DISSERTATION

TOWARDS USING NEURAL NETWORKS FOR GEOSCIENTIFIC DISCOVERY

Submitted by

Benjamin A. Toms

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2020

Doctoral Committee:

Advisor: Elizabeth A. Barnes

Imme Ebert-Uphoff
James W. Hurrell
David W. J. Thompson

ABSTRACT

TOWARDS USING NEURAL NETWORKS FOR GEOSCIENTIFIC DISCOVERY

How can we use computational methods to extract physically meaningful patterns from geoscientific data? This question has been asked in some form for decades within the geoscientific community, with many landmark discoveries resulting from the novel application of computational methods to a geoscientific dataset. For example, the Madden-Julian Oscillation was discovered through Fourier transforms of tropical time-series, while the defining structures of the Northern Hemispheric annular modes were first captured using principal component analysis. These discoveries rooted in computational methods have since driven decades of geoscientific research and innovation, and are only two of among many similar examples. It is therefore clear that computational science and geoscience are inextricably intertwined, and so the continued advancement of both fields in tandem is beneficial to future geoscientific discovery.

Many methods exist to discover patterns within geoscientific data, although each is limited by its own set of assumptions. The most common assumption is that of linearity, which oftentimes conflicts with our understanding that the earth system can be both dynamically and statistically nonlinear. However, a recently popularized subset of methods within the computer science community known as neural networks can identify nonlinear patterns and are therefore potentially powerful tools for geoscientific discovery. Neural networks learn how to map one dataset to another using a combination of nonlinear relationships, and are generalizable to a broad range of tasks including forecasting and identifying patterns within images. Regardless of the application, a common limitation of neural networks has been the difficulty to understand how and why they make their decisions. Therefore, while they have been used in geoscience for more than two decades, they have mostly been applied when accuracy is valued more than understanding, such as for making forecasts.

Within this dissertation, we first propose a framework for how neural networks can be used for geoscientific discovery by applying recently invented methods from the computer science community. We focus on methods that explain which aspects of the input dataset are useful for the neural network when making connections to the output dataset. This framework enables physical interpretations of how and why neural networks make decisions, since the geoscientist that designs the neural network is likely familiar with the physical meaning of each input.

In the first study of the dissertation, we outline the framework and apply it to two simple tasks to ensure the neural network interpretations abide by our current understanding of the earth system. The interpretable neural networks successfully identify the pattern of the El Niño Southern Oscillation and oceanic patterns that lend seasonal predictability, which lends confidence that the framework is reliable. In the second study, we then further test the methods by applying them to a more spatially and temporally complex oscillation called the Madden-Julian Oscillation (MJO). The interpretable neural networks correctly identify the known spatial structures and seasonality of the MJO, and also suggest that the MJO is nonlinear and expresses its nonlinearity through the uniqueness of each event. The final study assesses whether the proposed framework can be used to identify predictable patterns of earth-system variability within climate models through its application to decadal predictability. We find that the interpretable neural networks identify known modes of oceanic decadal variability that contribute to predictability of continental surface temperatures. The interpretations can also be used to identify distinct regimes of predictability, wherein spatially and temporally unique oceanic modes contribute predictability for the same location at different times.

From a broader perspective, these studies suggest that neural networks are a viable tool for geoscientific discovery and are particularly useful given their ability to capture nonlinear, time-evolving patterns. It is likely that new neural network algorithms and methods for their interpretation will continue to be developed by the computer science community, and so this research provides a guideline for how such methods can be gainfully applied within the geosciences.

ACKNOWLEDGEMENTS

As a kid, my parents taught me to find my passion and follow it. And that if I did that, then everything else would turn out okay. They've been right so far. Thanks, Mom and Dad, for supporting my passion and teaching me to never let that fire fade.

My work towards this dissertation was funded by the Department of Energy Computational Science Graduate Fellowship via grant DE-FG02-97ER25308. The topic of this dissertation may have been drastically different without the support of this fellowship, so I am grateful for the intellectual freedom it has granted me.

DEDICATION

This dissertation is dedicated to my family, past and present, for building a future so I could have the opportunity to earn this PhD.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

Computational methods are essential to modern geoscience. Discoveries of dominant earth-system patterns such as the annular modes and Madden-Julian Oscillation have relied on automated pattern detection techniques [4,5]. Assessments of future impacts of anthropogenic climate change depend on computationally demanding earth-system models run on the world's largest supercomputers [6–8]. Computational methods even accelerate observational data retrieval and dictate the types of information that can be extracted from observations [9–12]. This is not to say theoretical studies are unimportant to the past and future of geoscience, as theory has oftentimes guided the future development and application of computational methods. Regardless, it has become clear that the advancement of geoscience depends on the continued integration of novel computational methods into its domain.

