2S predictability can be sourced from the stratosphere (e.g. sudden stratospheric warmings) [13, 14], the land (e.g. surface moisture) [17, 18], and other tropical phenomena (e.g. East Asian summer monsoon) [2, 16], to name a few. Furthermore, certain sources are more applicable to different seasons and regions [17, 153]. Therefore, an exploration of these applications to other regions, seasons, and sources of predictability may shed further light on future predictability changes or may even identify new regions of importance for improved S2S prediction skill.

In addition to seasonal and regional dependence, lower frequency variability (i.e. decadal to multi-decadal variability) is also known to impact sources of S2S predictability [154–156] and their teleconnections [157]. Furthermore, Appendix B.2 of this dissertation shows that the neural network prediction skill fluctuates in the North Atlantic, depending on the number of ensemble members used for training. We hypothesize that these skill changes may be related to low frequency variability impacting how well the network learns. Given these results and previous research [154–157], the neural network predictability approach presented in this dissertation could be further applied to identify low frequency fluctuations in S2S predictability. Employing explainability techniques would allow for the examination of whether the relevance of certain S2S sources of predictability are impacted by low frequency variability. A possible approach to this analysis could be through the use of long preindustrial runs to understand decadal to multidecadal fluctuations in predictability in an unforced climate, or through a comparison of predictability between ensemble members in a large ensemble (e.g. CESM2-LE). The latter methodology could also be used to examine how decadal modulation of subseasonal predictability may change in the future, by comparing the variability in accuracy between a historical and future period (similar to the work in this dissertation). Further, the analysis could also be partitioned into accuracy across all predictions and network identified forecasts of opportunity (e.g. the 20% most confident predictions), to examine if the impact of low frequency variability is different between the two. However, additional analysis would be needed to determine whether the result is a product of internal multidecadal variability or from an externally forced climate change response.

As the climate continues to warm and plausible futures evolve, it is important to examine the evolution of and changes to predictability. This research provides a comprehensive approach for diagnosing S2S predictability changes near or at the end of the century, but further research should also examine if these changes are evident in other future greenhouse gas emission and stratospheric aerosol injection scenarios throughout the coming decades. In addition, this work focuses on predictability changes using the CESM2-LE and ARISE-SAI. There is a large spread in possible futures across climate models, due to model structural uncertainty, internal variability and emission scenario [158, 159]. Therefore, it is important to evaluate how consistent these predictability changes are across models, as well as the consistency of the evolution of predictability.

5.3 Concluding Thoughts

As the field of machine learning continues to develop, new machine learning and explainability techniques will likely emerge. It is important to continue to evaluate how these tools can be applied to the atmospheric sciences. Simultaneously, the S2S community is still working to identify, understand and harness all sources of predictability on S2S timescales. The work presented in this dissertation integrates these two scientific frontiers to provide an exciting guide and catalyst for continued research in both fields. This dissertation has introduced a systematic framework for assessing S2S predictability and its changes using machine learning and explainability techniques. The results presented in this dissertation also provide an outlook for S2S predictability across a range of futures and suggest new avenues for exploring predictability on S2S timescales and beyond.

Bibliography

[1] Christopher J White, Henrik Carlsen, Andrew W Robertson, Richard J T Klein, Jeffrey K Lazo, Arun Kumar, Frederic Vitart, Erin Coughlan de Perez, Andrea J Ray, Virginia Murray, Sukaina Bharwani, Dave MacLeod, Rachel James, Lora Fleming, Andrew P Morse, Bernd Eggen, Richard Graham, Erik Kjellström, Emily Becker, Kathleen V Pegion, Neil J Holbrook, Darryn McEvoy, Michael Depledge, Sarah Perkins-Kirkpatrick, Timothy J Brown, Roger Street, Lindsey Jones, Tomas A Remenyi, Indi Hodgson-Johnston, Carlo Buontempo, Rob Lamb, Holger Meinke, Berit Arheimer, and Stephen E Zebiak. Potential applications of subseasonal-to-seasonal (S2S) predictions. *Met. Apps*, 24(3):315–325, July 2017.

[2] Annarita Mariotti, Cory Baggett, Elizabeth A Barnes, Emily Becker, Amy Butler, Dan C Collins, Paul A Dirmeyer, Laura Ferranti, Nathaniel C Johnson, Jeanine Jones, Ben P Kirtman, Andrea L Lang, Andrea Molod, Matthew Newman, Andrew W Robertson, Siegfried Schubert, Duane E Waliser, and John Albers. Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Am. Meteorol. Soc.*, January 2020.

[3] W J Merryfield, J Baehr, L Batté, and others. Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the*, 2020.

[4] Christopher J White, Daniela I V Domeisen, Nachiketa Acharya, Elijah A Adefisan, Michael L Anderson, Stella Aura, Ahmed A Balogun, Douglas Bertram, Sonia Bluhm, David J Brayshaw, Jethro Browell, Dominik Büeler, Andrew Charlton-Perez, Xandre Chourio, Isadora Christel, Caio A S Coelho, Michael J DeFlorio, Luca Delle Monache, Francesca Di Giuseppe, Ana María García-Solórzano, Peter B Gibson, Lisa Goddard, Carmen González Romero, Richard J Graham, Robert M Graham, Christian M Grams, Alan Halford, W T Katty Huang, Kjeld Jensen, Mary Kilavi, Kamoru A Lawal, Robert W Lee, David MacLeod, Andrea Manrique-Suñén, Eduardo S P Martins, Carolyn J Maxwell, William J Merryfield, Ángel G Muñoz, Eniola Olaniyan, George Otieno, John A Oyedepo,

Lluís Palma, Ilias G Pechlivanidis, Diego Pons, F Martin Ralph, Dirceu S Reis, Tomas A Remenyi, James S Risbey, Donald J C Robertson, Andrew W Robertson, Stefan Smith, Albert Soret, Ting Sun, Martin C Todd, Carly R Tozer, Francisco C Vasconcelos, Ilaria Vigo, Duane E Waliser, Fredrik Wetterhall, and Robert G Wilson. Advances in the application and utility of subseasonal-to-seasonal predictions. *Bull. Am. Meteorol. Soc.*, -1(aop):1–57, November 2021.

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

[6] F Vitart, C Ardilouze, A Bonet, A Brookshaw, M Chen, C Codorean, M Déqué, L Ferranti, E Fucile, M Fuentes, H Hendon, J Hodgson, H-S Kang, A Kumar, H Lin, G Liu, X Liu, P Malguzzi, I Mallas, M Manoussakis, D Mastrangelo, C MacLachlan, P McLean, A Minami, R Mladek, T Nakazawa, S Najm, Y Nie, M Rixen, A W Robertson, P Ruti, C Sun, Y Takaya, M Tolstykh, F Venuti, D Waliser, S Woolnough, T Wu, D-J Won, H Xiao, R Zaripov, and L Zhang. The subseasonal to seasonal (S2S) prediction project database. *Bull. Am. Meteorol. Soc.*, 98(1):163–173, January 2017.

[7] K-C Tseng, E A Barnes, and E D Maloney. 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, January 2018.

[8] Cory F Baggett, Kyle M Nardi, Samuel J Childs, Samantha N Zito, Elizabeth A Barnes, and Eric D Maloney. Skillful subseasonal forecasts of weekly tornado and hail activity using the madden-julian oscillation. *J. Geophys. Res.*, 123(22):12,661–12,675, November 2018.

[9] Robert E Livezey and Marina M Timofeyeva. The first decade of Long-Lead U.S. seasonal forecasts: Insights from a skill analysis. *Bull. Am. Meteorol. Soc.*, 89(6):843–854, June 2008.

[10] Anthony G Barnston, Shuhua Li, Simon J Mason, David G DeWitt, Lisa Goddard, and Xiaofeng Gong. Verification of the first 11 years of IRI's seasonal climate forecasts. *J. Appl. Meteorol. Climatol.*, 49(3):493–520, March 2010.

[11] Nathaniel C Johnson, Dan C Collins, Steven B Feldstein, Michelle L L'Heureux, and Emily E Riddle. Skillful wintertime north american temperature forecasts out to 4 weeks based on the state of ENSO and the MJO. *Weather Forecast.*, 29(1):23–38, February 2014.

[12] Lei Wang and Andrew W Robertson. Week 3–4 predictability over the united states assessed from two operational ensemble prediction systems. *Clim. Dyn.*, 52(9):5861–5875, May 2019.

[13] Daniela I V Domeisen, Amy H Butler, Andrew J Charlton-Perez, Blanca Ayarzagüena, Mark P Baldwin, Etienne Dunn-Sigouin, Jason C Furtado, Chaim I Garfinkel, Peter Hitchcock, Alexey Yu Karpechko, Hera Kim, Jeff Knight, Andrea L Lang, Eun-pa Lim, Andrew Marshall, Greg Roff, Chen Schwartz, Isla R Simpson, Seok-woo Son, and Masakazu Taguchi. The role of the stratosphere in subseasonal to seasonal prediction: 2. predictability arising from stratosphere-troposphere coupling. *J. Geophys. Res.*, 125(2), January 2020.

[14] Chaim I Garfinkel, Chen Schwartz, Daniela I V Domeisen, Seok-Woo Son, Amy H Butler, and Ian P White. Extratropical atmospheric predictability from the Quasi-Biennial oscillation in subseasonal forecast models. *J. Geophys. Res. D: Atmos.*, 140:1, August 2018.

[15] Kyung-Ja Ha, Ye-Won Seo, June-Yi Lee, R H Kripalani, and Kyung-Sook Yun. Linkages between the south and east asian summer monsoons: a review and revisit. *Clim. Dyn.*, 51(11):4207–4227, December 2018.

[16] Kelsey Malloy and Ben P Kirtman. The summer Asia–North america teleconnection and its modulation by ENSO in community atmosphere model, version 5 (CAM5). *Clim. Dyn.*, March 2022.

[17] Zhichang Guo, Paul A Dirmeyer, and Tim DelSole. Land surface impacts on subseasonal and seasonal predictability. *Geophys. Res. Lett.*, 38(24), 2011.

[18] Paul A Dirmeyer, Subhadeep Halder, and Rodrigo Bombardi. On the harvest of predictability from land states in a global forecast model. *J. Geophys. Res.*, 123(23):13,111–13,127, December 2018.

[19] Stephanie A Henderson, Eric D Maloney, and Elizabeth A Barnes. The influence of the Madden–Julian oscillation on northern hemisphere winter blocking. *J. Clim.*, 29(12):4597–4616, June 2016.

[20] Cory F Baggett, Elizabeth A Barnes, Eric D Maloney, and Bryan D Mundhenk. Advancing atmospheric river forecasts into subseasonal-to-seasonal time scales. *Geophys. Res. Lett.*, 44(14):2017GL074434, July 2017.

[21] Cristiana Stan, David M Straus, Jorgen S Frederiksen, Hai Lin, Eric D Maloney, and Courtney Schumacher. Review of Tropical-Extratropical teleconnections on intraseasonal time scales: The subseasonal to seasonal (S2S) teleconnection Sub-Project. *Rev. Geophys.*, 55(4):902–937, December 2017.

[22] Cheng Zheng, Edmund Kar-Man Chang, Hye-Mi Kim, Minghua Zhang, and Wanqiu Wang. Impacts of the Madden–Julian oscillation on Storm-Track activity, surface air temperature, and precipitation over north america. *J. Clim*, 31, August 2018.

[23] Stephanie A Henderson and Eric D Maloney. The impact of the Madden–Julian oscillation on High-Latitude winter blocking during el niño–southern oscillation events. *J. Clim.*, 31(13):5293–5318, July 2018.

[24] M C Arcodia, B P Kirtman, and L S P Siqueira. How MJO teleconnections and ENSO interference impacts US precipitation. *J. Clim.*, 2020.

[25] Kai-Chih Tseng, Elizabeth A Barnes, and Eric Maloney. The importance of past MJO activity in determining the future state of the midlatitude circulation. *J. Clim.*, 33(6):2131–2147, February 2020.

[26] Roland A Madden and Paul R Julian. Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. *J. Atmos. Sci.*, 28(5):702–708, July 1971.

[27] Roland A Madden and Paul R Julian. Description of Global-Scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.*, 29(6):1109–1123, September 1972.

[28] Roland A Madden and Paul R Julian. Observations of the 40–50-day tropical Oscillation—A review. *Mon. Weather Rev.*, 122(5):814–837, May 1994.

