THESIS

# ERRORS OF OPPORTUNITY: USING NEURAL NETWORKS TO PREDICT ERRORS IN THE UNIFIED FORECAST SYSTEM (UFS) ON S2S TIMESCALES

Submitted by

Jack Cahill

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2023

Master's Committee:

Advisor: Elizabeth A. Barnes
Co-Advisor: Eric D. Maloney

Matthew Ross

ABSTRACT

ERRORS OF OPPORTUNITY: USING NEURAL NETWORKS TO PREDICT ERRORS IN THE UNIFIED FORECAST SYSTEM (UFS) ON S2S TIMESCALES

Making predictions of impactful weather on timescales of weeks to months (subseasonal to seasonal; S2S) in advance is incredibly challenging. Dynamical models often struggle to simulate tropical systems that evolve over multiple weeks such as the Madden Julian Oscillation (MJO) and the Boreal Summer Intraseasonal Oscillation (BSISO), and these errors can impact geopotential heights, precipitation, and other variables in the continental United States through teleconnections. While many data-driven S2S studies attempt to predict future midlatitude variables using current conditions, here we instead focus on post-processing of the National Oceanic and Atmospheric Association's (NOAA) Unified Forecast System (UFS) to predict UFS errors. Specifically, by looking at when/where there are errors in the UFS, neural networks can be used to understand what atmospheric conditions helped produce these errors via explainability methods. Our 'Errors of Opportunity' approach identifies phase 4 of the MJO and phases 1 and 2 of the BSISO as significant factors in aiding UFS error prediction across different regions and seasons. Specifically, we see high accuracy for underestimates of geopotential heights in the Pacific Northwest during Spring and as well as high accuracy for overestimates of geopotential heights in Northwest Mexico during Summer. Furthermore, we demonstrate enhanced error prediction skill for overestimates of Summer precipitation in the Midwest following BSISO phases 1 and 2. Most notably, our findings highlight that the identified errors stem from the UFS's failure to accurately forecast teleconnection patterns.

TABLE OF CONTENTS

# Chapter 1

# Introduction

For much of the modern forecasting era, significant improvements in prediction skill have been made on forecast lead times of less than two weeks and seasonal timescales (Alley et al., 2019; Klemm and McPherson, 2017). Unfortunately, the subseasonal to seasonal (S2S) timeframe is characterized by a prediction "gap" (Mariotti et al., 2018). Despite this, decision-making on S2S timescales is crucial in sectors such as public health, energy, and water management (White et al., 2017; Vitart et al., 2012). In recent years, considerable efforts have focused on bridging this gap and improving S2S predictions (Stan et al., 2022; Zhang et al., 2022; Merryfield et al., 2020; Pegion et al., 2019; Vitart et al., 2017).

One important aspect of S2S prediction lies in tropical-extratropical teleconnections. These teleconnections arise from the exchange of energy across large spatial and temporal scales, linking meteorological events together (Bjerknes, 1969). Many teleconnections affecting North America originate from processes in the tropical Pacific Ocean, such as the Madden Julian Oscillation (MJO) and El Niño Southern Oscillation (ENSO) (Cassou, 2008; Alexander et al., 2002). The MJO refers to the large-scale pattern of enhanced and suppressed convection that travels eastward across the tropical Indian and Pacific Oceans, varying on 30 to 60 day periods (Madden and Julian, 1971; 1972). In comparison, ENSO is an atmosphere-ocean coupled phenomenon denoted by periodic warming and cooling of the tropical East Pacific, occurring on multi-year timescales (Rasmusson and Wallace, 1983; Philander, 1983). Among their many effects, the MJO and ENSO can modulate precipitation frequency and intensity and temperature across the U.S. (e.g. Larson

et al., 2022; Cassou 2008; Stan et al., 2017). Teleconnections associated with the MJO and ENSO provide opportunities for enhanced prediction skill but also create additional avenues for forecast errors (Lee et al., 2023). To achieve accurate predictions, it is essential to successfully forecast the propagation of the oscillations as well as their associated atmospheric impacts, which span thousands of kilometers (Dias et al., 2019). Errors present at initialization can grow and propagate, massively impacting the accuracy of a forecast (Lorenz, 1969).

Through the lens of teleconnection prediction, "forecasts of opportunity" (FOOs) aim to improve S2S predictability by identifying specific geophysical conditions that contribute to enhanced forecast skill (Mariotti et al., 2020). By identifying these conditions, forecasts can be adjusted and improved. For example, enhanced convection associated with the MJO is responsible for the formulation of quasi-stationary Rossby waves that propagate into the midlatitudes and influence atmospheric circulations (Hoskins and Ambrizzi, 1993). The modulation of midlatitude circulations by the MJO is consistently observed on S2S timescales (Baggett et al., 2017, Sardeshmukh and Hoskins, 1988, Hoskins and Karoly, 1981). Particularly, following phase 2 when enhanced convection is most prevalent over the Indian Ocean, the MJO consistently modulates 500 hPa geopotential height (h500) anomalies in the North Pacific (Tseng et al., 2018). Moreover, following enhanced convection in the central Pacific (MJO phase 7), the prevalence of East Pacific atmospheric blocking increases (Henderson et al., 2016). Accurately identifying MJO phases and their intensity can result in enhanced forecast skill, serving as a prime example of a forecast of opportunity.

Over the past few decades, artificial neural networks (ANNs) have emerged as powerful tools in meteorology, showcasing forecast skill across a variety of atmospheric and oceanic variables (Hsieh, 2009; Zheng et al., 2021). While traditional forecast methods often rely on physical parameterizations and statistical linear methods to approximate complex interactions, ANNs excel in

the estimation of both linear and nonlinear spaces (Leutbecher and Palmer, 2008; Chen and Chen, 1995). Leveraging ANNs to predict meteorological quantities is not a new concept. In fact, ANNs have been forecasting precipitation, temperature, and wind speed for decades (e.g. Glahn 1964; French et al., 1992; Mihalakokou et al., 1998; Maqsood et al., 2004), and have been applied to seasonal forecasts, predicting monthly temperatures, yearly sea surface temperatures, and the ENSO (Toms et al., 2020; Krasnopolsky et al., 2010; Ham et al., 2019). In recent years, ANNs have found success in the S2S range as well, forecasting the MJO as well as pressure, temperature, and precipitation out to six weeks (Martin et al., 2022; Weyn et al., 2021; Vitart and Robertson, 2018; Cohen et al., 2019). Most notably, ANNs and explainability techniques have been leveraged to locate forecasts of opportunity on S2S timescales, specifically for pressure and temperature (Mayer et al., 2021, Breeden et al., 2022).

Much like how forecasts of opportunity help identify earth system conditions that enhance forecast skill, it is possible to identify specific conditions that propagate errors in forecast models. For example, soil moisture has been identified as a source of 2-meter temperature overestimates in late Summer for the European Centre for Medium-Range Weather Forecasts (ECMWF) (Dutra et al., 2021). However, the application of ANNs to predict meteorological errors on S2S timescales is a relatively recent strategy (Rasp and Lerch, 2018; van Straaten, 2023). In this study, we predict errors of the National Oceanic and Atmospheric Association's (NOAA) Unified Forecast System (UFS) using ANNs. We then analyze the most confident predictions in an attempt to identify the most predictable errors, a strategy we coin as "errors of opportunity" (EOO). To better understand these errors, we utilize explainable AI (xAI) methods and more standard climate composites to identify the meteorological conditions associated with the error source (Mamalakis et al., 2020). This approach allows us to gain an enhanced comprehension of where the errors occur and how

they propagate through space and time. As a result, we can identify situations where the forecast model can be improved, while also potentially gaining insights into methods to improve their performance.

# Chapter 2

# Data and Methods

## 2.1 Data

We use data from the Global Ensemble Forecast System (GEFSv12), an uncoupled version of the UFS released in September 2020 (Guan et al., 2022). Variables for the control run were output every 6 hours at 00, 06, 12, and 18 UTC, which we then average to create daily means. The reforecasts were initialized every 7 days, beginning on 01/05/2000 and ending on 12/18/2018, and integrated out to 35 days, resulting in a total of 1042 samples. Additionally, GEFSv12 daily-averaged reanalysis data is utilized over the same time frame. The daily reanalysis covers the entire 1042 weekly reforecast period, providing us with a total of 7305 reanalysis samples that we define as "truth". Reforecasts and reanalysis data are referred to as UFS forecasts and observations, respectively, throughout the rest of this paper.

Variables used in this study include outgoing longwave radiation (OLR), geopotential heights at 500mb (h500), and precipitation, all with 0.5 by 0.5 degree grid spacing. The seasonal cycle is removed from both the UFS forecasts and observations for all variables. For the observations, we compute the daily anomalies by subtracting the day-of-year climatological average from each value at each grid cell, allowing us to capture any deviations from the seasonal cycle. To ensure that the analysis focuses on subseasonal features, we smooth the computed seasonal cycle by removing the first two harmonics at each grid cell. The same approach is also applied at each lead time of the UFS forecast to facilitate the removal of the seasonal cycle as well as the day-of-year mean model

bias. However, each UFS forecast is not initialized on the same day of the year. To account for this, a "forecast day of year" approach is employed. Under this approach, the climatology for each sample is computed using all other samples initialized within a centered 6 day window around the forecast day-of-year. For instance, if a sample is initialized on January 5th, its climatology is calculated using samples initialized between January 2nd and 8th, without regard to the specific year. This strategy is applied separately for each lead time, resulting in a unique climatology for each lead calculated using twenty values, representing the twenty years of available data.

The neural networks take tropical OLR data as the input and predict h500 and precipitation errors. The OLR inputs span the Tropical Indian and Pacific Oceans from -25S to 25N and 60E to 60W. Before being fed into the neural networks, the OLR inputs undergo normalization to minimize estimation errors and training time. Each input is normalized by subtracting the mean and dividing by the standard deviation across the entire input region. The predictands are calculated by subtracting the h500 and precipitation forecasts from their respective observations, resulting in an error averaged over lead times of 10-14 days. The 10-14 day period represents the typical time it takes an Indo-Pacific warm pool tropical teleconnection to reach North America and the North Pacific (Cassou et al., 2008; Henderson et al., 2016; Stan et al., 2017; Dai et al., 2017).

Samples are classified into three distinct classes: UFS underestimates, UFS precise estimates, and UFS overestimates, all as a function of the day-of-year. Due to a limited number of samples, each individual sample is classified using a centered window of 30 calendar days. The 80 errors that occur in each window are then ordered sequentially and split into three classes of equal size. The classification of each individual sample is determined by these bounds. This approach ensures that each class contains an equal number of samples and that the classification thresholds vary as a function of the seasonal cycle, enabling effective analysis and evaluation of our neural networks.

Locating when and where the neural networks accurately predict classes represents a crucial step of our EOO process.