[29] Brian J Hoskins and Tercio Ambrizzi. Rossby wave propagation on a realistic longitudinally varying flow. *J. Atmos. Sci.*, 50(12):1661–1671, June 1993.

[30] Brian J Hoskins and David J Karoly. The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.*, 38(6):1179–1196, June 1981.

[31] Prashant D Sardeshmukh and Brian J Hoskins. The generation of global rotational flow by steady idealized tropical divergence. *J. Atmos. Sci.*, 45(7):1228–1251, April 1988.

[32] Kevin E Trenberth. The definition of el nino. *Bull. Am. Meteorol. Soc.*, 78(12):2771–2778, 1997.

[33] Benjamin Pohl and Adrian J Matthews. Observed changes in the lifetime and amplitude of the Madden–Julian oscillation associated with interannual ENSO sea surface temperature anomalies. *J. Clim.*, 20(11):2659–2674, June 2007.

[34] Jacob Bjerknes. Atmospheric teleconnections from the equatorial pacific. *Mon. Weather Rev.*, 97(3):163–172, 1969.

[35] Kai-Chih Tseng, Eric Maloney, and Elizabeth A Barnes. The consistency of MJO teleconnection patterns on interannual time scales. *J. Clim.*, 33(9):3471–3486, March 2020.

[36] Ja-Yeon Moon, Bin Wang, and Kyung-Ja Ha. ENSO regulation of MJO teleconnection. *Clim. Dyn.*, 37(5):1133–1149, September 2011.

[37] Paul E Roundy, Kyle MacRitchie, Jonas Asuma, and Timothy Melino. 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, August 2010.

[38] John D Horel and John M Wallace. Planetary-Scale atmospheric phenomena associated with the southern oscillation. *Mon. Weather Rev.*, 109(4):813–829, April 1981.

[39] Henry F Diaz, Martin P Hoerling, and Jon K Eischeid. ENSO variability, teleconnections and climate change. *Int. J. Climatol.*, 21(15):1845–1862, December 2001.

[40] Andréa S Taschetto, Caroline C Ummenhofer, Malte F Stuecker, Dietmar Dommenget, Karumuri Ashok, Regina R Rodrigues, and Sang-Wook Yeh. ENSO atmospheric teleconnections, November 2020.

[41] Bryan D Mundhenk, Elizabeth A Barnes, Eric D Maloney, and Cory F Baggett. 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, February 2018.

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

[43] Kai-Chih Tseng, Nathaniel C Johnson, Eric D Maloney, Elizabeth A Barnes, and Sarah B Kapnick. Mapping large-scale climate variability to hydrological extremes: An application of the linear inverse model to subseasonal prediction. *J. Clim.*, -1(aop):1–58, February 2021.

[44] John R Albers and Matthew Newman. A priori identification of skillful extratropical subseasonal forecasts. *Geophys. Res. Lett.*, 46(21):12527–12536, November 2019.

[45] John R Albers and Matthew Newman. Subseasonal predictability of the north atlantic oscillation. *Environ. Res. Lett.*, February 2021.

[46] William E Chapman, Luca Delle Monache, Stefano Alessandrini, Aneesh C Subramanian, F Martin Ralph, Shang-Ping Xie, Sebastian Lerch, and Negin Hayatbini. Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Mon. Weather Rev.*, -1(aop), October 2021.

[47] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775):568–572, September 2019.

[48] Stephan Rasp and Nils Thuerey. Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for WeatherBench. *J. Adv. Model. Earth Syst.*, 13(2), February 2021.

[49] Jonathan A Weyn, Dale R Durran, Rich Caruana, and Nathaniel Cresswell-Clay. Subseasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. Model. Earth Syst.*, June 2021.

[50] Michael Scheuerer, Matthew B Switanek, Rochelle P Worsnop, and Thomas M Hamill. Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over california. *Mon. Weather Rev.*, 148(8):3489–3506, August 2020.

[51] T Chen and H Chen. 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, 1995.

[52] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. June 2020.

[53] Benjamin A Toms, Elizabeth A Barnes, and Imme Ebert-Uphoff. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *J. Adv. Model. Earth Syst.*, 12(9), September 2020.

[54] Benjamin A Toms, Karthik Kashinath, Prabhat, and Da Yang. Testing the reliability of interpretable neural networks in geoscience using the Madden-Julian oscillation. August 2020.

[55] Zachary M Labe and Elizabeth A Barnes. Predicting slowdowns in decadal climate warming trends with explainable neural networks. *Geophys. Res. Lett.*, May 2022.

[56] Zane K Martin, Elizabeth A Barnes, and Eric Maloney. Using simple, explainable neural networks to predict the Madden-Julian oscillation. *J. Adv. Model. Earth Syst.*, May 2022.

[57] Frances V Davenport and Noah S Diffenbaugh. Using machine learning to analyze physical causes of climate change: A case study of U.S. midwest extreme precipitation. *Geophys. Res. Lett.*, July 2021.

[58] Emily M Gordon, Elizabeth A Barnes, and James W Hurrell. Oceanic harbingers of pacific decadal oscillation predictability in CESM2 detected by neural networks. *Geophys. Res. Lett.*, 48(21), November 2021.

[59] IPCC. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press., 2022.

[60] Eric D Maloney, Ángel F Adames, and Hien X Bui. Madden–Julian oscillation changes under anthropogenic warming. *Nat. Clim. Chang.*, 9(1):26–33, December 2018.

[61] Wenju Cai, Agus Santoso, Matthew Collins, Boris Dewitte, Christina Karamperidou, Jong-Seong Kug, Matthieu Lengaigne, Michael J McPhaden, Malte F Stuecker, Andréa S Taschetto, Axel Timmermann, Lixin Wu, Sang-Wook Yeh, Guojian Wang, Benjamin Ng, Fan Jia, Yun Yang, Jun Ying, Xiao-Tong Zheng, Tobias Bayr, Josephine R Brown, Antonietta Capotondi, Kim M Cobb, Bolan Gan, Tao Geng, Yoo-Geun Ham, Fei-Fei Jin, Hyun-Su Jo, Xichen Li, Xiaopei Lin, Shayne McGregor, Jae-Heung Park, Karl Stein, Kai Yang,

Li Zhang, and Wenxiu Zhong. Changing el niño–southern oscillation in a warming climate. *Nature Reviews Earth & Environment*, 2(9):628–644, August 2021.

[62] Jonathan D Beverley, Matthew Collins, F Hugo Lambert, and Robin Chadwick. Future changes to el niño teleconnections over the north pacific and north america. *J. Clim.*, 34(15):6191–6205, August 2021.

[63] Jingxuan Cui and Tim Li. Changes in MJO characteristics and impacts in the past century. *J. Clim.*, -1(aop):1–1, October 2021.

[64] Marie Drouard and Christophe Cassou. A modeling- and Process-Oriented study to investigate the projected change of ENSO-Forced wintertime teleconnectivity in a warmer world. *J. Clim.*, 32(23):8047–8068, December 2019.

[65] D R Fereday, R Chadwick, J R Knight, and A A Scaife. Tropical rainfall linked to stronger future ENSO-NAO teleconnection in CMIP5 models. *Geophys. Res. Lett.*, n/a(n/a):e2020GL088664, October 2020.

[66] Savini M Samarasinghe, Charlotte Connolly, Elizabeth A Barnes, Imme Ebert-Uphoff, and Lantao Sun. Strengthened causal connections between the MJO and the north atlantic with climate warming. *Geophys. Res. Lett.*, 48(5), March 2021.

[67] Wenyu Zhou, Da Yang, Shang-Ping Xie, and Jing Ma. Amplified Madden–Julian oscillation impacts in the Pacific–North america region. *Nat. Clim. Chang.*, 10(7):654–660, July 2020.

[68] Gerald A Meehl and Haiyan Teng. Multi-model changes in el niño teleconnections over north america in a future warmer climate. *Clim. Dyn.*, 29(7-8):779–790, October 2007.

[69] Zhen-Qiang Zhou, Shang-Ping Xie, Xiao-Tong Zheng, Qinyu Liu, and Hai Wang. Global Warming–Induced changes in el niño teleconnections over the north pacific and north america. *J. Clim.*, 27(24):9050–9064, December 2014.

[70] Engineering National Academies of Sciences and Medicine. *Reflecting Sunlight: Recommendations for Solar Geoengineering Research and Research Governance*. The National Academies Press, Washington, DC, 2021.

[71] Ben Kravitz, Douglas G MacMartin, Simone Tilmes, Jadwiga H Richter, Michael J Mills, Wei Cheng, Katherine Dagon, Anne S Glanville, Jean-Francois Lamarque, Isla R Simpson, Joseph Tribbia, and Francis Vitt. Comparing surface and stratospheric impacts of geoengineering with different SO 2 injection strategies. *J. Geophys. Res.*, 124(14):7900–7918, July 2019.

[72] Jadwiga Richter, Daniele Visioni, Douglas MacMartin, David Bailey, Nan Rosenbloom, Walker Lee, Mari Tye, and Jean-Francois Lamarque. Assessing responses and impacts of solar climate intervention on the earth system with stratospheric aerosol injection (arise-sai). April 2022.

[73] Kirsten J Mayer and Elizabeth A Barnes. Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophys. Res. Lett.*, May 2021.

[74] Kirsten J Mayer and Elizabeth A Barnes. Quantifying the effect of climate change on midlatitude subseasonal prediction skill provided by the tropics. *Geophys. Res. Lett.*, 49(14), July 2022.

[75] Annarita Mariotti, Paolo M Ruti, and Michel Rixen. Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *npj Climate and Atmospheric Science*, 1(1):1–4, March 2018.

[76] Andrew W Robertson, Arun Kumar, Malaquias Peña, and Frederic Vitart. Improving and promoting subseasonal to seasonal prediction. *Bull. Am. Meteorol. Soc.*, 96(3):ES49–ES53, March 2015.

[77] Kathy Pegion, Ben P Kirtman, Emily Becker, Dan C Collins, Emerson LaJoie, Robert Burgman, Ray Bell, Timothy DelSole, Dughong Min, Yuejian Zhu, Wei Li, Eric Sinsky,

Hong Guan, Jon Gottschalck, E Joseph Metzger, Neil P Barton, Deepthi Achuthavarier, Jelena Marshak, Randal D Koster, Hai Lin, Normand Gagnon, Michael Bell, Michael K Tippett, Andrew W Robertson, Shan Sun, Stanley G Benjamin, Benjamin W Green, Rainer Bleck, and Hyemi Kim. The subseasonal experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Am. Meteorol. Soc.*, 100(10):2043–2060, October 2019.

[78] John Abbot and Jennifer Marohasy. Input selection and optimisation for monthly rainfall forecasting in queensland, australia, using artificial neural networks. *Atmos. Res.*, 138:166–178, March 2014.

[79] Benjamin A Toms, Elizabeth A Barnes, and Imme Ebert-Uphoff. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *J. Adv. Model. Earth Syst.*, 12(9), September 2020.

[80] B. Liebmann and C.A. Smith. Description of a complete (interpolated) outgoing longwave radiation dataset. *Bull. Am. Meteorol. Soc.*, 77:1275–1277, 1996.

[81] Dick P Dee, S M Uppala, A J Simmons, Paul Berrisford, P Poli, S Kobayashi, U Andrae, M A Balmaseda, G Balsamo, d P Bauer, and Others. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, 137(656):553–597, 2011.

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

[83] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise explanations for Non-Linear classifier decisions by Layer-Wise relevance propagation. *PLoS One*, 10(7):e0130140, July 2015.

[84] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-Wise relevance propagation: An overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors,

Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pages 193–209. Springer International Publishing, Cham, 2019.

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

[86] Kai-Chih Tseng, Eric Maloney, and Elizabeth Barnes. The consistency of MJO teleconnection patterns: An explanation using linear rossby wave theory. *J. Clim.*, 32(2):531–548, January 2019.

[87] Hai Lin, Gilbert Brunet, and Jacques Derome. An observed connection between the north atlantic oscillation and the Madden–Julian oscillation. *J. Clim.*, 22(2):364–380, January 2009.

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

[89] Xiaolu Shao, David Martin Straus, Shuanglin Li, Erik Swenson, Priyanka Yadav, and Jie Song. Forcing of the MJO-related indian ocean heating on the intraseasonal lagged NAO. 2020.