## 2.2 ANN Architecture

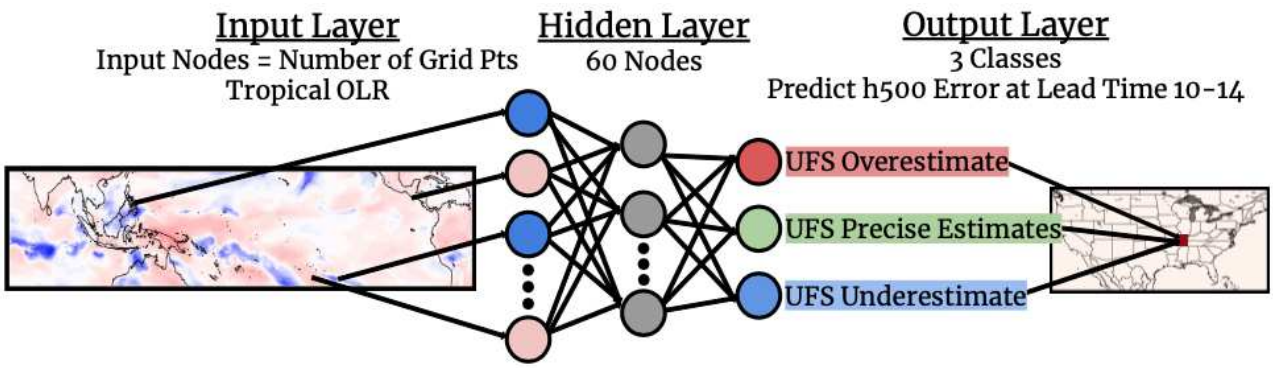


**Figure 2.1:** Architecture of the artificial neural network for the prediction of geopotential height errors over various continental United States locations at an averaged 10-14 days following OLR anomalies in the tropical Pacific. The network consists of one hidden layer of sixty nodes and one output layer of three nodes representing three types of errors: underestimates, precise estimates, and overestimates.

Figure 2.1 illustrates the architecture of the neural networks trained to classify the UFS error. The input layer consists of OLR across the tropics resulting in a total of 48,581 vectorized and normalized grid points/input nodes. To account for the large number of input nodes we apply a "dropout layer" following the input layer (Srivastava et al., 2014). Our dropout layer randomly excludes 30% of the input nodes for each epoch during training. The inputs are then passed through a singular hidden layer of 60 nodes using a rectified linear unit as the activation function, followed by an output layer of three nodes, representing the three classes of error: UFS underestimates, UFS precise estimates, and UFS overestimates. Finally, a softmax activation function is applied to the output layer which remaps the values of the three-node output such that they sum to one. The largest value of the three nodes is then defined as the network's predicted class. Additionally, we

associate the value of the winning class with its predicted value, which we call "model confidence". For instance, we interpret an output of 0.9 to be more confident than a value of 0.6. The batch size is set to 32 and the neural network is run for 100 epochs with a learning rate of 0.0001.

The network is trained on 16 years of data (834 or 833 samples depending on the presence of leap years) and the remaining 4 years of data (208 or 209 samples) are used as testing data. In an attempt to ameliorate potential issues of a small testing set, a five-fold cross-validation technique is employed. This approach incorporates each consecutive testing fold, resulting in 1042 testing samples for analysis. Figure S1 demonstrates our cross-validation method. To ensure robustness, the neural network is run for six different random initialization seeds of starting weights for each training-testing fold. Unless otherwise stated, our analysis shown here is performed solely on the testing data averaged across all cross-validation folds and seeds. This average is computed after all folds and seeds are run. This setup is applied to all 156 grid points across the North Pacific and continental United States such that each location is trained using 30 different networks.

## 2.3 Explainability via Integrated Gradients

Neural networks are frequently criticized for their lack of interpretability, as their predictions are often not easily understood by humans (Goodman and Flaxman, 2017; Adadi and Barrada, 2018). To address this issue, various explainable artificial intelligence (xAI) methods have been developed (Bach et al., 2015; Lundberg and Lee, 2017; Montavan et al., 2018; Shrikumar et al., 2017). One such method, 'Integrated Gradients', provides a valuable strategy to attribute feature importance in neural networks and has demonstrated successful reconstruction of ground truths in past climate prediction problems (Mamalakis et al., 2022). Specifically, this method integrates the network's gradients along a path from a chosen 'baseline' to input. The baseline serves as a

reference point to compare to when attributing feature importance. Since our inputs are normalized anomalies and naturally centered around zero, using a baseline of zero can appropriately capture relevant features (Sundararajan et al., 2017). By integrating to the input, Integrated Gradients quantifies the contribution of each input feature to the output via an attribution score which is then reconstructed to create a map of relevance. However, the resulting relevance map can be noisy, and so, we apply a Gaussian filter with a sigma of 2 to the positive gradients when presenting the results (Figure 3.5b and Figure 3.7b).

# Chapter 3

# Results

Throughout this study we leverage neural networks to predict UFS forecast errors and investigate their origins using xAI methods. We focus on the most confident predictions as they provide valuable opportunities to uncover underlying mechanisms driving forecast errors. Therefore, we coin this approach, "errors of opportunity" (EOO). By pairing this strategy with the analysis of the progression of the predicted variable, we gain insights into why these errors occur and how they develop within UFS. By doing so, we can gain an enhanced understanding of the meteorological states associated with errors, facilitating the potential identification of where forecast adjustments are needed.

## 3.1 H500 error predictions across the North Pacific and Continental United States

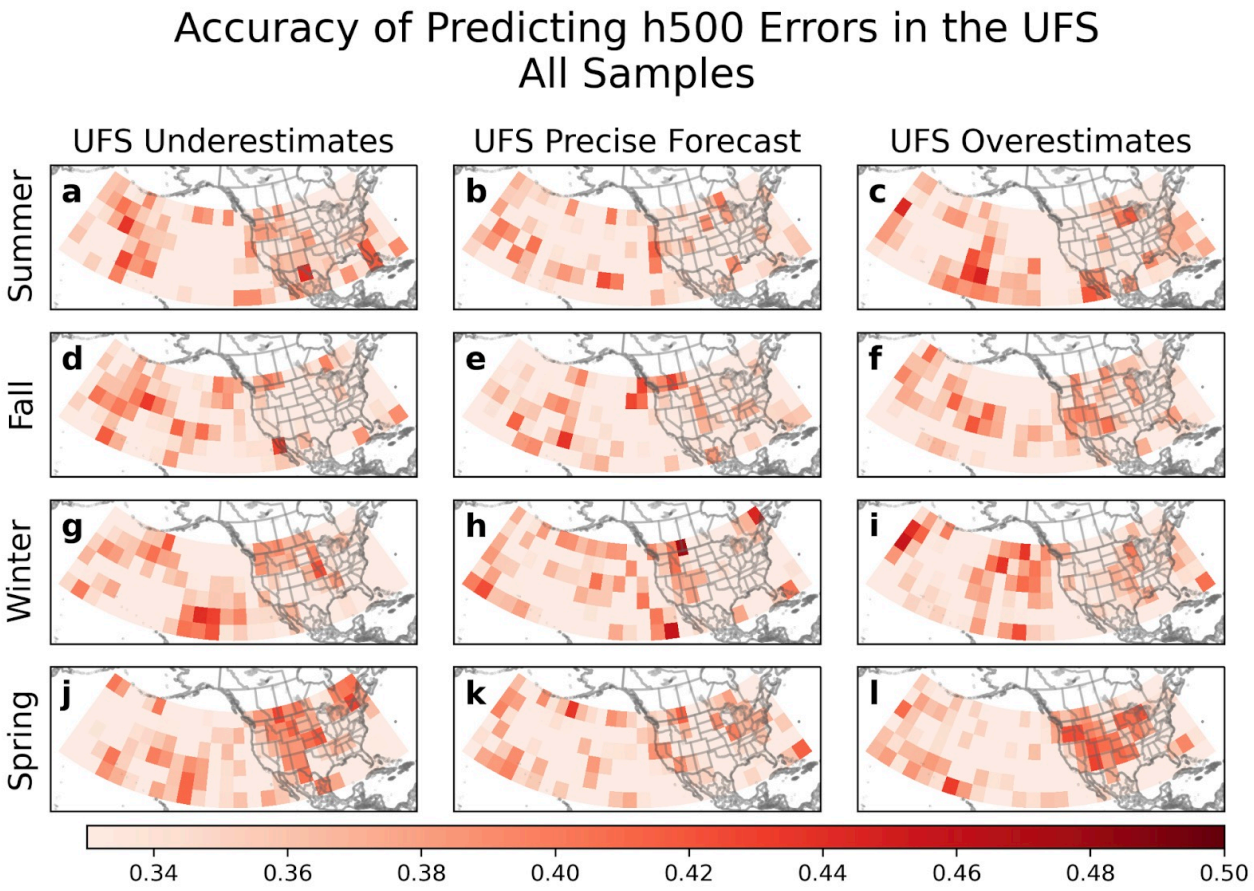


**Figure 3.1:** 500-hPa geopotential height error prediction accuracy at lead time 10-14. As constructed, random chance would yield an accuracy of 0.33 when averaged across all seasons.

Neural networks with the architecture presented in Figure 2.1 are trained to predict h500 errors with a 10-14 day lead across the North Pacific and continental United States. The overall accuracy of the testing data, split by season and true class, is shown in Figure 3.1. The accuracy is computed by counting the number of correct predictions across all samples and dividing that by the total number of predictions. With 1042 testing samples and six seeds, this means that the accuracy at each location is calculated across 6252 predictions. Given an approximately equal distribution of classes and days in a season, the accuracy across each subfigure of Figure 3.1 is calculated using roughly the same number of samples. However, since classes are defined on a location by location basis, the specific samples used in each accuracy calculation vary by grid cell. By constructing

a three-class network, a neural network of pure random chance would have an accuracy of 33%, thus the lower accuracy limit in Figure 3.1. Notably, certain seasons and classes (Figures 3.1a,g,j,l) exhibit spatially robust accuracies greater than 33%. For instance, in Figure 3.1l there exists consistent and above random prediction skill across much of the Midwest and western United States during Spring. Since h500 is dominated by synoptic-scale variations, spatial coherence of a similar scale is not surprising. In fact, due to their above random accuracy and spatial robustness, we consider these situations to be of particular interest and serve as the focus of our analysis.

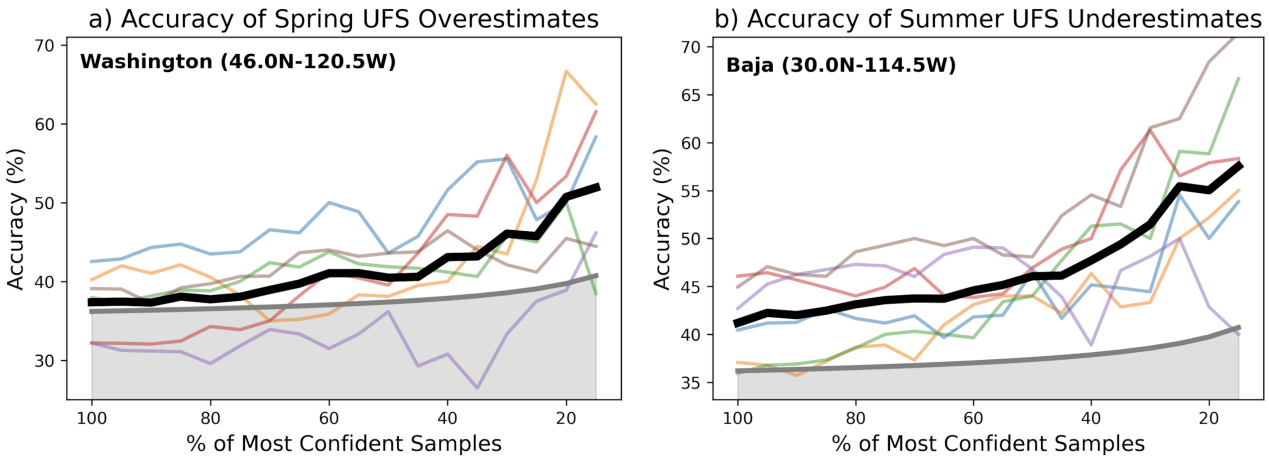


**Figure 3.2:** Neural network accuracy as a function of confidence for a singular location in (a) Washington State (46°N - 120.5°W) and (b) Baja California (30°N - 114.5°W). Light colored lines represent the six neural networks with cross-validation included, the thick black line represents the neural network average across the six networks, and the gray region represents the 95% confidence interval based on random chance using the number of samples at each 'percent most confident'.