[90] Hai Lin and Gilbert Brunet. 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), January 2011.

[91] Amy McGovern, Ryan Lagerquist, David John Gagne, G Eli Jergensen, Kimberly L Elmore, Cameron R Homeyer, and Travis Smith. Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Am. Meteorol. Soc.*, 100(11):2175–2199, November 2019.

[92] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[93] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

[94] Michael A Nielsen. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, 2015.

[95] Harry H Hendon, Chidong Zhang, and John D Glick. Interannual variation of the Madden–Julian oscillation during austral summer. *J. Clim.*, 12(8):2538–2550, August 1999.

[96] William S Kessler. EOF representations of the Madden–Julian oscillation and its connection with ENSO. *J. Clim.*, 14(13):3055–3061, July 2001.

[97] Jerome Namias. Persistence of flow patterns over north america and adjacent ocean sectors. *Mon. Weather Rev.*, 114(7):1368–1383, July 1986.

[98] Chiharu Takahashi and Ryuichi Shirooka. Storm track activity over the north pacific associated with the Madden-Julian oscillation under ENSO conditions during boreal winter. *J. Geophys. Res.*, 119(18):10,663–10,683, September 2014.

[99] Kai-Chih Tseng, Eric Maloney, and Elizabeth A Barnes. The consistency of MJO teleconnection patterns on interannual time scales. *J. Clim.*, 33(9):3471–3486, May 2020.

[100] Aditi Sheshadri, Marshall Borrus, Mark Yoder, and Thomas Robinson. Midlatitude error growth in atmospheric GCMs: The role of eddy growth rate. *Geophys. Res. Lett.*, 48(23), December 2021.

[101] Keith B Rodgers, Sun-Seon Lee, Nan Rosenbloom, Axel Timmermann, Gokhan Danabasoglu, Clara Deser, Jim Edwards, Ji-Eun Kim, Isla Simpson, Karl Stein, and Others. Ubiquity of human-induced changes in climate variability. *Earth System Dynamics Discussions*, pages 1–22, 2021.

[102] Min-seop Ahn, Daehyun Kim, Daehyun Kang, Jiwoo Lee, Kenneth R Sperber, Peter J Gleckler, Xianan Jiang, Yoo-geun Ham, and Hyemi Kim. MJO propagation across the

maritime continent: Are CMIP6 models better than CMIP5 models? *Geophys. Res. Lett.*, 47(11):741, June 2020.

[103] Jiabao Wang, Hyemi Kim, and Michael J DeFlorio. Future changes of PNA-like MJO teleconnections in CMIP6 models: underlying mechanisms and uncertainty. *Journal of Climate*, pages 1–40, 2022.

[104] G Danabasoglu, J -F Lamarque, J Bacmeister, D A Bailey, A K DuVivier, J Edwards, L K Emmons, J Fasullo, R Garcia, A Gettelman, C Hannay, M M Holland, W G Large, P H Lauritzen, D M Lawrence, J T M Lenaerts, K Lindsay, W H Lipscomb, M J Mills, R Neale, K W Oleson, B Otto-Bliesner, A S Phillips, W Sacks, S Tilmes, L Kampenhout, M Vertenstein, A Bertini, J Dennis, C Deser, C Fischer, B Fox-Kemper, J E Kay, D Kinnison, P J Kushner, V E Larson, M C Long, S Mickelson, J K Moore, E Nienhouse, L Polvani, P J Rasch, and W G Strand. The community earth system model version 2 (CESM2). *J. Adv. Model. Earth Syst.*, 12(2):106, February 2020.

[105] Ncar climate data guide. https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni, 2020. Accessed: 2022-2-11.

[106] Masato Mori and Masahiro Watanabe. The growth and triggering mechanisms of the PNA: A MJO-PNA coherence. . *2*, 86(1):213–236, 2008.

[107] John M Wallace and David S Gutzler. Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Weather Rev.*, 109(4):784–812, April 1981.

[108] Yuan Zhang, John M Wallace, and Naoto Iwasaka. Is climate variability over the north pacific a linear response to ENSO? *J. Clim.*, 9(7):1468–1478, July 1996.

[109] Emily E Riddle, Marshall B Stoner, Nathaniel C Johnson, Michelle L L'Heureux, Dan C Collins, and Steven B Feldstein. The impact of the MJO on clusters of wintertime circulation anomalies over the north american region. *Clim. Dyn.*, 40(7):1749–1766, April 2013.

[110] Stephanie A Henderson, Daniel J Vimont, and Matthew Newman. The critical role of Non-Normality in partitioning tropical and extratropical contributions to PNA growth. *J. Clim.*, 33(14):6273–6295, June 2020.

[111] Hege-Beate Fredriksen, Judith Berner, Aneesh C Subramanian, and Antonietta Capotondi. How does el niño–southern oscillation change under global warming—a first look at CMIP6. *Geophys. Res. Lett.*, 47(22), November 2020.

[112] Gerald A Meehl and Haiyan Teng. Multi-model changes in el niño teleconnections over north america in a future warmer climate. *Clim. Dyn.*, 29(7-8):779–790, October 2007.

[113] Müller and Roeckner. ENSO teleconnections in projections of future climate in ECHAM5/MPI-OM. *Climate Dynamics*, 31:533–549, 2008.

[114] Jong-Seong Kug, Soon-Il An, Yoo-Geun Ham, and In-Sik Kang. Changes in el niño and la niña teleconnections over north Pacific–America in the global warming simulations. *Theor. Appl. Climatol.*, 100(3):275–282, May 2010.

[115] Brandon O Wolding, Eric D Maloney, Stephanie Henderson, and Mark Branson. Climate change and the madden-julian oscillation: A vertically resolved weak temperature gradient analysis. *J. Adv. Model. Earth Syst.*, 9(1):307–331, March 2017.

[116] Andrea M Jenney, David A Randall, and Elizabeth A Barnes. Drivers of uncertainty in future projections of Madden–Julian oscillation teleconnections. *Weather Clim. Dynam.*, 2(3):653–673, July 2021.

[117] B Wang and H Rui. Synoptic climatology of transient tropical intraseasonal convection anomalies: 1975–1985. *Meteorol. Atmos. Phys.*, 44(1):43–61, March 1990.

[118] Antonios Mamalakis, Imme Ebert-Uphoff, and Elizabeth A Barnes. Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. March 2021.

[119] Jerome H Friedman. Fast sparse regression and classification. *Int. J. Forecast.*, 28(3):722–738, July 2012.

[120] Isla R Simpson, Clara Deser, Karen A McKinnon, and Elizabeth A Barnes. Modeled and observed multidecadal variability in the north atlantic jet stream and its connection to sea surface temperatures. *J. Clim.*, 31(20):8313–8338, October 2018.

[121] Christophe Cassou, Yochanan Kushnir, Ed Hawkins, Anna Pirani, Fred Kucharski, In-Sik Kang, and Nico Caltabiano. Decadal climate variability and predictability: Challenges and opportunities. *Bull. Am. Meteorol. Soc.*, 99(3):479–490, March 2018.

[122] B L Welch. The generalisation of student's problems when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[123] Wake Smith. The cost of stratospheric aerosol injection through 2100. *Environ. Res. Lett.*, 15(11):114004, October 2020.

[124] A Robock, A Marquardt, B Kravitz, and others. Benefits, risks, and costs of stratospheric geoengineering. *Geophys. Res. Lett.*, 2009.

[125] Wake Smith, Umang Bhattarai, Donald C Bingaman, James L Mace, and Christian V Rice. Review of possible very high-altitude platforms for stratospheric aerosol injection. *Environ. Res. Commun.*, 4(3):031002, March 2022.

[126] Douglas G MacMartin and Ben Kravitz. The engineering of climate engineering. *Annu. Rev. Control Robot. Auton. Syst.*, 2(1):445–467, May 2019.

[127] Douglas G MacMartin, Peter J Irvine, Ben Kravitz, and Joshua B Horton. Technical characteristics of a solar geoengineering deployment and implications for governance. *Clim. Policy*, 19(10):1325–1339, November 2019.

[128] Simone Tilmes, Jadwiga H Richter, Michael J Mills, Ben Kravitz, Douglas G MacMartin, Rolando R Garcia, Douglas E Kinnison, Jean-francois Lamarque, Joseph Tribbia, and Fran-

cis Vitt. Effects of different stratospheric SO_2 injection altitudes on stratospheric chemistry and dynamics. *J. Geophys. Res.*, 123(9):4654–4673, May 2018.

[129] Antara Banerjee, Amy H Butler, Lorenzo M Polvani, Alan Robock, Isla R Simpson, and Lantao Sun. Robust winter warming over eurasia under stratospheric sulfate geoengineering – the role of stratospheric dynamics. *Atmos. Chem. Phys.*, 21(9):6985–6997, May 2021.

[130] Andy Jones, Jim M Haywood, Adam A Scaife, Olivier Boucher, Matthew Henry, Ben Kravitz, Thibaut Lurton, Pierre Nabat, Ulrike Niemeier, Roland Séférian, Simone Tilmes, and Daniele Visioni. The impact of stratospheric aerosol intervention on the north atlantic and Quasi-Biennial oscillations in the geoengineering model intercomparison project (GeoMIP) g6sulfur experiment. *Atmos. Chem. Phys.*, 22(5):2999–3016, March 2022.

[131] Jiu Jiang, Long Cao, Douglas G MacMartin, Isla R Simpson, Ben Kravitz, Wei Cheng, Daniele Visioni, Simone Tilmes, Jadwiga H Richter, and Michael J Mills. Stratospheric sulfate aerosol geoengineering could alter the high-latitude seasonal cycle. *Geophys. Res. Lett.*, 46(23):14153–14163, December 2019.

[132] Ben Kravitz, Douglas G. MacMartin, Michael J. Mills, Jadwiga H. Richter, Simone Tilmes, Jean-Francois Lamarque, Joseph J. Tribbia, and Francis Vitt. First simulations of designing stratospheric sulfate aerosol geoengineering to meet multiple simultaneous climate objectives. 2017.

[133] Alan Robock, Luke Oman, and Georgiy L Stenchikov. Regional climate responses to geoengineering with tropical and arctic SO2injections. *J. Geophys. Res.*, 113(D16), August 2008.

[134] C F Ropelewski and M S Halpert. Global and regional scale precipitation patterns associated with the el niño/southern oscillation. *Mon. Weather Rev.*, 115(8):1606–1626, August 1987.

[135] Kevin E. Trenberth, Grant W. Branstator, David Karoly, Arun Kumar, Ngar-Cheung Lau, and Chester Ropelewski. Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. 1998.

[136] Michael J McPhaden, Stephen E Zebiak, and Michael H Glantz. ENSO as an integrating concept in earth science. *Science*, 314(5806):1740–1745, December 2006.

[137] C F Ropelewski and M S Halpert. North american precipitation and temperature patterns associated with the el niño/southern oscillation (ENSO). *Mon. Weather Rev.*, 114(12):2352–2362, December 1986.

[138] Michael Winton. Amplified arctic climate change: What does surface albedo feedback have to do with it? *Geophys. Res. Lett.*, 33(3), 2006.

[139] James A Screen. Arctic amplification decreases temperature variance in northern mid- to high-latitudes. *Nat. Clim. Chang.*, 4(7):577–582, June 2014.

[140] Tapio Schneider, Tobias Bischoff, and Hanna Płotka. Physics of changes in synoptic midlatitude temperature variability. *J. Clim.*, 28(6):2312–2331, March 2015.

[141] Tao Zhang, Martin P Hoerling, Judith Perlwitz, De-Zheng Sun, and Donald Murray. Physics of U.S. surface temperature response to ENSO. *J. Clim.*, 24(18):4874–4887, September 2011.

[142] James W Hurrell. Influence of variations in extratropical wintertime teleconnections on northern hemisphere temperature. *Geophys. Res. Lett.*, 23(6):665–668, March 1996.

[143] James W Hurrell and Clara Deser. North atlantic climate variability: The role of the north atlantic oscillation. *J. Mar. Syst.*, 79(3):231–244, February 2010.

[144] Mark P Baldwin, Blanca Ayarzagüena, Thomas Birner, Neal Butchart, Amy H Butler, Andrew J Charlton-Perez, Daniela I V Domeisen, Chaim I Garfinkel, Hella Garny, Edwin P

Gerber, Michaela I Hegglin, Ulrike Langematz, and Nicholas M Pedatella. Sudden stratospheric warmings. *Rev. Geophys.*, 59(1), March 2021.

[145] LeCun, Y., Bottou, L., Bengio, Y. , and Haffner, P. Gradient-Based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[146] Neena Aloysius and M Geetha. A review on deep convolutional neural networks. In *2017 International Conference on Communication and Signal Processing (ICCSP)*, pages 0588–0592, April 2017.

[147] Yann LeCun and Yoshua Bengio. Convolutional networks for images speech and Time-Series. 1995.

[148] Elizabeth A Barnes, Randal J Barnes, and Nicolas Gordillo. Adding uncertainty to neural network regression tasks in the geosciences. September 2021.

[149] Emily M Gordon and Elizabeth A Barnes. Incorporating uncertainty into a regression neural network enables identification of decadal State-Dependent predictability. March 2022.

[150] Elizabeth A Barnes, Randal J Barnes, and Mark DeMaria. Sinh-arcsinh-normal distributions to add uncertainty to neural network regression tasks: applications to tropical cyclone intensity forecasts. July 2022.

[151] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014.

[152] Antonios Mamalakis, Elizabeth A Barnes, and Imme Ebert-Uphoff. Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. February 2022.

[153] A M Jenney, K M Nardi, E A Barnes, and D A Randall. The seasonality and regionality of MJO impacts on north american temperature. *Geophys. Res. Lett.*, 46(15):9193–9202, August 2019.

[154] Panini Naidu C Dasgupta and M K Roxy. Exploring the long-term changes in the madden julian oscillation using machine learning. *London*, 10(1):s41598–020, 2020.

[155] Nina Schuhen, Nathalie Schaller, Hannah C Bloomfield, David J Brayshaw, Llorenç Lledó, Irene Cionni, and Jana Sillmann. Predictive skill of teleconnection patterns in twentieth century seasonal hindcasts and their relationship to extreme winter temperatures in europe. *Geophys. Res. Lett.*, May 2022.

[156] Zhen Fu, Pang-Chi Hsu, Juan Li, Jian Cao, Young-Min Yang, and Fei Liu. Multidecadal changes in zonal displacement of tropical pacific MJO variability modulated by north atlantic SST. *J. Clim.*, -1(aop):1–42, May 2022.

[157] Yao Ge and Dehai Luo. Impacts of the different types of el niño and PDO on the winter sub-seasonal north american zonal temperature dipole via the variability of positive PNA events. *Clim. Dyn.*, July 2022.

[158] Flavio Lehner, Clara Deser, Nicola Maher, Jochem Marotzke, Erich M Fischer, Lukas Brunner, Reto Knutti, and Ed Hawkins. Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth Syst. Dyn.*, 11(2):491–508, May 2020.

[159] Clara Deser. "certain uncertainty: The role of internal climate variability in projections of regional climate change and risk management". *Earths Future*, 8(12), December 2020.

Appendix A: Chapter 2 Supporting Information

A.1 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.

A.2 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 [42, 87]. 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.

A.3 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 A.1) 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).

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

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.

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

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:

$$\tilde{y_i} = \frac{exp(x_i)}{\sum_j exp(x_j)} \tag{A.3}$$

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.

$$loss = -\sum_i y_i log(\tilde{y_i}) \tag{A.4}$$

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 [93], 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.

A.4 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.

A.5 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) [83, 84]. 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

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

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:

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

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 et al. (2020) and McGovern et al. (2019).

A.6 K-Means Clustering

K-Means cluster analysis [88] 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.

S1

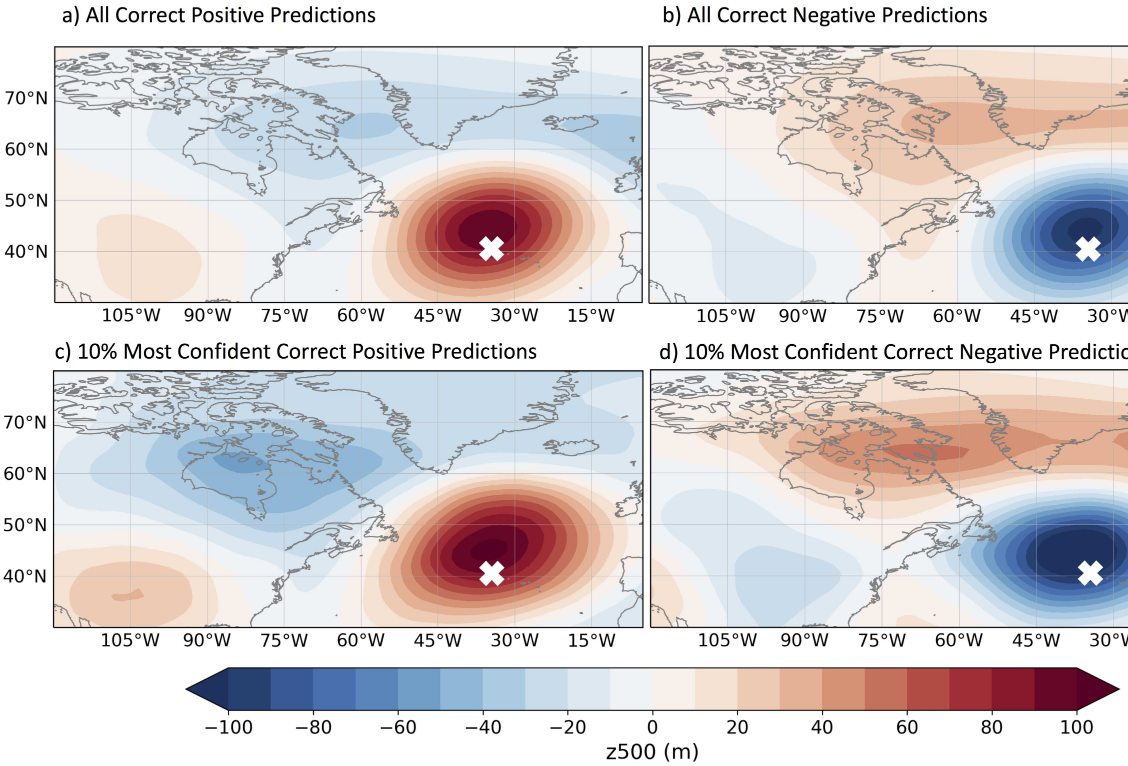

a) All Correct Positive Predictions
b) All Correct Negative Predictions
c) 10% Most Confident Correct Positive Predictions
d) 10% Most Confident Correct Negative Predictions
70°N
60°N
50°N
40°N
105°W
90°W
75°W
60°W
45°W
30°W
15°W
−100
−80
−60
−40
−20
0
20
40
60
80
100
z500 (m)

Figure A.1: 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).

S2

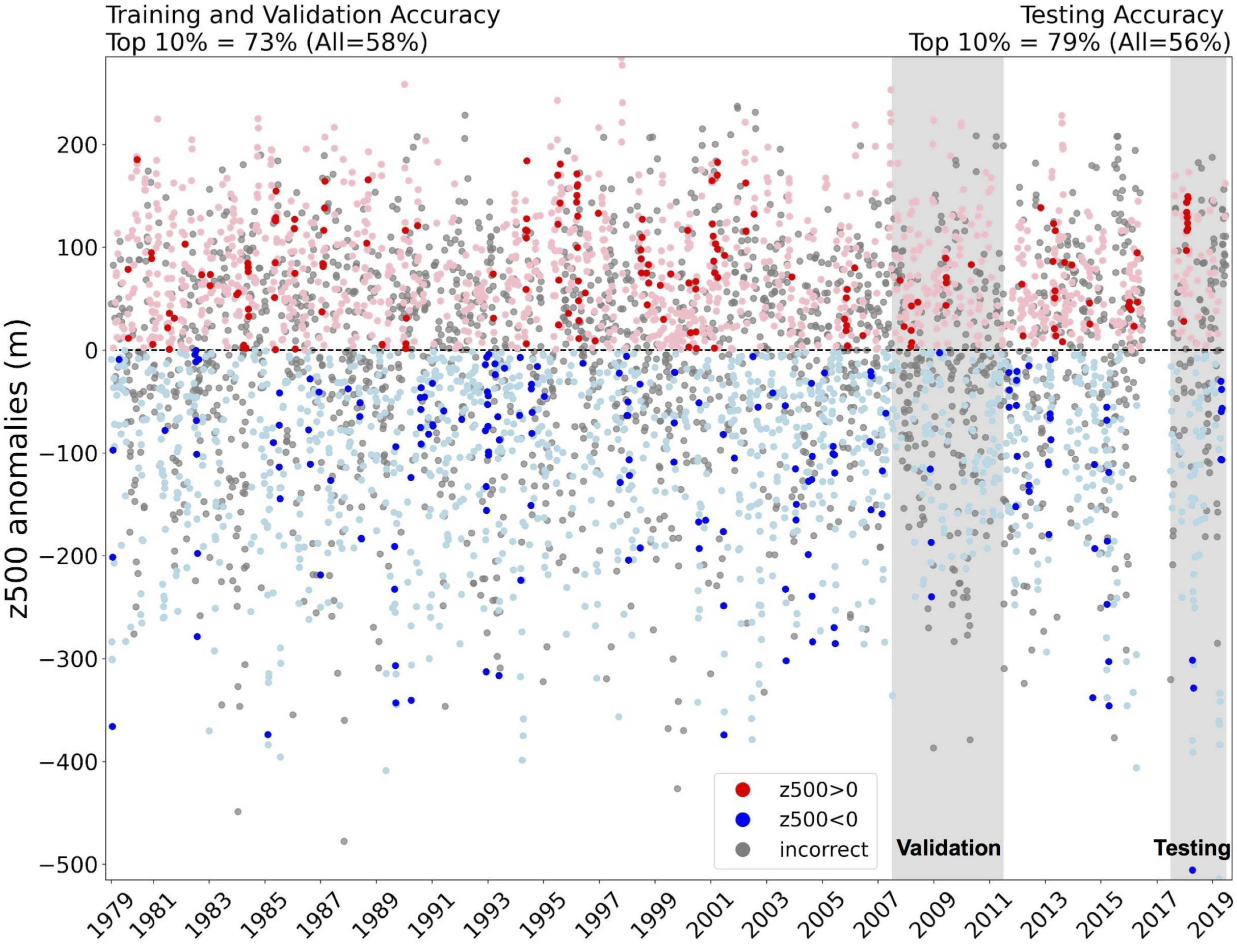


Figure A.2: Timeseries of z500 anomalies shaded by the sign of the ANN predictions. Blue dots represent correct negative predictions, red dots represent correct positive predictions, and dark colored dots indicate forecasts of opportunities (i.e. 10% most confident predictions). Grey dots represent incorrect 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.

S3

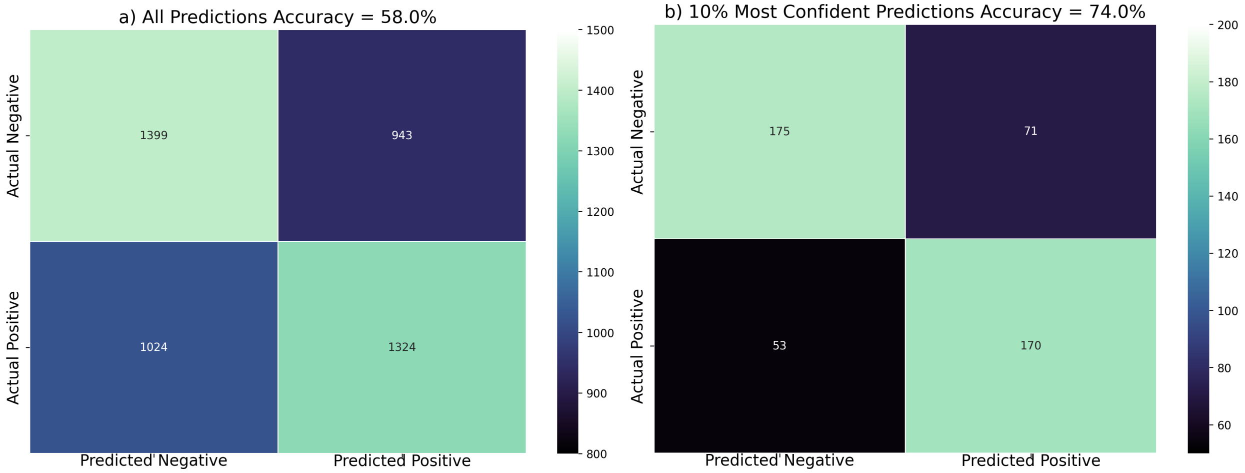
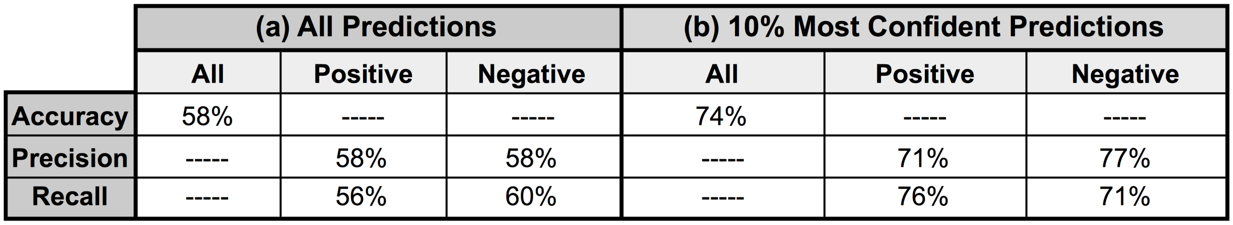


Figure A.3: Confusion matrix of training, validation, and testing data for (a) all predictions and (b) the 10% most confident predictions, where the accuracy is located at the top of each plot and the shading and the values inside each box represents the sample size for each category.