Regions of strong spatial robustness shown in Figure 3.1l and Figure 3.1a are further examined in Figure 3.2. Specifically, Figure 3.2a and Figure 3.2b display the accuracy versus confidence for Spring overestimates in Washington State (46°N - 120.5°W) and Summer underestimates in Baja California (30°N - 114.5°W), respectively. The plotted confidence levels range from the 100% (all samples) to the 20% most confident samples. Levels below 20% confidence are not

plotted due to the limited number of samples. To ensure years with exceptionally high confidences are properly accounted for, confidence levels are calculated across all 20 years of testing data, rather than after each cross-validation subset. The six different random initializations, or seeds, are presented with various colors with the solid black line representing the average across all six of these seeds. Additionally, the gray region indicates the 95% confidence interval, which is determined based on a random chance of 33% and the number of samples. In both Figure 3.2a and Figure 3.2b, the average accuracy improves with confidence, emphasizing the potential to identify EOOs. Highlighting the heightened skill of a higher confidence, Figure 3.3 presents an example of three consecutive Spring seasons in Washington State (46°N - 120.5°W) where eleven of the fourteen 30% most confident testing samples are predicted correctly. (Note that Figure 3.3 shows results for a single fold and initialization seed). Additionally, nine of the fourteen most confident predictions are overestimates, with six of them occurring in 2004, exemplifying how the most confident samples may not be equally distributed across class or season. By specifically focusing on the most confident samples, we next delve into their underlying meteorological conditions to better understand why they result in skillful error predictions.

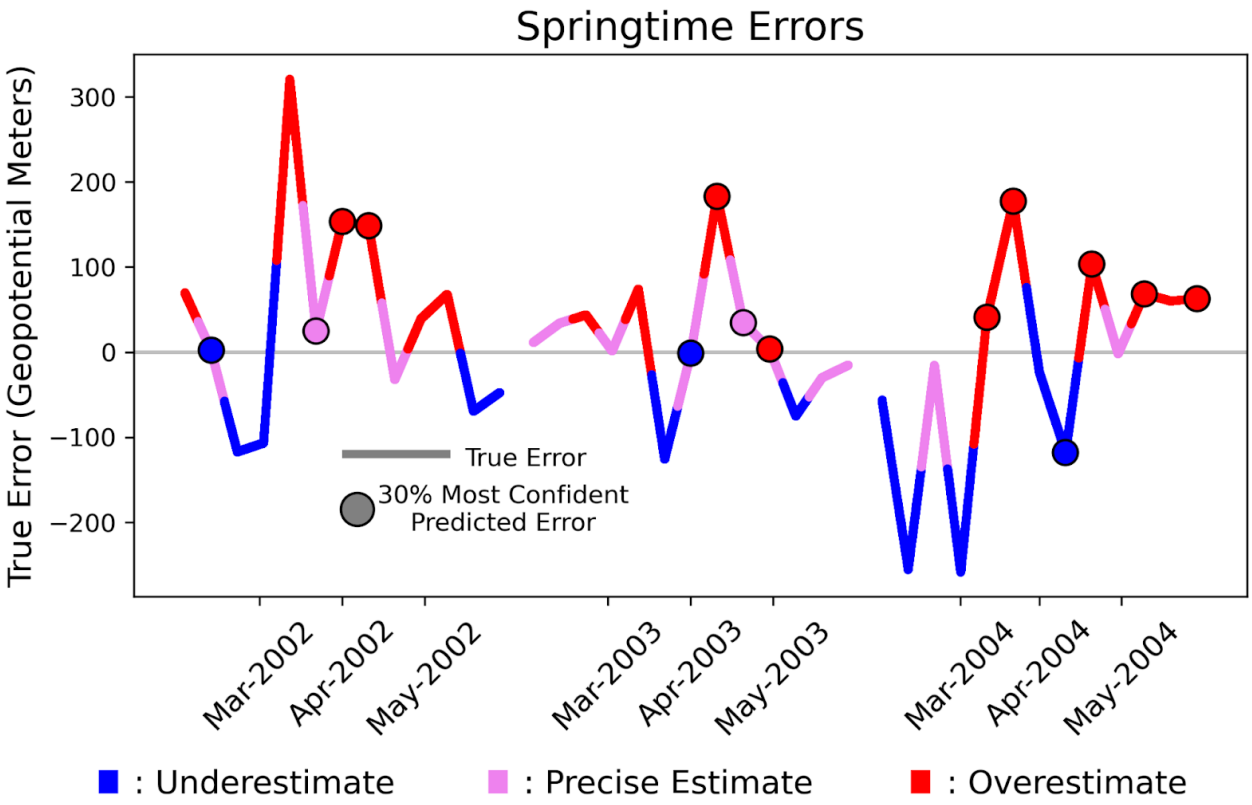


**Figure 3.3:** 500-hPa geopotential height testing errors for the Spring seasons of 2002-2004 for a location in Washington State (46°N - 120.5°W). Solid, colored lines indicate the true error and the colored circles indicate the network's predicted error for the 30% most confident samples. The three colors denote the three classes. Instances where the true error (lines) and the predicted error (circles) are the same color represent a correct prediction, while instances where the colors differ denote that the network was confident but wrong.

Exploring periods of enhanced confidence allows us to uncover specific features within the inputs that lead to more accurate predictions. Highlighting the increase in accuracy with confidence, Figure 3.4 shows a recreation of Figure 3.1, but for the 30% most confident samples. Here we see distinct regions of spatial robustness and improved accuracy relative to Figure 3.1. This is particularly evident in the North Pacific (Figures 3.4c,d,g,l) and over the western United States (Figures 3.4a,j,l). For our case studies, Figure 3.4a and Figure 3.4l display improved accuracy in Northwest Mexico and the West Coast of America, respectively.

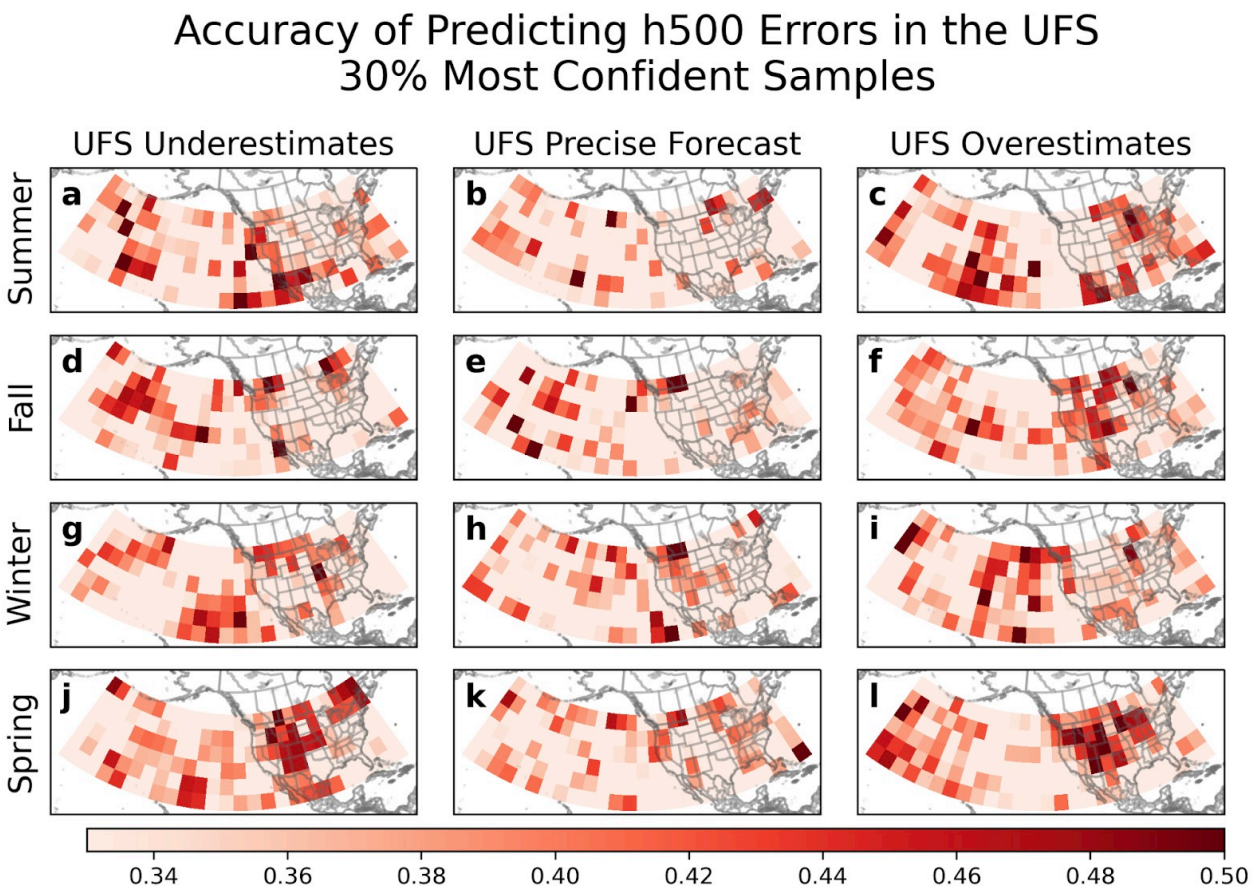


**Figure 3.4:** As in Figure 3.1, but for the 30% most confident samples.

## 3.2 Case Study 1: Pacific Northwest

As presented in Figure 3.1l and Figure 3.4l, Spring overestimates represent a situation of higher skill for many locations in the continental United States, particularly for the 30% most confident samples. Therefore, our first case study analyzes Spring overestimates for a four grid cell average over Washington State (46N-48N, 122.5W-120.5W). This region is of particular interest due to well-documented tropical-extratropical teleconnections affecting the western United States (Mo et al., 1997; Jones et al., 1998; Bond and Vecchi 2003; Moon et al., 2011). By examining this area, we aim to better understand EOOs that may be associated with these teleconnections. To document

the robustness of our results, we also performed a similar analysis for Spring overestimates across four grid cells in Colorado (37N-39N, 107.5W-105.5W), and the results are shown in Figure S2.