S1



	(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%

Figure A.4: Table of accuracy, precision, and recall for (a) all predictions and (b) the 10% most confident predictions using training, validation, and testing data.

Appendix B: Chapter 3 Supporting Information

B.1 Overview

In the supporting information, we provide details on the robustness of our results to changes in the number of training ensemble members and variations in the neural network architecture and hyperparameters. The sensitivity of the results to the choice of members used for validation and testing are examined as well. We also include information about the random chance and bootstrapping analysis, and provide additional information on the neural network explainability technique and the seasonal filtering results, along with the corresponding figures. Confidence versus accuracy diagrams for seasonal predictions are also included.

B.2 Network Sensitivity to the Number of Training Members

To test whether 8 training ensemble members (members #1-8) are sufficient for this analysis, 100 neural networks are trained with different sized training sets, starting with only 1 member and increasing to 8 members iteratively (moving left to right in Figure B.1). Figure B.1a,b includes the accuracies of the testing member (#10) for all predictions and Figure B.1c,d includes the accuracies for the corresponding 20% most confident predictions. In the North Pacific (Figure B.1a,c), the skill for the historical *and* future periods plateaus at about 5 training members for both all and the most confident predictions. The North Atlantic (Figure B.1b,d) shows more skill variability with training size, but generally maintains the same range of skill for each period when 3 or more ensemble members are used. The skill variability in the North Atlantic may be related to different multidecadal variability states within each ensemble member [120]. Different lower frequency/background states can ultimately impact how well and what the network learns and thus, impact the skill of the network. For example, the skill of the network is reduced for historical predictions in the North Atlantic by adding ensemble member #5 to the training dataset (Figure B.1b,d). However, the skill then rebounds to comparable values to the smaller training datasets when members #1-8 are used for training. This suggests that the network has enough data to learn about these various multidecadal predictability states simulated in the previous individual ensemble members. While the Pacific is also impacted by longer timescales, it is more prominently impacted by decadal variability instead [121].

B.3 Network Architecture and Sensitivity to Hyperparameters

The neural network used in this study consists of two layers of 128 and 8 nodes, respectively. The rectified linear unit “relu" activation function is applied to the hidden layers. Categorical cross entropy is used for the loss function and the batch size is set to 256 samples. Adam [93] is used as an optimizer with a learning rate of 0.001. We reduce the learning rate exponentially by $e^{-0.1}$ for each epoch after 10 epochs to assist the network in minimizing the loss. To reduce overfitting on the training data, ridge regression (L_2 = 1.0) [119] is applied to the first hidden layer and early stopping is implemented. Ridge regression is used to direct the network to account for spatial autocorrelation within the input field (tropical precipitation). Early stopping monitors the validation prediction accuracy, so when the validation prediction accuracy does not increase for more than 20 epochs, the network stops training and reverts back to the network weights from 20 epochs before. Otherwise, the network concludes training at 100 epochs. We find that a patience of 20 epochs is useful for this problem to reduce overfitting since the network never trains for the full 100 epochs when early stopping is implemented. The output layer consists of 2 nodes and uses the softmax activation function. The softmax activation function converts the output into two numbers which sum to 1 and can be interpreted as the likelihood of a given prediction, referred to as “model confidence”. A more detailed description of network training for a similar artificial neural network is provided in the supporting information of Mayer and Barnes (2021) and additional information on artificial neural networks in general can be found in Nielsen (2015) or Goodfellow et al. (2016).

To test the sensitivity of our conclusions to the network architecture and hyperparameter choice, the learning rate, ridge regression parameter, nodes per layer and the number or layers were all varied and the validation accuracy compared (Figure B.3-B.4). Figure B.3 (B.4) shows results for the North Pacific (Atlantic), where the validation member #9 accuracy of 10 trained models with different initial weights are shown for each hyperparameter variation and time period. The network hyperparameters and architecture for this analysis were ultimately chosen because it has some of the highest validation skill for both the historical and the future time period in the North Atlantic, but also performs well in the North Pacific (Figure B.3-B.4). We initially focus on the skill of the network in the North Atlantic because it is more difficult for the network to predict than the North Pacific. We also see that slight variations of these hyperparameters show similar skill to the network chosen.

We note that for the North Pacific hyperparameter sweep (Figure B.3), validation member #9 shows a decrease in skill for all predictions between the historical and the future period which is not seen with the testing member (Figure 2a). We believe that the decrease in skill in the validation data between the two time periods is likely a result of slight overfitting of the validation during the historical time period due to its use for early stopping (not shown).

B.4 Random Chance Analysis

To calculate random chance for each network across all confidence levels, 1000 time series are created with an equal number of 0s and 1s to represent ‘truth’. Corresponding ‘predicted’ time series are created by randomly selecting values between 0 and 1 (confidence values) and subsequently, creating time series of the predicted class (0 or 1) from these confidence values. The accuracy of these predictions is then calculated for each time series at each percent confidence threshold. The 95th percentile and below of this distribution is shaded in light blue in Figure 2.2a,c.

B.5 Accuracy Bootstrapping Analysis

Due to the computational costs of training 100 networks for each grid point in the Northern Hemisphere, 10 neural networks are trained for each location instead. To check whether these changes identified in the North Pacific and North Atlantic with 100 networks can be seen using only 10 networks, and to provide a reference of the magnitude of significant skill changes for the other grid points in Figure 3, we used the 100 models trained for both the North Pacific and North Atlantic to conduct a bootstrapping analysis.

For each location, from the 100 models trained, 10 models are randomly selected and the top three networks are chosen, defined using the three highest 20% most confident validation accuracies. The mean of the 20% most confident testing accuracies is then calculated for these three models, identical to the method used to calculate the testing accuracy for each grid point in Figure 3. This is repeated 1000 times for each time period with the resulting distributions plotted in Figure B.5. For each region, we find that the direction of change in skill for 10 networks is the same as that for 100 networks, and the future accuracy is statistically different than the historical time period using a one-sided Welch's t-test [122] at a 95% confidence level (p-value < 0.0001). For the North Pacific (Atlantic), we test whether the future period is statistically less (greater) than the historical. Therefore, we find that 10 networks is sufficient for identifying these subseasonal prediction skill changes.

B.6 Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) [83,84] is a neural network (attribution) explainability technique that creates a heatmap of the estimated “relevance" of the input for a given prediction. Here, we use the LRP_z rule, which has been shown to perform well for specific geoscience applications. This is especially true compared to methods which neglect negative preactivations (e.g. LRP alpha=1, beta=0). These methods have been shown to impact the relevance magnitude and assign the same sign relevance independent of whether they contribute positively or negatively to the final prediction [118]. For an individual prediction, LRP_z backpropagates relevance information from an output node through the network to create a heatmap of the estimated regions of the input that the network found most relevant for its prediction, where positive (negative) relevance denotes positive (negative) contributions to the final output. The softmax activation function is removed before back propagation and the heatmap for each prediction is normalized by dividing by the absolute maximum relevance value in that map. Figure B.6 shows the average of the correct and confident predictions' heatmaps for an example neural network. We find that other networks produce similar LRP maps to this example.

In both the North Pacific and North Atlantic, the differences in relevance between the two time periods are most evident in the equatorial Pacific. In the North Pacific, the relevance of this region is generally reduced and the focus shifts westward in the future period. In the North Atlantic, the relevance of this region increases in the future, mainly over the western equatorial Pacific. The change in relevance of the equatorial Pacific corresponds with the change in prediction skill for both regions and suggests that the network's changing focus in the equatorial Pacific is related to the changes in subseasonal prediction skill.

B.7 Seasonal Filtering Analysis

To further examine the possible role of seasonal variability influencing future subseasonal prediction skill, we task the neural network to predict the sign of z500 anomalies using only z500 variability on *shorter* than 60 day (subseasonal) timescales. The z500 anomalies are filtered by removing the forward 60 day running mean. This filtering is used to direct the network to focus on tropical precipitation specifically related to midlatitude subseasonal variability in the z500 anomalies. Thus, changes to prediction skill between the two time periods are a result of changes in the ability to specifically predict midlatitude variability with shorter than 60 day periods. We use this approach to identify if the changes in skill are mainly related to changes in midlatitude subseasonal variability or if the changes are related purely to changes in seasonal variability, or a combination of the two. We note that this filtering analysis could have been conducted for the main paper, however, we wanted to retain seasonal variability information to identify possible skill changes that could be seen in a typical subseasonal forecast. For each time period and region, 100 networks are again trained and their accuracies across model confidence thresholds are computed (Figure B.7).

Overall, the removal of seasonal variability reduces the information the network can use for its predictions, so the filtering leads to a decrease in skill for both time periods compared to the unfiltered predictand. In the North Pacific (Figure B.7a-b), there is virtually no difference between the historical and future period when seasonal variability is removed because the historical skill decreases more than the future skill, resulting in similar accuracies across model confidence thresholds. This implies that the historical period relies more on seasonal variability for subseasonal prediction than the future period, consistent with the LRP analysis. The lack of skill change between the two time periods also implies that the *change* in subseasonal prediction skill seen in the unfiltered analysis is related to midlatitude seasonal variability instead of subseasonal. In the North Atlantic, we see that the future period still has higher prediction skill compared to the historical, although, the overall skill for both time periods is reduced (Figure B.7c-d). A reduction in skill for both time periods is expected because the LRP maps suggest that both time periods rely, at least partially, on the ENSO regions for the predictions. However, even with midlatitude seasonal variability removed from z500, there is still an increase in skill from the historical to the future time period over the North Atlantic, suggesting there are other shorter timescale variability contributors to the increase in midlatitude subseasonal prediction skill in the future.

B.8 Seasonal Predictions

To check whether the neural networks are using more than seasonal information for their predictions, we train 100 neural networks for East Asia, the North Pacific and the North Atlantic for leads of 60 and 90 days (Figure B.8-B.9). East Asia is also analyzed here because of the unexpected increase in subseasonal prediction skill in the future (Figure 3). By training the networks at seasonal lead times, we can assess whether the prediction skill in each region *only* comes from seasonal variability. In other words, if only seasonal variability is contributing to the prediction skill, there should be no difference in skill between a lead of 21 days and a lead of 60 or 90 days.

We see that in East Asia (Figure B.8a-b, B.9a-b) the neural networks have similar skill whether trained at a lead of 21, 60 or 90 days. This suggests that the skill seen at 21 days is likely skill from seasonal variability alone. On the other hand, the North Pacific (Figure B.8c-d, B.9c-d) and the North Atlantic (Figure B.8e-f, B.9e-f) both show higher skill at a lead of 21 days, suggesting that in these regions the neural network is using more than seasonal variability for its predictions. Lastly, Figures B.8 and B.9 demonstrate that the *changes* in skill between the historical and future periods at a lead of 21 days (Figure 2.2), are similar to those for seasonal lead predictions, particularly in the North Pacific. In the North Atlantic, the change in skill is larger for the seasonal lead predictions than the 21 day lead. This implies that the networks for the 21 day lead prediction use sources of predictability other than seasonal variability to make predictions, ultimately impacting how much the skill changes between time periods. Overall, this analysis again suggests that seasonal variability is playing a role in the changes to subseasonal prediction skill, but the magnitude of the seasonal influence varies by region.

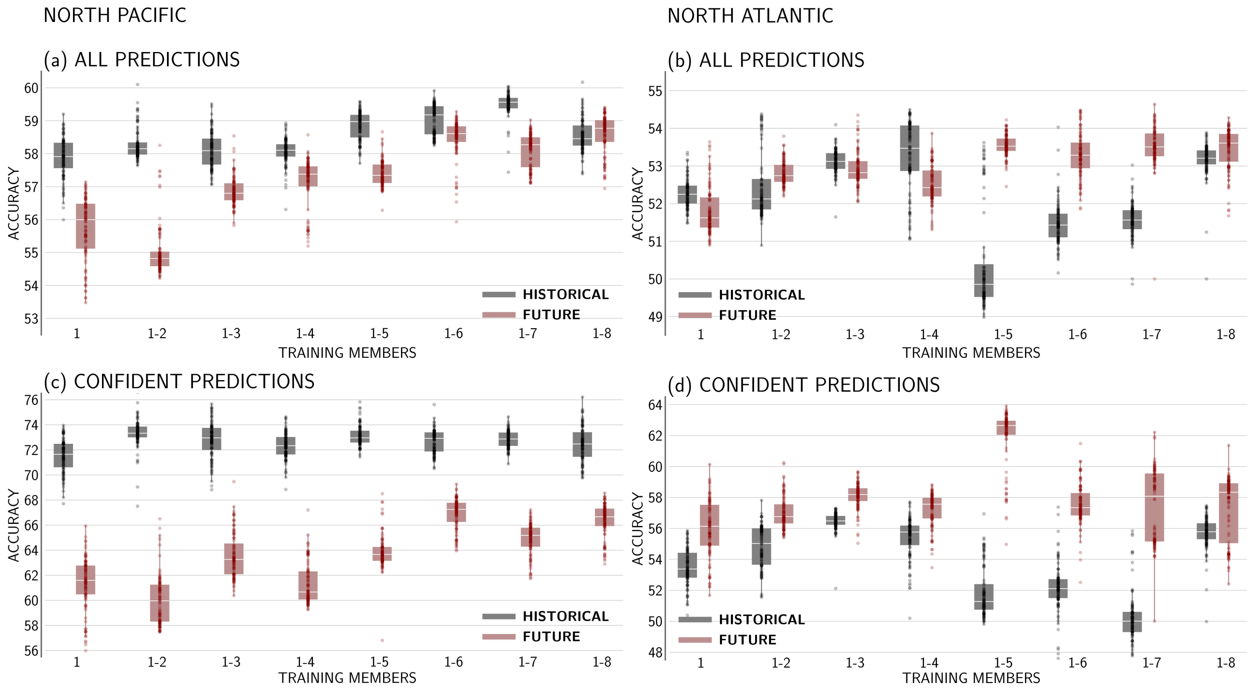


Figure B.1: Box and whisker plots of (a,b) all prediction and (c,d) the 20% most confident prediction accuracies for testing ensemble member #10 for the (a,c) North Pacific and (b,d) North Atlantic using increasing numbers of ensemble members for training. Training members #1-8 are used for the main analysis. The black (red) denotes the historical (future) period and the x-axis are the members used to train. The dots indicate individual accuracy for each of the 100 models trained. The white line across each box is the median of the models and the edges of the boxes are the 25th and 75th percentiles.

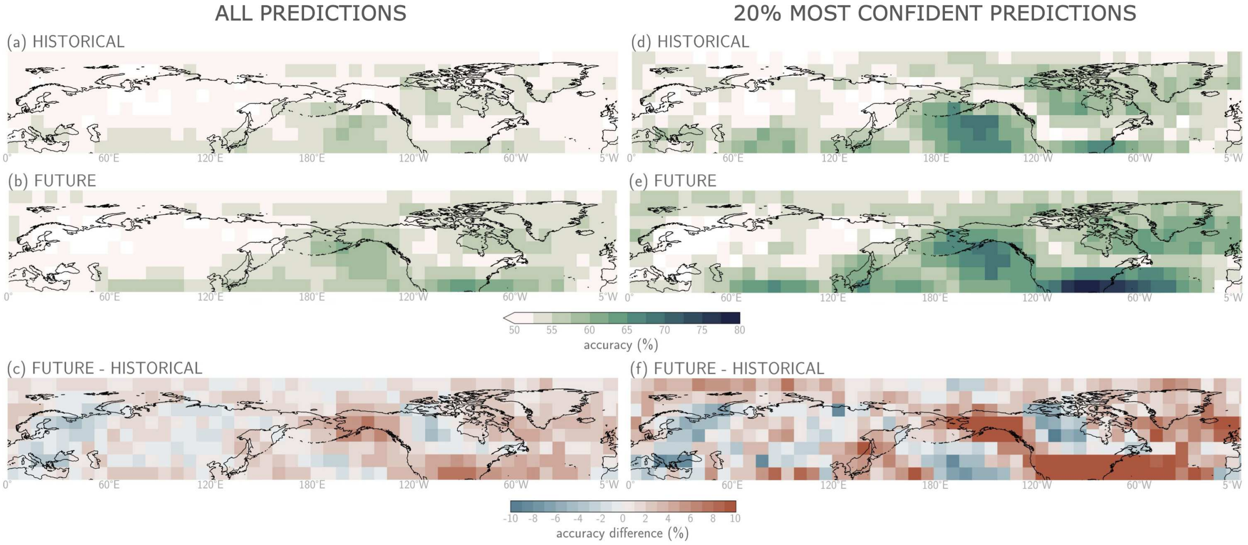


Figure B.2: As in main text Figure 3, but with ensemble members #3-10 for training, member #2 for validation and member #1 for testing.

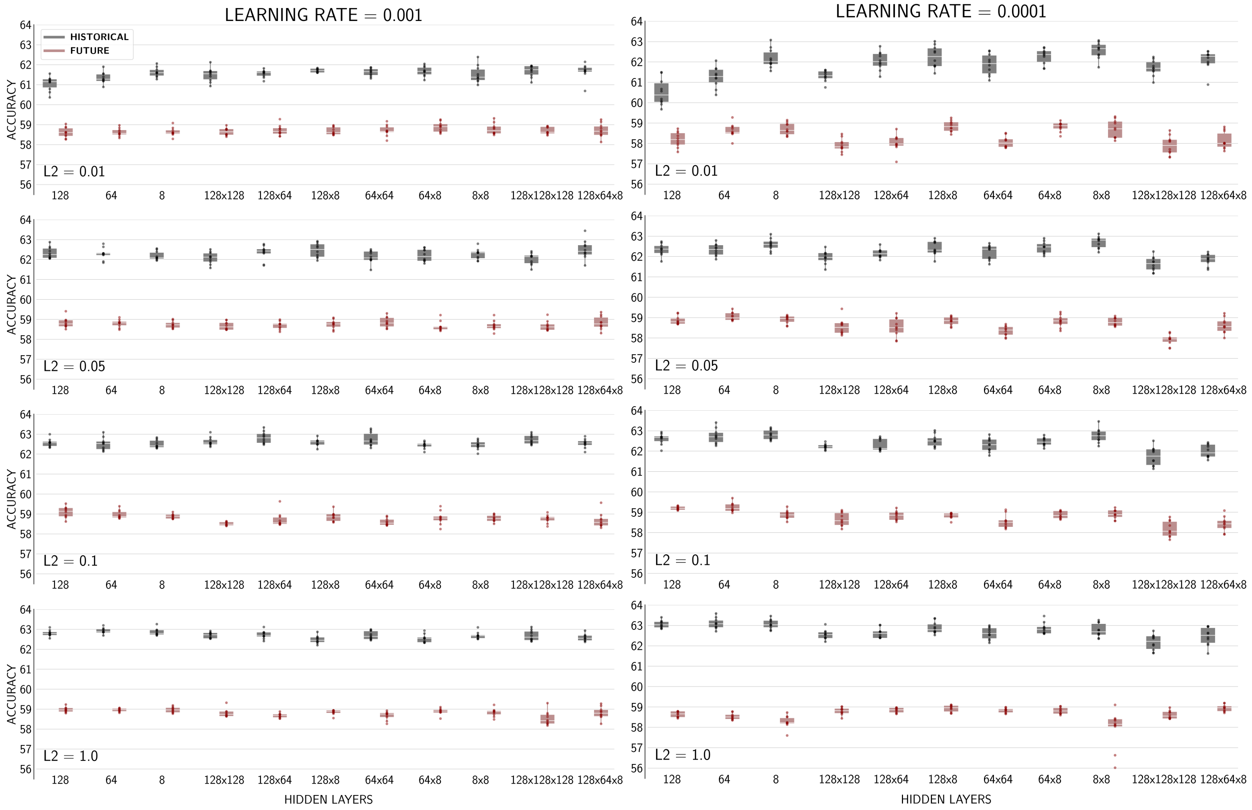


Figure B.3: Validation (member #9) box and whisker plots of accuracies for 10 trained models in the North Pacific for variations combinations of the learning rate, ridge regression (L2), nodes per layer, and number of layers. Networks accuracies for a learning rate of 0.001 (0.0001) are in the left (right) column. Ridge regression values (denoted in the bottom left of each figure) increase from top to bottom and the network depth increases from left to right, where the number(s) represent the number of nodes per layer.

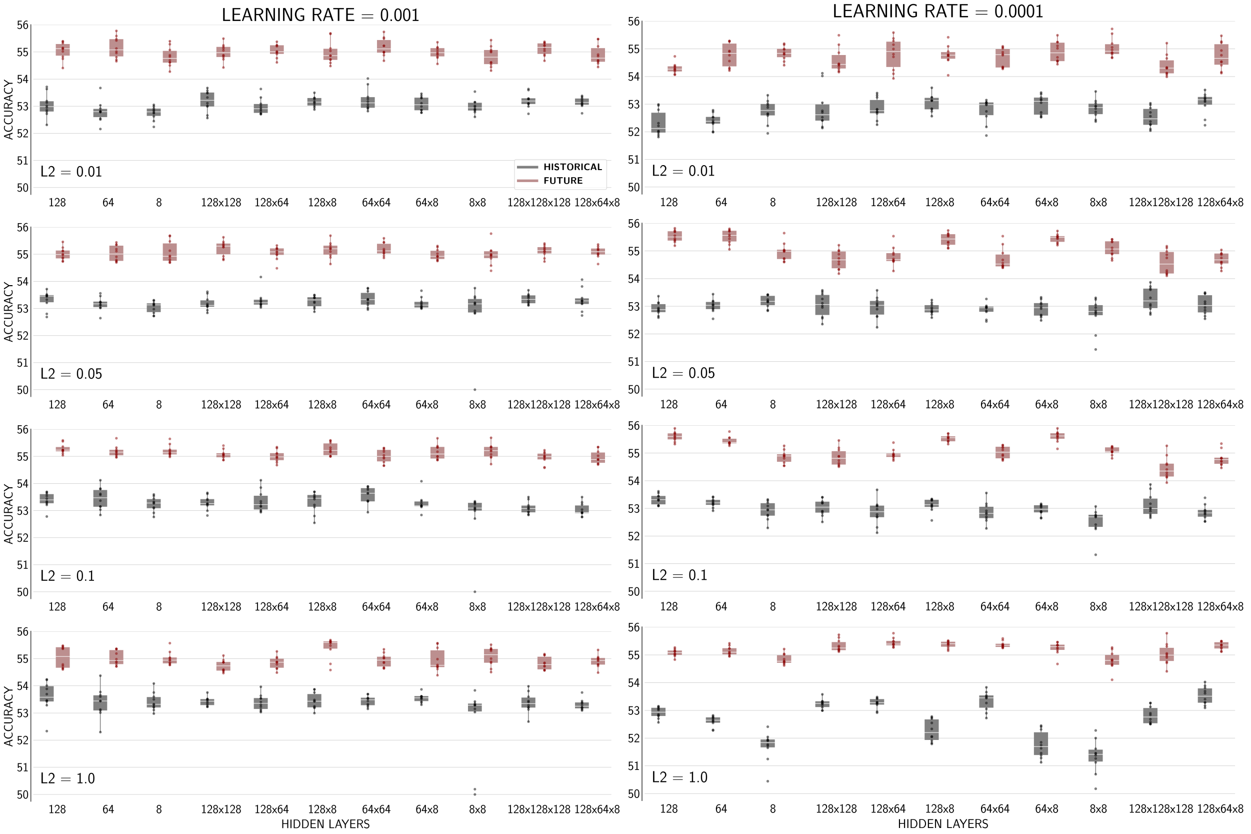


Figure B.4: As in Figure B.2, but for the North Atlantic.