We begin by taking a composite of the normalized OLR input maps for the 30% most confident samples over Washington State (Figure 3.5a). When the networks are confident, we find a distinct OLR dipole of negative anomalies spanning from the Malay Archipelago to the Indian Ocean and positive anomalies west of New Guinea. This dipole resembles phases 3-5 of the MJO, which is further supported by the distribution of MJO phases for these samples presented in Figure 3.5. Here, the overall counts of the MJO phases for these samples are represented by the blue bar graph, and the relative frequency of the phases to all Spring samples is represented by the black line plot. While Figures 3.5a,c provide information through composites, Figure 3.5b provides insights into the decision making process of the neural network itself; its relevance map highlights the MJO region relative to other regions of our input. Corresponding to our input composite dipole, the relevance map emphasizes regions of the Western Indian Ocean and north/southwest of New Guinea. Based on our findings, we conclude our networks can accurately predict Spring overestimates in Washington State following MJO phase 4. Interestingly, relevance maps for Colorado overestimates in Figure S2 put more emphasis on the tropical west Pacific than for Washington State overestimates, highlighting that teleconnection error origins for distinct U.S. regions may originate in different parts of the tropics.

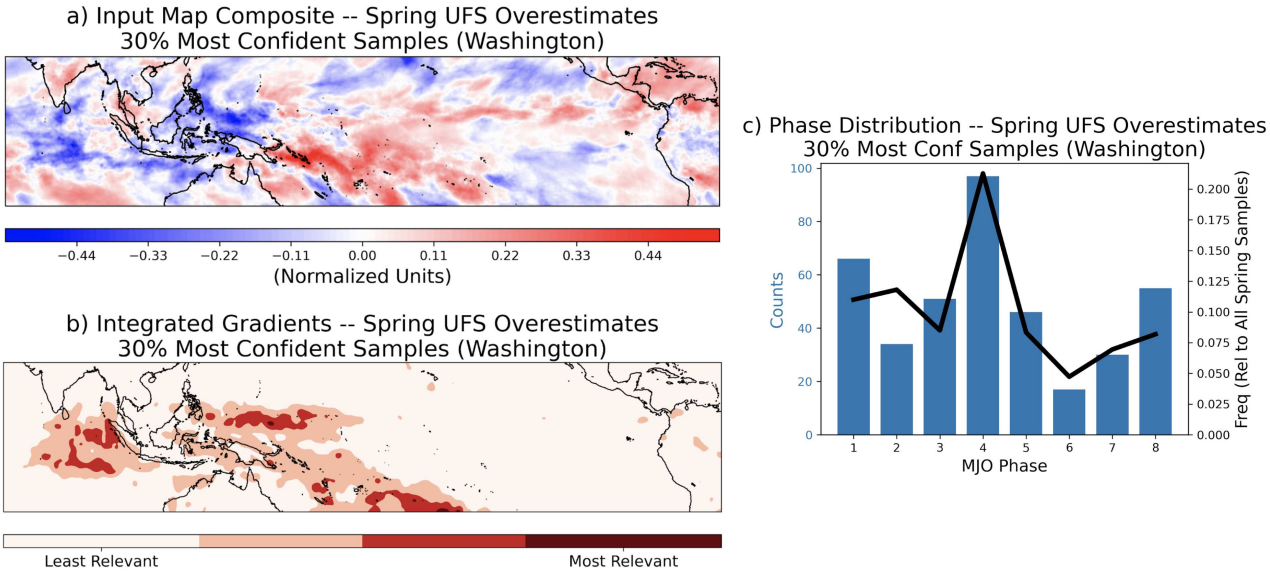


**Figure 3.5:** a) Composite of normalized OLR input maps, b) integrated gradients relevance map, and c) a distribution of MJO phases associated with the input maps for the 30% most confident Spring, underestimate samples predicted across four locations in Washington State. The bar chart represents the total counts of each MJO phase and the black line represents the phase frequency relative to all Spring samples.

Results from Figure 3.5 suggest that the network has confidently identified UFS h500 underestimates for particular MJO phase 4 samples. Figure 3.6 displays the progression of the h500 anomalies and UFS errors for these samples for lead times of 0 to 17 days. Distinct differences between the UFS and observations are seen in the evolution of the positive anomalies that begin in the Northwest Pacific (Figures 3.6a and 3.6g). These anomalies progress eastward more quickly in the UFS than in the observations. This discrepancy is particularly evident between 6 and 11 day leads, where the positive anomalies in the UFS progress from the North Pacific to the Pacific Northwest (Figures 3.6c-3.6d) while they remain in the North Pacific in the observations (Figures 3.6i-3.6j). The slower progression of anomalies in the observations compared to the UFS is also emphasized in Figures 3.6j-3.6k with the negative anomalies mostly stagnating in the Pacific Northwest over a 6 day period.

The faster progression speed within the UFS compared to the observations is most likely due to UFS errors in MJO strength. A reproduction of Figure 3.6, but for tropical OLR anomalies is shown in Figure S3. In both the UFS (Figures S3a and S3f) and the observations (Figures S3g and S3l), the MJO progresses eastward across the Maritime Continent. However, as the MJO progresses the UFS exhibits significantly stronger and more sustained OLR anomalies compared to the observations, consistent with the faster progression of h500 anomalies by the UFS. Past research has identified similar interactions of the MJO influencing midlatitude h500 progression. For instance, dynamical models have difficulty in simulating the MJO teleconnection to the North Pacific as manifest in h500 anomalies (Stan et al., 2022). Additionally, Henderson et al. (2016) demonstrated that tropical-extratropical teleconnections can affect the eastward progression of North Pacific h500 anomalies during the Winter.

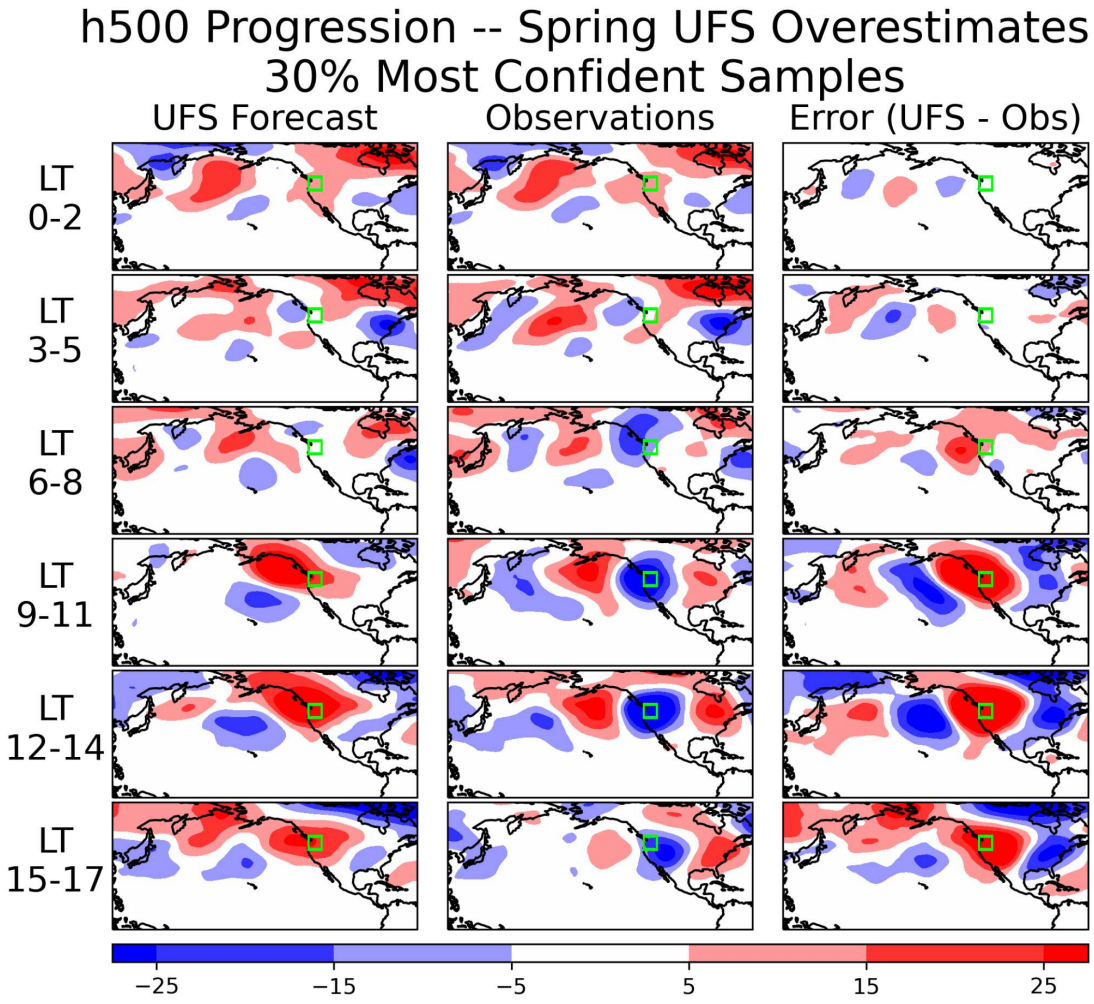


**Figure 3.6:** 500-hPa geopotential height composites of the 30% most confident spring, underestimate samples. Composites develop from lead 0 - associated with the neural network input - until lag 17. Columns are separated into the UFS Forecast, reanalysis, and their corresponding error. The green box indicates the predicted location in Washington State.

## 3.3 Case Study 2: Baja California

Another instance of higher prediction skill for EOOs occurs in Northwest Mexico for Summer underestimates (Figure 3.1a and Figure 3.4a). During the Summer, the MJO is often referred to as the Boreal Summer Intraseasonal Oscillation (BSISO) where OLR anomalies are found at more northerly latitudes, and exhibit both northward and eastward propagation (Jiang et al., 2018). Following a similar strategy as the Washington State case, Figure 3.7a shows strong positive OLR

anomalies north of the Maritime Continent and strong negative anomalies south of Mexico. These features share a strong resemblance with phases 1 and 2 of the BSISO, as shown in Figure S4 (see also Kikuchi 2021). Furthermore, the network also appears to highlight the BSISO activity as the most relevant region for its confident predictions of UFS underestimates (Figure 3.7b), identifying regions north of the Maritime Continent and south of Mexico, aligning with the strongest OLR features of Figure 3.7a. Finally, Figure 3.7b also exhibits a thin band of relevance across the Pacific around 7°N. This band is likely related to the active phase of the BSISO, during which regions of convection propagate across the Intertropical Convergence Zone towards the East Pacific in as little as two weeks (Kikuchi 2021). We further explore the importance of this band next.

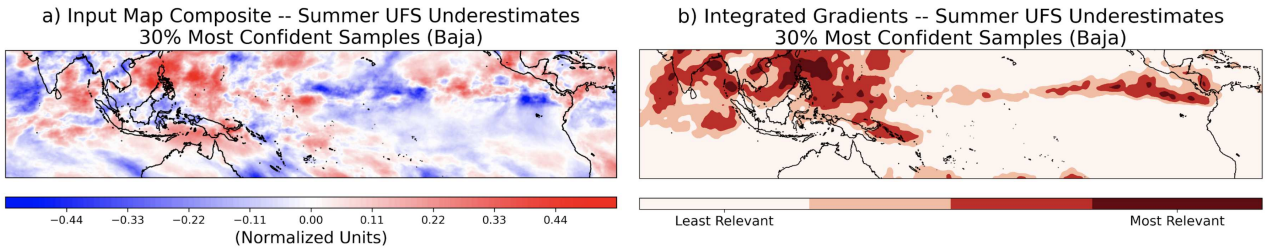


**Figure 3.7:** 30% most confident Summer, overestimate samples predicted to Baja California, represented by composites of a) input maps and b) the network's integrated gradient.