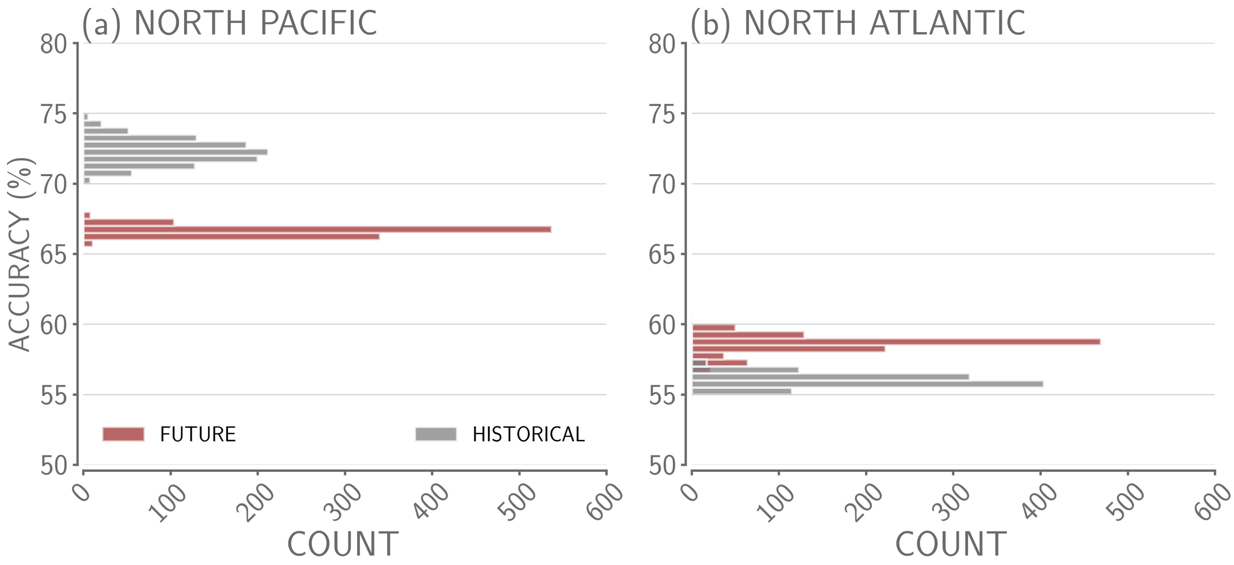


Figure B.5: Histograms of bootstrapped top 3 models' mean 20% most confident testing accuracies with a bin size of 0.5% for (a) the North Pacific and (b) the North Atlantic, where grey and red refer to the historical and future, respectively.

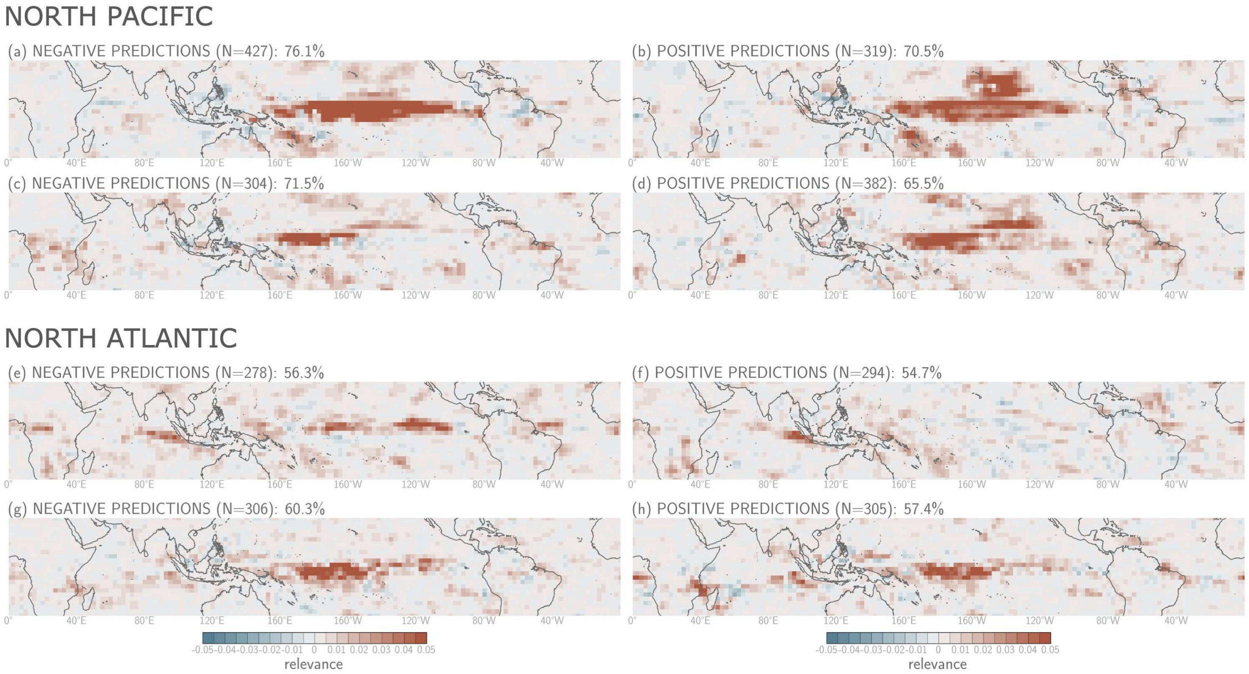

NORTH PACIFIC
(a) NEGATIVE PREDICTIONS (N=427): 76.1%
(b) POSITIVE PREDICTIONS (N=319): 70.5%
(c) NEGATIVE PREDICTIONS (N=304): 71.5%
(d) POSITIVE PREDICTIONS (N=382): 65.5%
NORTH ATLANTIC
(e) NEGATIVE PREDICTIONS (N=278): 56.3%
(f) POSITIVE PREDICTIONS (N=294): 54.7%
(g) NEGATIVE PREDICTIONS (N=306): 60.3%
(h) POSITIVE PREDICTIONS (N=305): 57.4%
0°
40°E
80°E
120°E
160°E
160°W
120°W
80°W
40°W
-0.05 -0.04 -0.03 -0.02 -0.01 0 0.01 0.02 0.03 0.04 0.05
relevance

Figure B.6: Example average layer-wise relevance plots for the 20% most confident and correct predictions in the North Pacific (a-d) and the North Atlantic (e-h). The top two panels for each locations (a-b, e-f) are the historical period and the bottom two panels for each location (c-d, g-h) are the future period. The left column includes heatmaps for the negative predictions and the right column includes heatmaps for the positive predictions. Red (blue) colors indicate the location had a positive (negative) contribution to the correct prediction. The percentage at the top of each panel is the precision of each sign prediction and 'N' is the number of samples in each average.

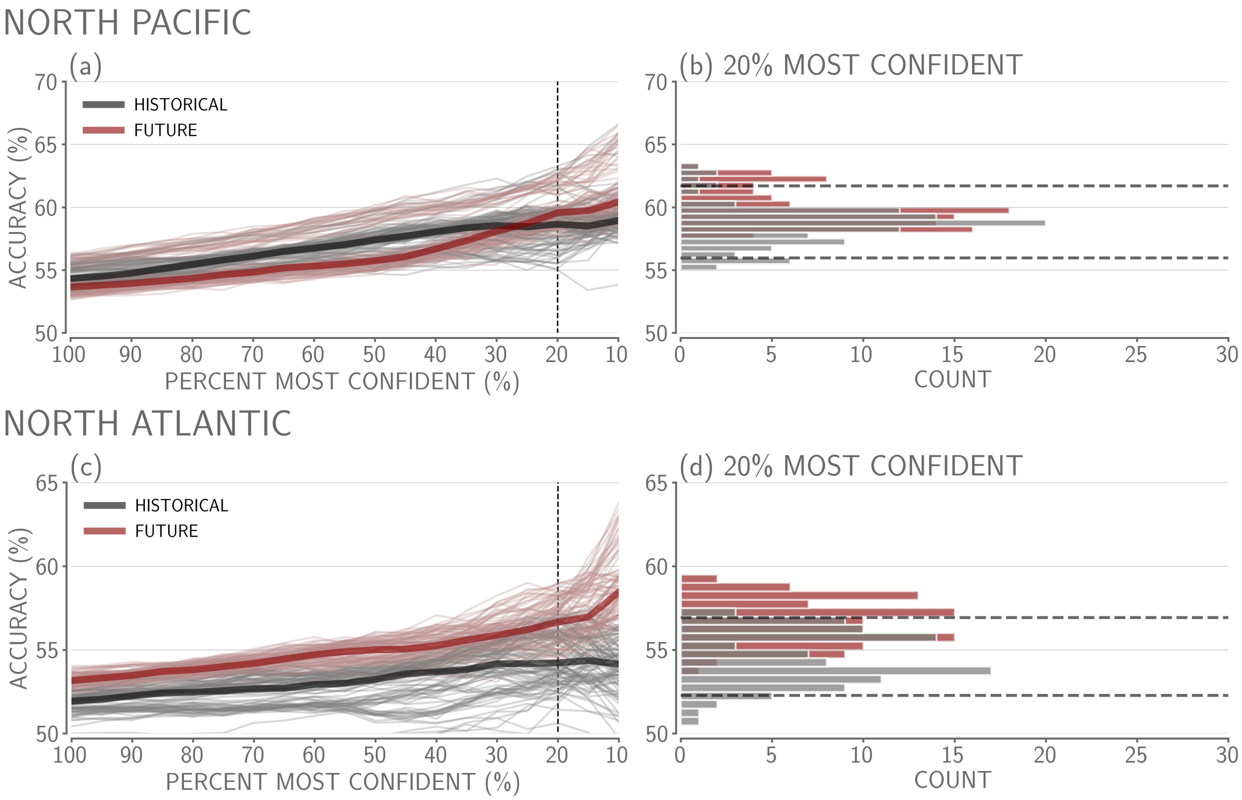


Figure B.7: As in Figure 2 in the main text, but with 60+ day z500 anomaly variability removed from the predictand.