Instead of differences in progression speeds as seen for Spring overestimates (Figure 3.6), the h500 anomaly patterns for the UFS and observations generally propagate at the same pace in the Baja California case for the 30% most confident Summer underestimates (Figure 3.8). The h500 patterns between the UFS and observations are quite similar over the North Pacific, but a distinct difference lies over our predicted region illustrated by the green box in Figure 3.8. The observations show the development of positive anomalies in the predicted region (Figures 3.8i-3.8l) that are not replicated by the UFS (Figures 3.8c-3.8f). Similar panels, but for OLR show that the UFS predicts

negative OLR anomalies near Baja California from lead days 9-17 (Figures S5d-S5f), which are not present in the observations (Figures S5j-S5l). The presence of negative OLR, often associated with convection, may hinder the formation of positive geopotential anomalies (Iribarne and Cho, 1980). Our EOO approach thus effectively identifies a unique interaction between the BSISO and h500 in the East Pacific, but future investigation is required to fully recognize and improve these UFS errors.

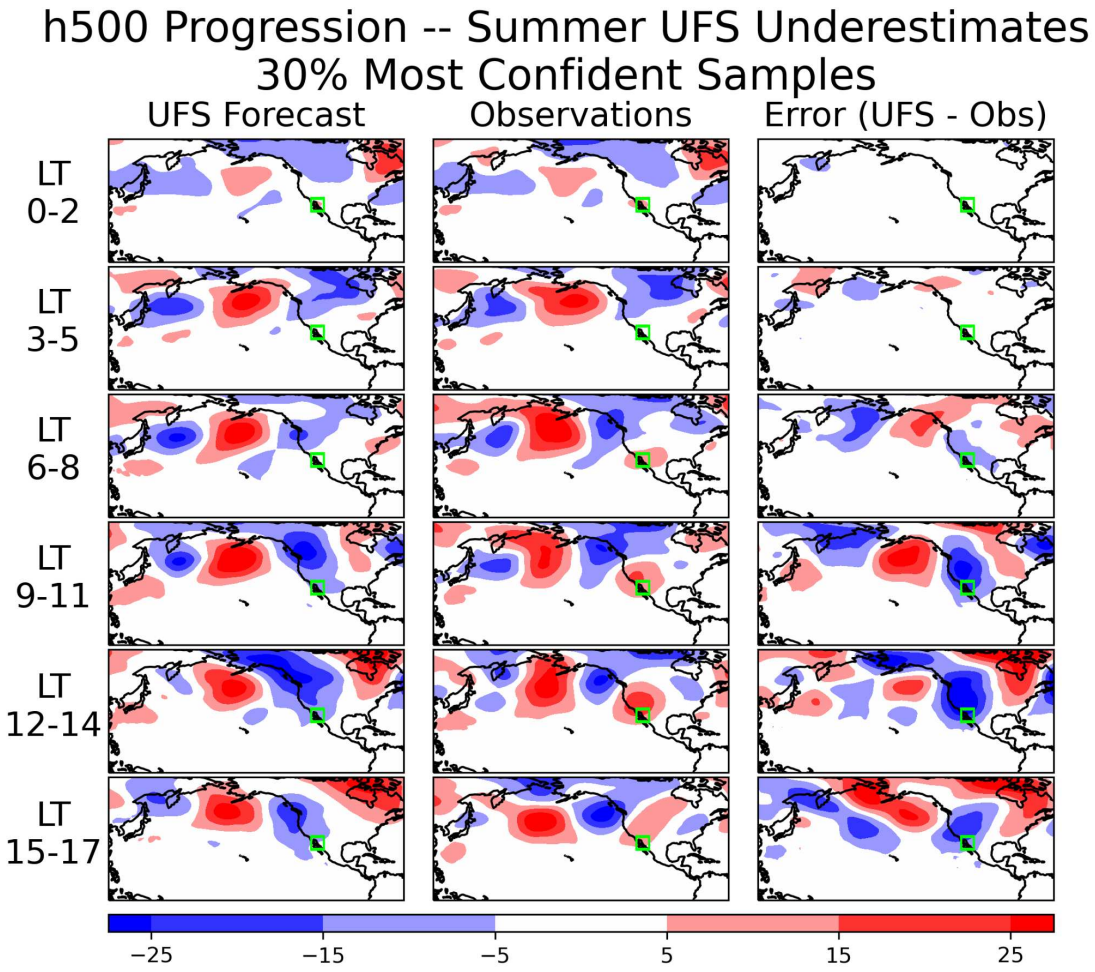


**Figure 3.8:** As in Figure 3.6, but for the 30% most confident Summer, overestimate samples predicted to Baja California.

## 3.4 Case Study 3: Great Plains Summertime Precipitation Errors

Since BSISO phases 1 and 2 are associated with h500 differences between the UFS and observations, it suggests the possibility that these h500 errors have downstream impacts on other atmospheric variables. Specifically, we next show precipitation errors following the 30% most confident Summer h500 underestimates based on Baja California predictions, a period favoring BSISO phases 1 and 2 (Figure 3.9). At a lead time of 12-16 days in Figure 3.9c, precipitation is overestimated across much of the northern Midwest and underestimated in the Gulf of Mexico. Previous research has identified Summertime tropical-extratropical teleconnections that affect this region, including those originating from the BSISO (Lang et al., 2020; Weaver and Nigam, 2008). Weaver and Nigam's study identified a teleconnection pattern that has a modulating effect on the Great Plains Low-level Jet (GPLLJ), which is instrumental in the meridional water flux along 100°W from the Gulf of Mexico to the Canadian border. Compellingly, when the teleconnection is active they found positive precipitation anomalies in the North Midwest and negative precipitation anomalies in the Gulf of Mexico, similar to the precipitation errors in Figure 3.9c. Additionally, during the active teleconnection period, diabatic heating anomalies closely resemble phases 1 and 2 of the BSISO, and a 200hPa height pattern closely resembles Figures 3.8j-3.8k.

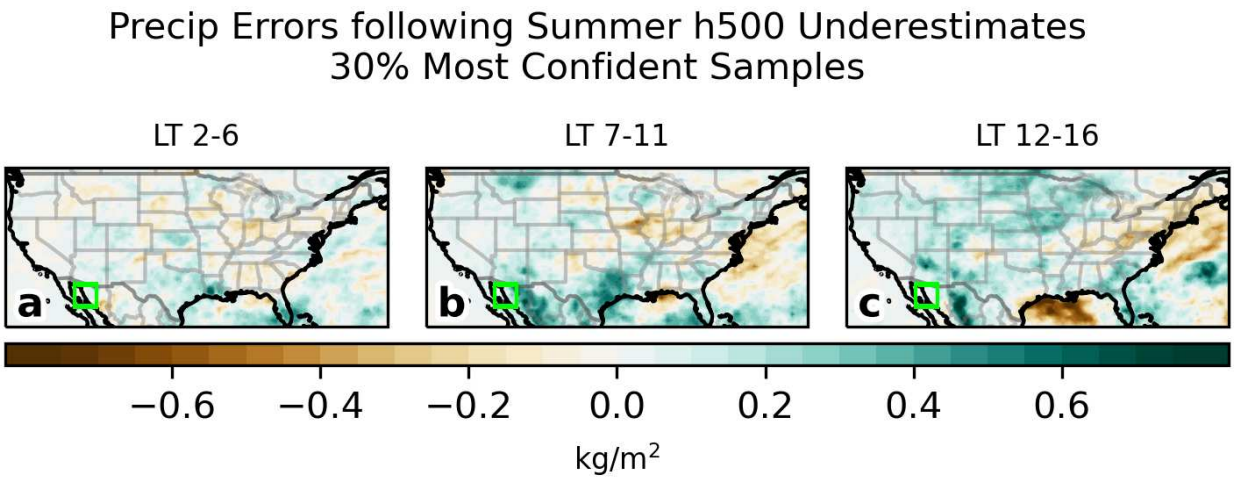


**Figure 3.9:** U.S. precipitation errors following the 30% most confident Summer, h500 overestimate samples predicted at a location in Baja California spanning lead times 2-16.

Following the same general neural network approach as for h500, we train neural networks to classify UFS Summertime precipitation at 12-16 day lead times using tropical OLR as the sole input. Since the North Midwest exhibits the largest errors over land in Figure 3.9, we predict precipitation errors in this region (Figure 3.10a). Coinciding with the largest overestimates in Figure 3.9c, the networks are most skillful at predicting precipitation overestimates in North Dakota, Northwest Minnesota, and southern Wisconsin. Much like Figure 3.7b, the network places the most relevance on the BSISO region but now for northern Minnesota (48°N, 96°W) precipitation errors (Figure 3.10b). Furthermore, the relevance map emphasizes the East Pacific convective region. This region corresponds to BSISO phases 1 and 2, which prior work has shown to modulate the flow associated with the GPLLJ (Maloney and Hartmann, 2000; Small et al., 2011). Maloney and Hartmann's research found that East Pacific convection modified the flow in a way consistent with decreased moisture flux from the Gulf of Mexico to the Great Plains, which provides a plausible explanation for the underestimation of northern Midwest precipitation by the UFS. In our

study, this can be attributed to enhanced East Pacific convection predicted by the UFS to the south of Mexico (Figures S5d-S5f), which is not seen in the observations (Figures S5d-S5f).