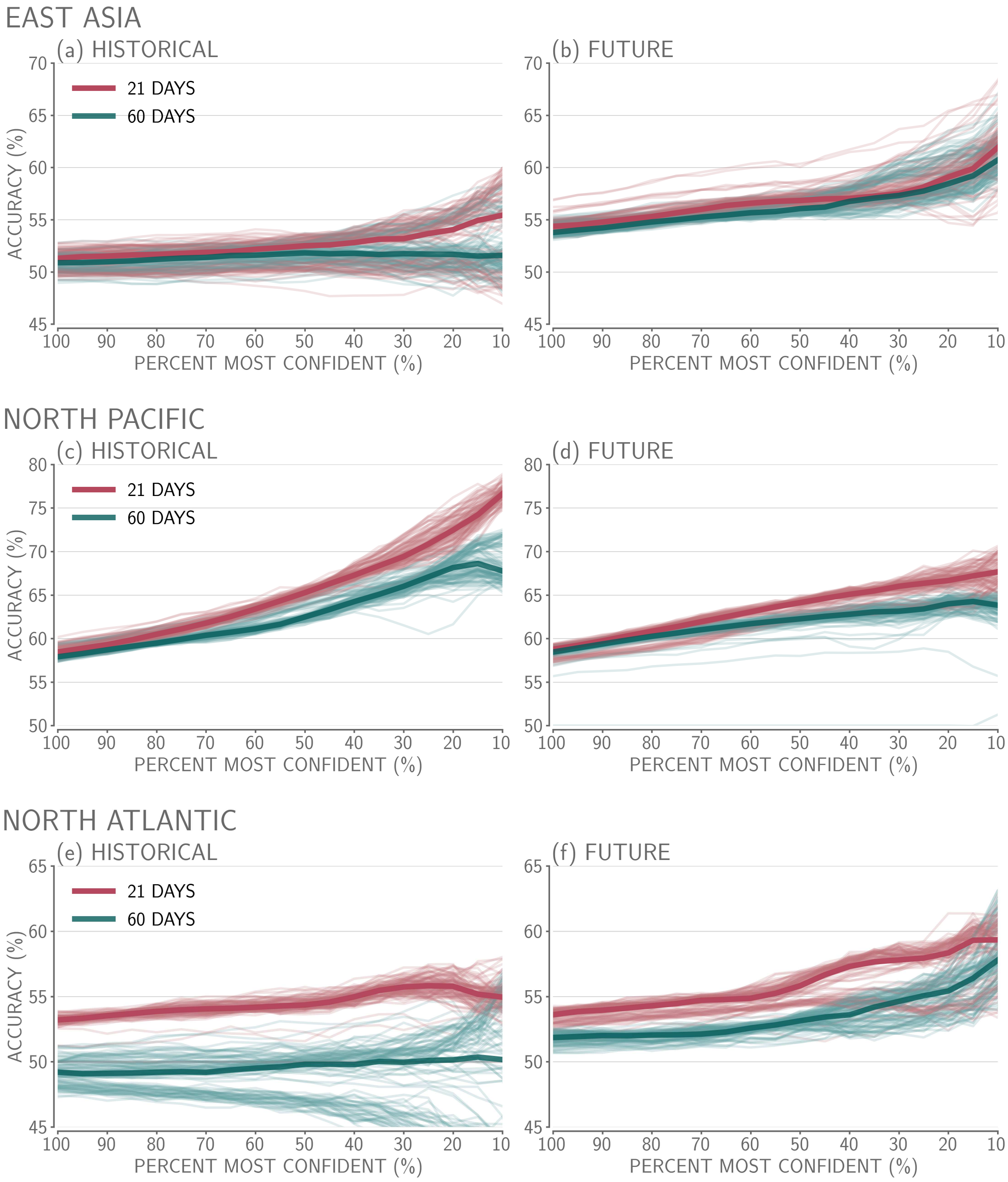


Figure B.8: Accuracy versus confidence for 100 trained networks for the (left) historical and (right) future time period at leads of 21 (pink) and 60 (teal) days in (a,b) East Asia, (c,d) the North Pacific and (e,f) the North Atlantic. Accuracies are calculated using the testing member #10 and the thicker lines denote the median accuracy across the 100 networks at each confidence threshold. The pink lines are the same as the red/grey lines included in Figure 2 for the respective location and time period.

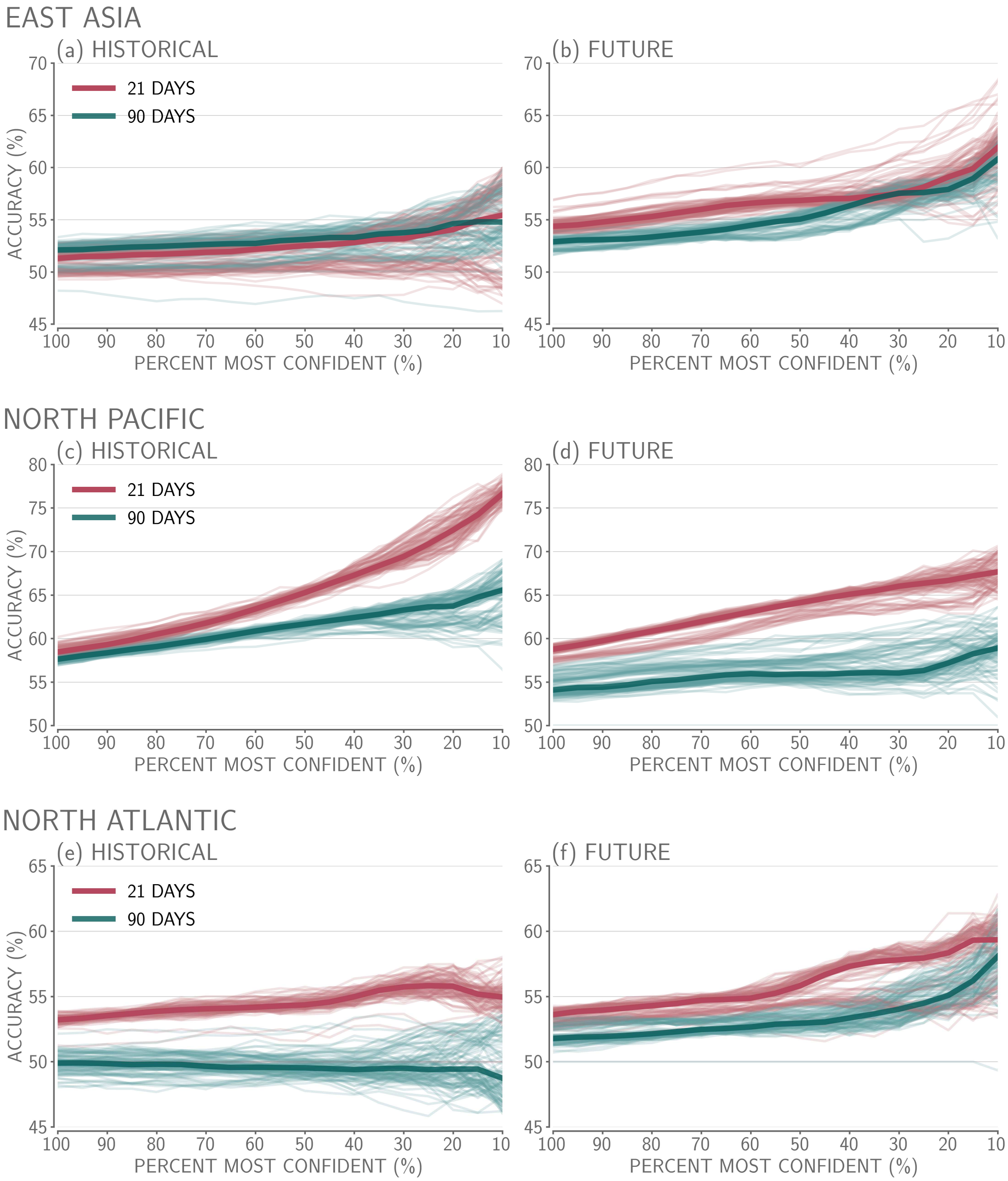


Figure B.9: As in Figure B.8, but for a lead of 90 days.

Appendix C: Chapter 4 Supporting Information

C.1 Overview

In the supporting information, we provide details on the hyperparameter sweep applied to identify the final neural networks used in the predictability analysis as well as further information on the final neural networks' hyperparameters selected. In addition, we provide the details of the significance tests for the variability, ENSO teleconnection, and prediction analyses, and a figure depicting the relationship between (confident) predictions and ENSO events.

C.2 Hyperparameter Sweep and Network Architectures

To find the final neural network architectures, a hierarchical hyperparameter selection is implemented for each scenario. An initial hyperparameter sweep is conducted to find a base neural network to begin the hyperparameter selection. This base network consists of one hidden layer with 8 nodes, a learning rate of 0.001, and a ridge regression parameter of 0.25. For each hyperparameter and value, cross validation is conducted and the hyperparameter value that has the highest validation accuracy across the most validation members is selected. Using the base network architecture, we first tune to find an ideal batch size using this selection method. After selecting the batch size and changing the base network accordingly, the same method is then systematically applied to find the learning rate, the number of hidden layers and nodes per each layer, and the ridge regression parameter. Below includes a list of hyperparameters examined for this analysis.

Batch Size: ***32***, 64, 128

Learning Rate: 0.0001, ***0.001***, 0.01

Hidden Layers: 4, *8*, 16, 8x4, 16x4, **16x8**

Ridge Regression: ***0.25***, 0.5, 1.0

Bold values indicate the final hyperparameters used for the SAI scenario and the italicized values indicate the final hyperparameters used for the SSP2-4.5 scenario. However, we find that the accuracy of the network only changes on the order of a couple of percentages on average across values within each hyperparameter tested.

For both network architectures, we use categorical cross entropy for the loss function and Adam [93] as the optimizer. The rectified linear unit 'relu' is applied to the hidden layer(s), where the specified ridge regression parameter [119] is applied to the first hidden layer to force the network to account for spatial autocorrelation in the input SST field. In addition, early stopping on the validation data is used to reduce overfitting to the training data. In particular, the network monitors the validation loss and if the loss does not decrease for more than 20 epochs, it reverts back to the network weights from 20 epochs prior. The learning rate is also decreased by 10% for each epoch after 10 epochs to help the network minimize loss.

C.3 Variance Significance Test

To determine whether the ratio of monthly boreal winter variance between the SAI and SSP2-4.5 scenario is statistically significant (Figure 4.3c), we conduct a bootstrapping analysis. The 2m temperature data for both scenarios are shuffled together, and then, this shuffled data is randomly split into 'SAI' and 'SSP2-4.5' months (without replacement). The monthly variance of these two randomly drawn samples is calculated and then this process is repeated for each ensemble member. Finally, the ensemble mean is calculated. We repeat this analysis 500 times, and the ratios of 'SAI' to 'SSP2-4.5' of these 500 variances is taken. Any value below (above) the 2.5th (97.5th) percentile of this resulting ratio distribution for each grid point is considered significant.

C.4 ENSO Teleconnection Significance Test

To calculate whether the difference in ENSO teleconnection consistency between the SAI and SSP2-4.5 scenarios is statistically different (Figure 4.4c,f), we apply a two-sided difference of

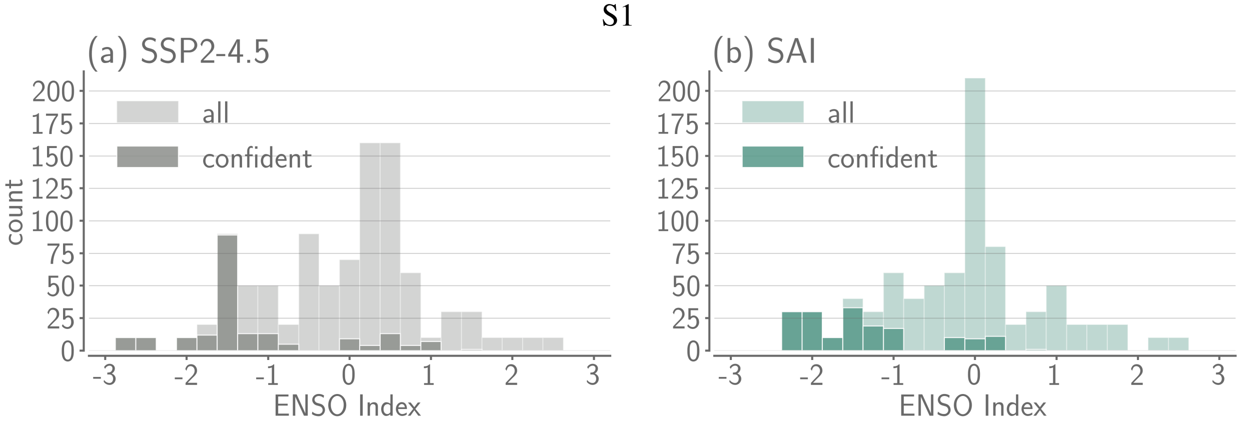


Figure C.1: Histograms of number of (confident) predictions across the ENSO Index for (a) SSP2-4.5 and (b) SAI predictions using a bin size of 0.25σ. The light (dark) shading indicates all (20% most confident) predictions.

means t-test. To calculate the ENSO teleconnection consistency in the main analysis, we calculate a frequency of a positive sign anomaly 2 months following an ENSO event for each of the ten ensemble members before taking the ensemble mean, and therefore, each scenario mean includes ten samples. Where a grid point's t-value exceeds the critical t-value (degrees of freedom = 18 and significance level of $\alpha = 0.05$), the difference is deemed significant.

C.5 Prediction Skill Significance Test

To calculate whether the mean 20% most confident accuracy under SAI is larger than that under SSP2-4.5, we use a one sided difference of means t-test. There is little variability in skill across network seed, and therefore, for each ensemble member, we treated the different network seeds as *dependent* samples. As a result, we use an effective sample size of 10 (the number of ensemble members). Using a 97.5th confidence interval, we are able to reject the null hypothesis that the mean accuracy under SAI is less than or equal to that under SSP2-4.5.