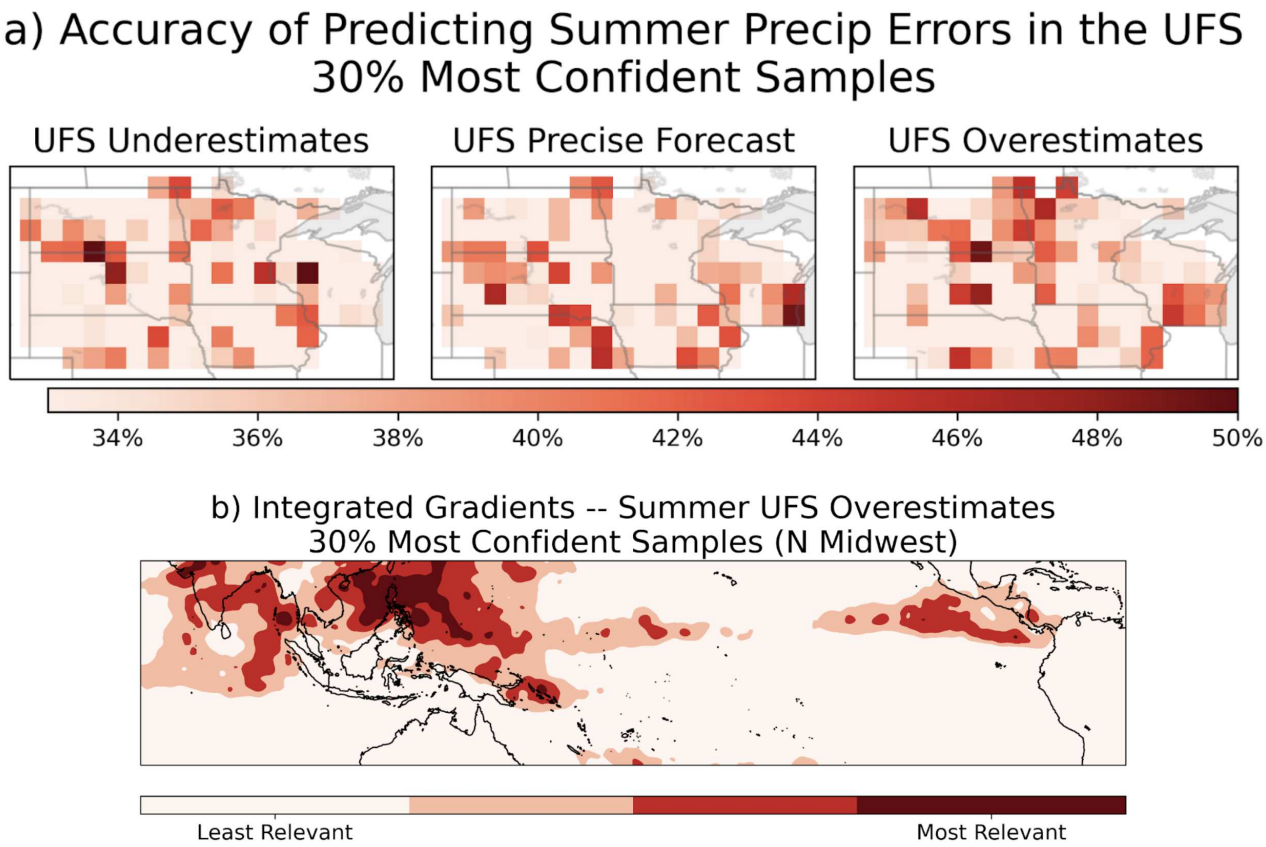


**Figure 3.10:** a) Prediction accuracy of precipitation errors at lead time 12-16 for various locations in the North Midwest averaged across 6 networks of different initializations divided by predicted class and b) a relevance map at one of the locations (48°N, 96°W) for Summer precipitation overestimates.

# Chapter 4

# Conclusions

This work demonstrates a novel approach to address challenges within the subseasonal to seasonal prediction period by utilizing artificial neural networks. With tropical OLR as the input, we predict "errors of opportunity" in NOAA's Unified Forecast System in North America. Specifically, we analyzed the most confident neural network predictions by compositing the input maps, analyzing progressions of the predicted variable, and utilizing the xAI method 'Integrated Gradients.' Compared to looking at all samples, examining solely the most confident samples proved to be an effective strategy at uncovering UFS's shortcomings and the meteorological conditions behind them.

Specifically, using our 'errors of opportunity' approach we found case studies where the networks had enhanced skill at identifying errors following phase 4 of the MJO and phases 1 and 2 of the BSISO for different regions and seasons. Our networks produced an increased accuracy for boreal Spring, h500 underestimates in the Pacific Northwest following phase 4. During the composite analysis, we observed a faster progression of h500 anomalies in the UFS compared to the observations; a disparity likely caused by the UFS's stronger and more sustained OLR anomalies throughout the eastward propagation of the MJO. Furthermore, our networks yielded high accuracy for boreal summer, h500 overestimates following phases 1 and 2 of the BSISO. During these periods, the UFS underestimates h500 in western Mexico and overestimates precipitation in the northern Midwest. The precipitation errors are likely caused by the UFS failing to capture tropical-extratropical teleconnections affecting the Great Plains low-level jet.

It is important to acknowledge that validating specific conclusions is challenging when training a neural network on a limited sample size of approximately 1000 samples. However, our study serves as a crucial first step in identifying predictable errors of the UFS over North America, and as reforecast data grows in availability, our findings can become solidified. Additionally, our research focused on a singular forecast model and a limited number of case studies, but by broadening our scope, the EOO approach can provide a better understanding of the challenges and opportunities associated with S2S predictions. Ultimately, our errors of opportunity approach demonstrates that by using a neural network we can predict forecast errors on S2S timescales, and by leveraging xAI and other strategies we can better understand the physical mechanisms behind the errors.

# Chapter 5

# Future Research and Thesis Summary

## 5.1 Future Research Directions

While applying the Errors of Opportunity (EOO) approach to three case studies yielded valuable results, there remain additional opportunities that should be further explored. Despite the achievements of our three case studies, Figure 5 presents several other regions and periods of enhanced accuracy and spatial coherence. Any one of these other regions could provide additional insights into how and why the forecast errors occurred. Additionally, the success of the EOO approach within NOAA's Unified Forecast System warrants the exploration into additional forecast models. This analysis has the potential to reveal common weaknesses across multiple models, as well as specific weaknesses unique to each model.

Furthermore, it is worth exploring different configurations of the ANN architecture since much of it remained unchanged throughout the study. For instance, there exist numerous other regions, inputs and target variables to analyze. For example, the utilization of temperature anomalies over North America and the North Pacific could provide a better understanding of how errors originate from meteorological phenomena within the region such as the North Atlantic Oscillation (NAO) (Hurrell et al., 2003). A more complex strategy could involve using multiple input maps of different variables, each spanning the entire globe. Under this framework, the prediction method would resemble a typical weather forecast that incorporates multiple variables and the full range of the Earth. In addition, while the use of our one-layer neural networks showed promise, investigating

the skill of multi-layered or deep learning architectures is crucial. Such exploration may lead to improved prediction skill, enabling the identification of additional sources of errors and an overall more reliable framework (Sarkar et al., 2020; van Straaten et al., 2023).

One of the weaknesses of this study is the limited number of training samples. Despite this, each weekly UFS forecast contains ten ensemble runs in addition to the singular control run. Meaning that instead of just over 1000 training samples, we could have well over 10,000, increasing the networks' robustness and potential accuracy. However, it is important to acknowledge that there are some drawbacks to this strategy. While aimed at capturing forecast uncertainty, the introduction of perturbations in the ensemble runs result in inherent biases that are not present in the control runs (Cui et al., 2012). Nonetheless, it would be a worthwhile endeavor to compare the skill of the error predictions in each framework.

An additional way to expand this research could involve advancing the error predictions by taking the next step and attempting to correct the forecast errors. By constructing the ANN as a linear regression problem instead of a classification problem, our networks can predict the magnitude and direction of the errors (Maulud and Abdulazeez, 2020). Based on the error prediction, the forecast prediction can be adjusted, resulting in new forecasts. This approach allows us to compare the accuracy of the UFS forecast to the accuracy of our new adjusted forecast to determine if improvements have been made. Similar to the error prediction analysis in this study, we could split our results by season to examine the situations where the most success is achieved.

## 5.2 Summary of Thesis

When reflecting on my thesis, it is only fitting to begin by revisiting when the project was first proposed to me. At the time, I was in the final year of my undergraduate degree and frankly

not enthusiastic about pursuing graduate school. I had seemingly lost my passion for atmospheric science and approached each school interview with a sense of detachment. However, it was at this time that I met one of my current advisors, who outlined a project that reignited that same fire I thought I lost. I no longer felt constrained to applying complex equations to even more complex concepts or engaging in concepts that did not interest me. Instead, I could spend my days applying cutting-edge machine learning and data science techniques within a field I had forgotten how much I loved. I was excited once again.

From that point forward, the process of completing my thesis became increasingly more enjoyable. I was fortunate enough to have not just one, but two advisors who complemented each other perfectly. One advisor provided invaluable guidance in devising clever approaches to solve problems and validate results, while the other helped me consider the big picture and connect my machine learning outcomes to the underlying physical science. In addition to my advisors, I felt incredibly grateful for my large research group as they helped me identify potential issues with my thesis early on. For instance, when I initially constructed my ANN from scratch, it was strongly recommended from members of my group that I instead utilize prebuilt programming packages. In retrospect, this was an incredibly important suggestion as it simplified the neural network process in the long term. Outside of academia, I was fortunate to gain experience with a weather and climate company. They helped me look at my research problem with new perspectives, while also confirming my desire to pursue a career in the private sector following the completion of my thesis.

It is important to acknowledge that not every aspect of writing my thesis was ideal, as there were certain things I wish I approached differently. Firstly, I wish I voiced my concerns over what appeared to be minor issues. For instance, I realized pretty early on that my neural networks were outputting too many "perfect" confidence scores. Although I initially dismissed the problem, it

ultimately resurfaced and required significant effort to fix during the later stages of my research. Additionally, I wish I spent more time exploring other variables, such as surface temperature. While the overall accuracy when predicting temperature errors was low, I realize now that I did not spend the proper time exploring alternative avenues within the ANN architecture. Despite these few limitations, creating this thesis has been an incredibly rewarding experience, and I am immensely proud to have accomplished it. It has undoubtedly prepared me for my future, and I cannot wait to see where this experience takes me.

# Bibliography


Adadi, A., & Berrada, M. 2018: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Alexander, M. A., Bladé, I., Newman, M., Lanzante, J. R., Lau, N.-C., & Scott, J. D. 2002: The Atmospheric Bridge: The Influence of ENSO Teleconnections on Air–Sea Interaction over the Global Oceans. *J. Climate*, *15*(16), 2205–2231.

Alley, R. B., Emanuel, K. A., & Zhang, F. 2019: Advances in weather prediction. *Science*, *363*(6425), 342–344.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. 2015: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS One*, *10*(7), e0130140.

Baggett, C. F., Barnes, E. A., Maloney, E. D., & Mundhenk, B. D. 2017: Advancing atmospheric river forecasts into subseasonal-to-seasonal time scales. *Geophys. Res. Lett.*, *44*(14), 7528–7536.

Bjerknes, J. 1969: Atmospheric Teleconnections From The Equatorial Pacific. *Mon. Wea. Rev.*, *97*(3), 163–172.

Bond, N. A., & Vecchi, G. A. 2003: The Influence of the Madden–Julian Oscillation on Precipitation in Oregon and Washington. *Wea. Forecasting*, *18*(4), 600–613.

Breeden, M. L., Albers, J. R., Butler, A. H., & Newman, M. 2022: The Spring Minimum in Subseasonal 2-m Temperature Forecast Skill over North America. *Mon. Wea. Rev.*, *150*(10), 2617–2628.

Cassou, C. 2008: Intraseasonal interaction between the Madden–Julian Oscillation and the North Atlantic Oscillation. *Nature*, *455*(7212), 523–527.

Chen, T., & Chen, H. 1995: Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Trans. on*

*Neural Networks / a Publ. of the IEEE Neural Networks Counc.*, *6*(4), 911–917.

Cohen, J., Coumou, D., Hwang, J., Mackey, L., Orenstein, P., Totz, S., & Tziperman, E. 2019: S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdiscip. Rev. Climate Change*, *10*(2), e00567.

Cui, B., Toth, Z., Zhu, Y., & Hou, D. (2012). Bias Correction for Global Ensemble Forecast. *Wea. and Forecasting*, *27*(2), 396–410.

Dai, Y., Feldstein, S. B., Tan, B., & Lee, S. 2017: Formation Mechanisms of the Pacific–North American Teleconnection with and without Its Canonical Tropical Convection Pattern. *J. Climate*, *30*(9), 3139–3155.

Dias, J., & Kiladis, G. N. 2019: The influence of tropical forecast errors on higher latitude predictions. *Geophys. Res. Lett.*, *46*(8), 4450–4459.

Dutra, E., Johannsen, F., & Magnusson, L. 2021: Late Spring and Summer Subseasonal Forecasts in the Northern Hemisphere Midlatitudes: Biases and Skill in the ECMWF Model. *Mon. Wea. Rev.*, *149*(8), 2659–2671.

French, M. N., Krajewski, W. F., & Cuykendall, R. R. 1992: Rainfall forecasting in space and time using a neural network. *J. of Hydrology*, *137*(1), 1–31.

Glahn, H. R. 1964: An Application of Adaptive Logic to Meteorological Prediction. *J. Appl. Meteor. Climatol.*, *3*(6), 718–725.

Goodman, B., & Flaxman, S. 2017: European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Mag.*, *38*(3), 50–57.

Guan, H., and Coauthors, 2022: GEFSv12 Reforecast Dataset for Supporting Subseasonal and Hydrometeorological Applications. *Mon. Wea. Rev.*, *150*(3), 647–665.

Hamill, T. M., and Coauthors, 2022: The Reanalysis for the Global Ensemble Forecast System, Version 12. *Mon. Wea. Rev.*, *150*(1), 59–79.

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

Henderson, S. A., Maloney, E. D., & Barnes, E. A. 2016: The Influence of the Madden–Julian Oscillation on Northern Hemisphere Winter Blocking. *J. Climate*, *29*(12), 4597–4616.

Hoskins, B. J., & Ambrizzi, T. 1993: Rossby Wave Propagation on a Realistic Longitudinally Varying Flow. *J. Atmos. Sci.*, *50*(12), 1661–1671.

Hoskins, B. J., & Karoly, D. J. 1981: The Steady Linear Response of a Spherical Atmosphere to Thermal and Orographic Forcing. *J. Atmos. Sci.*, *38*(6), 1179–1196.

Hsieh, W. W. 2009: *Mach. Learn. Methods in the Environ. Sci.: Neural Networks and Kernels*. Cambridge University Press.

Hurrell, J. W., Kushnir, Y., Ottersen, G., & Visbeck, M. (2003). An overview of the north Atlantic oscillation. In *The North Atlantic Oscillation: Climatic Signif. and Environ. Impact* (pp. 1–35). American Geophysical Union.

Iribarne, J. V., & Cho, H.-R. 1980: Atmospheric Dynamics. In J. V. Iribarne & H.-R. Cho (Eds.), *Atmos. Phys.* (pp. 149–197). Springer Netherlands.

Jiang, X., Adames, Á. F., Zhao, M., Waliser, D., & Maloney, E. 2018: A Unified Moisture Mode Framework for Seasonality of the Madden–Julian Oscillation. *J. Climate*, *31*(11), 4215–4224.

Jones, C., Waliser, D. E., & Gautier, C. 1998: The Influence of the Madden–Julian Oscillation on Ocean Surface Heat Fluxes and Sea Surface Temperature. *J. Climate*, *11*(5), 1057–1072.

Kikuchi, K. 2021: The Boreal Summer Intraseasonal Oscillation (BSISO): A Review. *J. of the Meteor. Soc. of Japan. Ser. II*, *advpub*, 2021–2045.

Klemm, T., & McPherson, R. A. 2017: The development of seasonal climate forecasting for agricultural producers. *Agric. and For. Meteor.*, *232*, 384–399.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., Hou, Y. T., Lord, S. J., & Belochitski, A. A. 2010: Accurate and Fast Neural Network Emulations of Model Radiation for the NCEP Coupled Climate Forecast System: Climate Simulations and Seasonal Predictions. *Mon. Wea. Rev.*, *138*(5), 1822–1842.

Larson, S. M., Okumura, Y., Bellomo, K., & Breeden, M. L. 2022: Destructive Interference of ENSO on North Pacific SST and North American Precipitation Associated with Aleutian Low Variability. *J. Climate*, *35*(11), 3567–3585.

Lee, Y., Kim, H.-R., Noh, N., Kim, K.-Y., & Kim, B.-M. 2023: Enhancing Forecast Skill of Winter Temperature of East Asia Using Teleconnection Patterns Simulated by GloSea5 Seasonal Forecast Model. *Atmosphere*, *14*(3), 438.

Leutbecher, M., & Palmer, T. N. 2008: Ensemble forecasting. *J. of Comput. Phys.*, *227*(7), 3515–3539.

Lorenz, E. N. 1969: Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *J. Atmos. Sci.*, *26*(4), 636–646.

Lundberg, S. M., & Lee, S.-I. 2017: A unified approach to interpreting model predictions. *Proc. of the 31st Int. Conf. on Neural Inf. Processing Syst.*, 4768–4777.

Madden, R. A., & Julian, P. R. 1971: Detection of a 40–50 Day Oscillation in the Zonal Wind in the Tropical Pacific. *J. Atmos. Sci.*, *28*(5), 702–708.

Madden, R. A., & Julian, P. R. 1972: Description of Global-Scale Circulation Cells in the Tropics with a 40–50 Day Period. *J. Atmos. Sci.*, *29*(6), 1109–1123.

Maloney, E. D., & Hartmann, D. L. 2000: Modulation of Eastern North Pacific Hurricanes by the Madden–Julian Oscillation. *J. Climate*, *13*(9), 1451–1460.

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. 2020: Explainable Artificial Intelligence in Meteorology and Climate Science: Model Fine-Tuning, Calibrating Trust and Learning New Science. *xxAI - Beyond Explainable AI: Int. Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, 315–339.

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. 2022: Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environ. Data Sci.*, *1*, e8.

Maqsood, I., Khan, M. R., & Abraham, A. 2004: An ensemble of neural networks for weather forecasting. *Neural Comput. & Appl.*, *13*(2), 112–122.

Mariotti, A., and Coauthors, 2020: Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond. *Bull. Amer. Meteor. Soc*, *101*(5), E608–E625.

Mariotti, A., Ruti, P. M., & Rixen, M. 2018: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *Npj Climate and Atmos. Sci.*, *1*(1), 1–4.

Martin, Z. K., Barnes, E. A., & Maloney, E. 2022: Using simple, explainable neural networks to predict the madden‐Julian oscillation. *J. of Adv. in Model. Earth Syst.*, *14*(5). https://doi.org/10.1029/2021ms002774

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *J.of Appl. Sci. and Technol. Trends*, *1*(4), 140–147.

Mayer, K. J., & Barnes, E. A. 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophys. Res. Lett.*. https://doi.org/10.1029/2020gl092092

Merryfield, W. J., and Coauthors, 2020: Current and Emerging Developments in Subseasonal to Decadal Prediction. *Bull. Amer. Meteor. Soc*, *101*(6), E869–E896.

Mihalakakou, G., Santamouris, M., & Asimakopoulos, D. 1998: Modeling ambient air temperature time series using neural networks. *J. of Geophys. Res.*, *103*(D16), 19509–19517.

Mo, K. C., Nogues Paegle, J., & Wayne Higgins, R. 1997: Atmospheric Processes Associated with Summer Floods and Droughts in the Central United States. *J. Climate*, *10*(12), 3028–3046.

Montavon, G., Samek, W., & Müller, K.-R. 2018: Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

Moon, J.-Y., Wang, B., & Ha, K.-J. 2011: ENSO regulation of MJO teleconnection. *Climate Dyn.*, *37*(5), 1133–1149.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A Multimodel Subseasonal Prediction Experiment. *Bull. Amer. Meteor. Soc*, *100*(10), 2043–2060.

Philander, S. G. H. 1983: El Niño southern oscillation phenomena. *Nature*, *302*(5906), 295–301.

Rasmusson, E. M., & Wallace, J. M. 1983: Meteorological aspects of the el nino/southern oscillation. *Science*, *222*(4629), 1195–1202.

Rasp, S., & Lerch, S. 2018: Neural Networks for Postprocessing Ensemble Weather Forecasts. *Mon. Wea. Rev.*, *146*(11), 3885–3900.

Sardeshmukh, P. D., & Hoskins, B. J. 1988: The Generation of Global Rotational Flow by Steady Idealized Tropical Divergence. *J. Atmos. Sci.*, *45*(7), 1228–1251.

Sarkar, P. P., Janardhan, P., & Roy, P. (2020). Prediction of sea surface temperatures using deep learning neural networks. *SN Appl. Sci.*, *2*(8), 1458.

Shrikumar, A., Greenside, P., & Kundaje, A. 2017: Learning important features through propagating activation differences. *Proc. of the 34th Int. Conf. on Mach. Learn. - Vol. 70*, 3145–3153.

Small, R. J., Xie, S.-P., Maloney, E. D., de Szoeke, S. P., & Miyama, T. 2011: Intraseasonal variability in the far-east pacific: investigation of the role of air–sea coupling in a regional coupled model. *Climate Dyn.*, *36*(5), 867–890.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. of Mach. Learning Res.*, *15*(56), 1929–1958.

Stan, C., Straus, D. M., Frederiksen, J. S., Lin, H., Maloney, E. D., & Schumacher, C. 2017: Review of tropical‐extratropical teleconnections on intraseasonal time scales. *Rev. of Geophys.*, *55*(4), 902–937.

Stan, C., and Coauthors, 2022: Advances in the Prediction of MJO Teleconnections in the S2S Forecast Systems. *Bull. Amer. Meteor. Soc*, *103*(6), E1426–E1447.

Sundararajan, M., Taly, A., & Yan, Q. 2017: Axiomatic attribution for deep networks. In *arXiv [cs.LG]*. arXiv. http://proceedings.mlr.press/v70/sundararajan17a/sundararajan17a.pdf

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. 2020: Physically interpretable neural networks

for the geosciences: Applications to earth system variability. *J. of Adv. in Model. Earth Syst.*, *12*(9). https://doi.org/10.1029/2019ms002002

Tseng, K.-C., Barnes, E. A., & Maloney, E. D. 2018: Prediction of the Midlatitude Response to Strong Madden-Julian Oscillation Events on S2S Time Scales. *Geophys. Res. Lett.*, *45*(1), 463–470.

van Straaten, C., Whan, K., Coumou, D., van den Hurk, B., & Schmeits, M. 2023: Correcting sub-seasonal forecast errors with an explainable ANN to understand misrepresented sources of predictability of European summer temperatures. *Artif. Intell. Earth Syst.*, 1–49.

Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bull. Amer. Meteor. Soc*, *98*(1), 163–173.

Vitart, F., & Robertson, A. W. 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *Npj Climate and Atmos. Sci.*, *1*(1), 1–7.

Vitart, F., Robertson, A. W., & Anderson, D. L. T. 2012: Subseasonal to Seasonal Prediction Project: Bridging the gap between weather and climate. *Bull. of the World Meteor. Organ.*, *61*(2), 23.

Weaver, S. J., & Nigam, S. 2008: Variability of the Great Plains Low-Level Jet: Large-Scale Circulation Context and Hydroclimate Impacts. *J. Climate*, *21*(7), 1532–1551.

Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. 2021: Sub‐seasonal forecasting with a large ensemble of deep‐learning weather prediction models. *J. of Adv. in Model. Earth Syst.*, *13*(7). https://doi.org/10.1029/2021ms002502

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, *24*(3), 315–325.

Zhang, Y.-F., and Coauthors, 2022: Subseasonal-to-Seasonal Arctic Sea Ice Forecast Skill Improvement from Sea Ice Concentration Assimilation. *J. Climate*, *35*(13), 4233–4252.

Zheng, L., Lin, R., Wang, X., & Chen, W. 2021: The Development and Application of Machine Learning in Atmospheric Environment Studies. *Remote Sens.*, *13*(23), 4839.

# Appendix

| | 2000-2003 | 2004-2007 | 2008-2011 | 2012-2015 | 2016-2019 |
|---|---|---|---|---|---|
| Fold 1 | Testing | Training | Training | Training | Training |
| Fold 2 | Training | Testing | Training | Training | Training |
| Fold 3 | Training | Training | Testing | Training | Training |
| Fold 4 | Training | Training | Training | Testing | Training |
| Fold 5 | Training | Training | Training | Training | Testing |

Figure A1: Training and testing splits for the 5 different neural network set-ups.

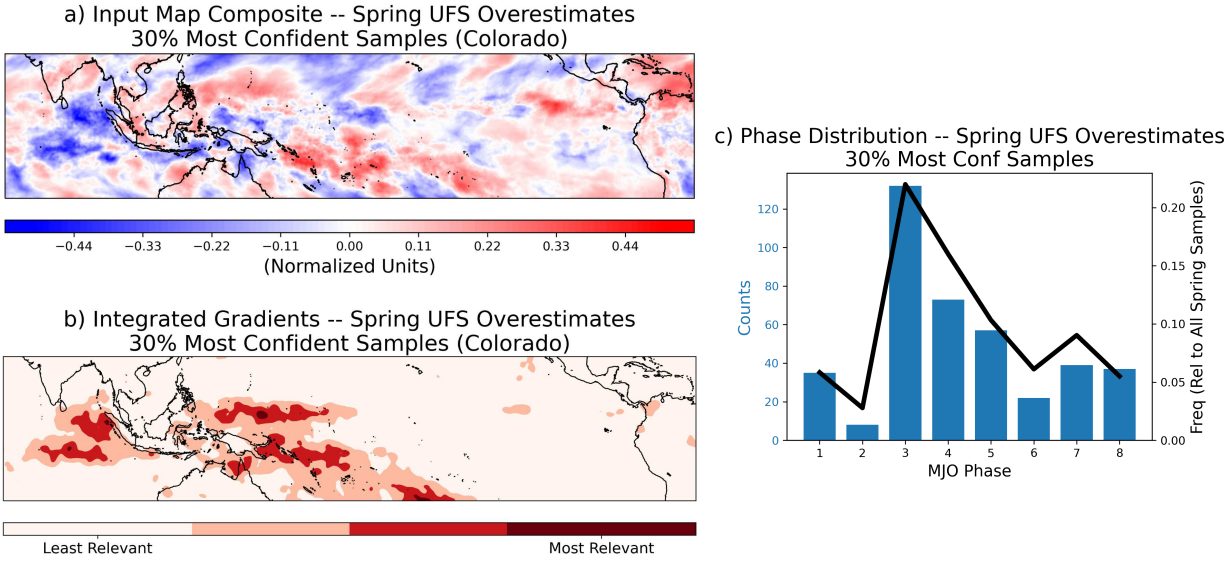


Figure A2: As in Figure 6, but for a region in Colorado (37N-39N, 107.5W-105.5W).

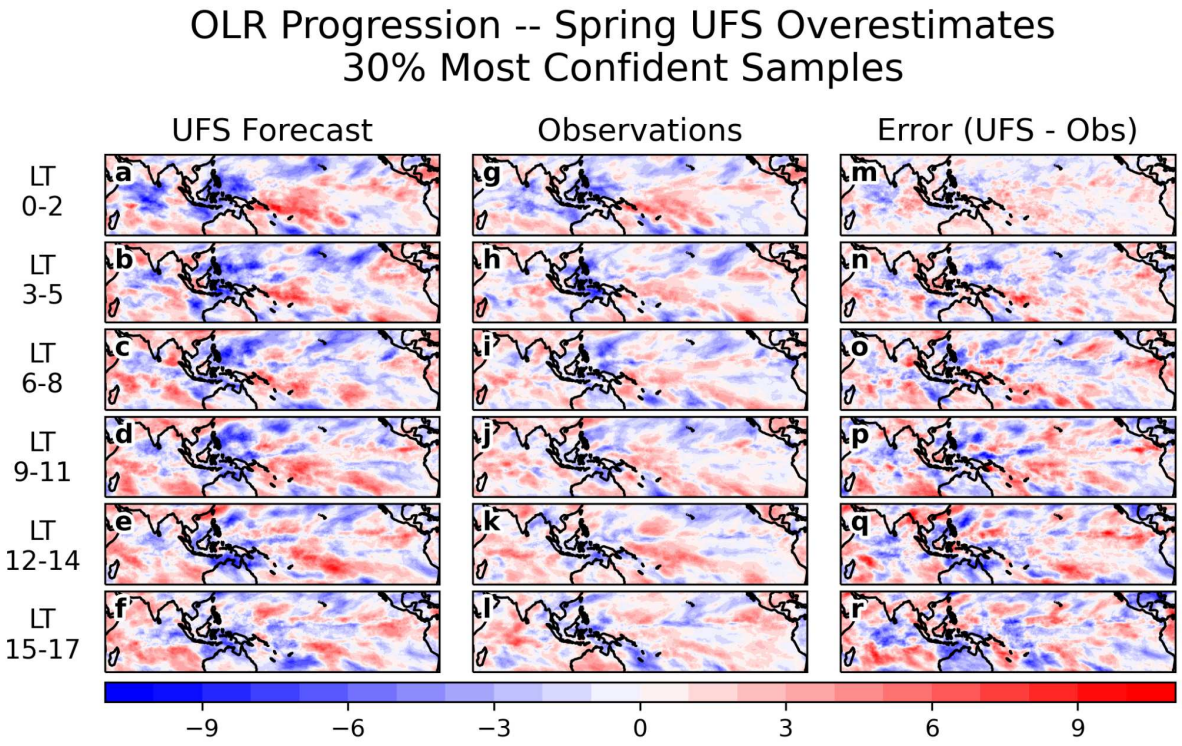


Figure A3: OLR composites of the 30% most confident Spring, underestimate samples. Composites develop from lead 0 - associated with the neural network input - until lag 17, forecasted to Washington State. Columns are separated into the UFS Forecast, reanalysis, and their corresponding error.

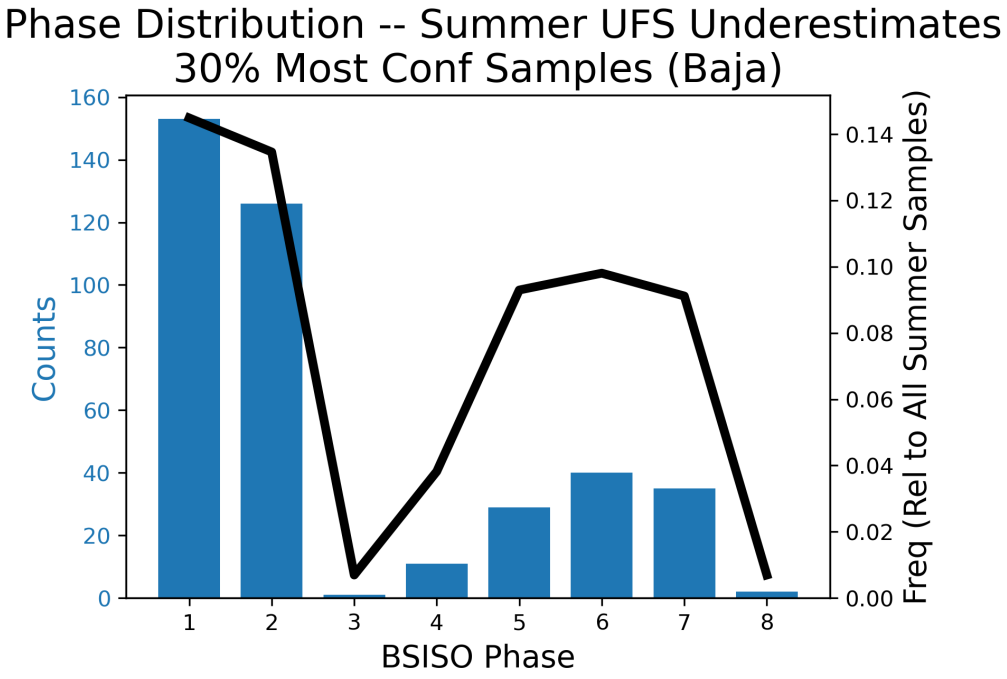


Figure A4: Distribution of BSISO phases for the 30% most confident summer, overestimate samples predicted at a location in Baja California.

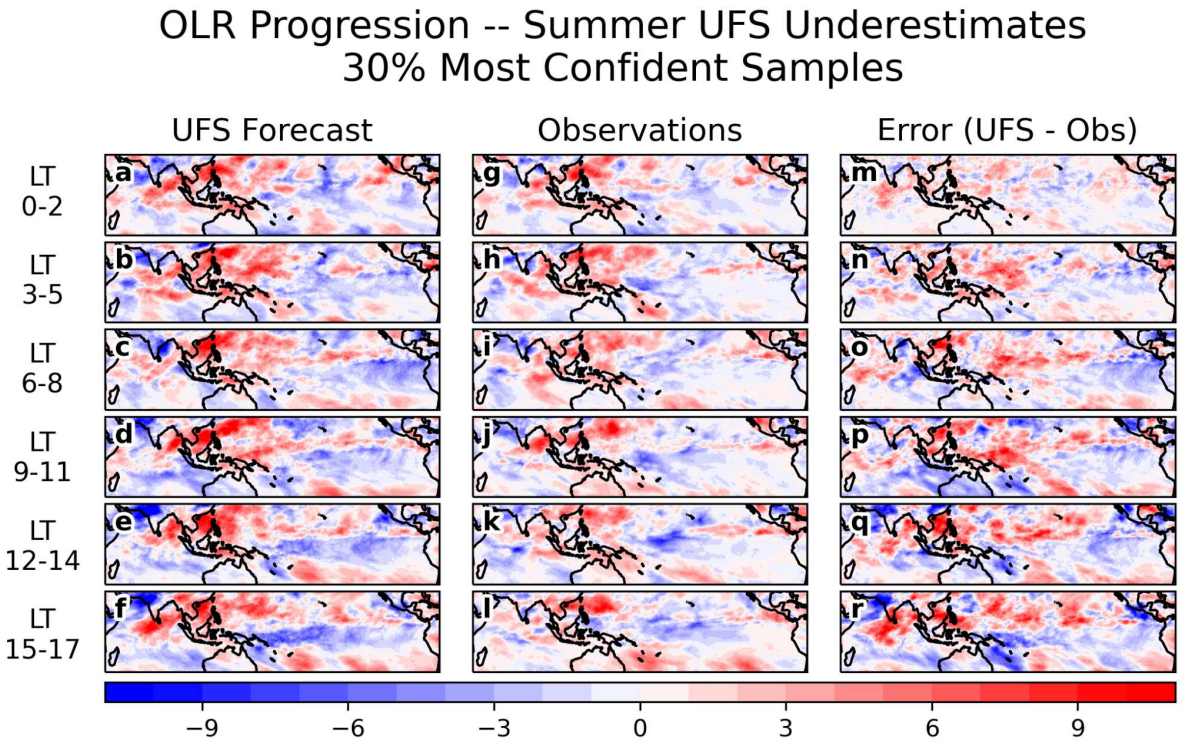


Figure S5: As in Figure S3, but for Summer Underestimates predicted to Baja California.