Most recently, machine learning methods have emerged as a powerful computational tool across many areas of geoscience [13, 14], including marine science [15], solid earth science [16], and atmospheric science [17–20]. This revolution in machine learning within the geosciences has been spurred by the coincident introduction of novel algorithms, an influx of large quantities of high-quality data, and an increase in computational power for processing immense quantities of data simultaneously. The algorithmic advances were originally inspired within the computer science community, for applications that are quite different from those within geoscience.

A decade ago, the computer science community introduced a dataset called Imagenet, which focused efforts in the community towards more advanced algorithms for image recognition [21,22]. At the surface, the task of the Imagenet challenge appears straight-forward: the participants must build an algorithm to accurately categorize which type of object is contained within images, with a total of 1,000 types of objects contained in the database. The complexity of the task lies in the fact that images of each type of object within the dataset contain different perspectives, orientations, and sizes of the object. In turn, this challenge led to the focused, rapid expansion of machine

learning algorithms capable of identifying and extracting useful information from images [23]. If the developed algorithms can distinguish between images of different objects within the Imagenet database, then it is possible that similar algorithms may be used for geoscientific problems such as automated feature detection within satellite images [24–26] or even the detection of predictable climate patterns [27].

Conveniently, a vast quantity of geoscientific datasets are also available in the form of images. From climate models to satellite images, the geosciences are ripe with image-like datasets from which useful information can be extracted. Extracting information from spatial patterns is not new to the geosciences, however, as the field has already used various methods over the past decades to extract meaningful information from these data. For example, the discovery and identification of annular atmospheric modes, tropical intraseasonal oscillations, and global modes of oceanic variability have utilized principal component analysis, which is a decades-old method for extracting information from many forms of data, including vectorized images [4, 28, 29]. The surge of machine learning usage within geoscience therefore is not the result of a lack of methods for processing image-like datasets. Rather, the newfound algorithms incepted by efforts targeted originally towards the Imagenet challenge may offer a new pathway for discovering geoscientific patterns.

Neural networks, the main type of algorithm advanced by the Imagenet challenge and tangential computer science research, offer a way to assess the nonlinear, time-evolving characteristics of geoscientific patterns [30]. As will be discussed in the subsequent chapters, the complexity of a neural network can be varied depending on the application, which enables a flexible framework for modeling the nonlinear relationship between datasets. This versatility is what drives the usage of neural networks in this dissertation. Can we use neural networks to identify new patterns in earth-system variability that are spatially or temporally nonlinear? This question is non-trivial. We must first assess whether neural networks are viable tools for identifying geoscientific patterns. If this is possible, we must then be able to thoroughly interpret what such models have learned about the patterns they have been trained to identify. It is well known that neural networks are

difficult to interpret and have therefore been colloquially referred to as “black boxes", because of the challenges in understanding how and why they make decisions [31, 32]. If the networks are interpretable, we must then thoughtfully use our geoscientific expertise to ensure the patterns the neural networks identify are physically meaningful. And, if the patterns are not yet known by the geoscientific community, the task then becomes particularly challenging when it must be tested whether the novel patterns are physically meaningful or spurious artifacts of a chaotic earth system. These challenges are immense, and this dissertation contributes a step towards understanding how neural networks can drive such geoscientific discovery.

This dissertation is organized into four subsequent chapters, the first three of which detail individual efforts to advance the usage of interpretable neural networks in geoscience, and the final of which summarizes these efforts. Each chapter focuses on different patterns of earth-system variability, the details of which are described within each chapter such that the reader can appreciate the unique importance of each pattern to each individual study.

The second chapter proposes a framework for extracting physically meaningful interpretations from neural networks using methods originally proposed within the computer science community. This chapter marks the first attempt to make interpreting what a neural network has learned the primary focus of using a neural network for a geoscientific study, and thereby provides a new framework for using such methods within the field. The proposed framework is then applied throughout the remaining chapters to further assess its reliability and to make new insights into the nonlinearities of geoscientific patterns. This manuscript has been published in the open-access American Geophysical Union journal, the Journal of Advances in Modeling Earth Systems.

The Madden-Julian Oscillation is the focus of the third chapter, in which myself and my collaborators assess whether the proposed framework can identify known structures of the spatio-temporally complex, global-scale tropical disturbance. We more thoroughly challenge the proposed framework through its application to a more complex problem, and also make a first attempt at its usage for geoscientific discovery. We find that interpretable neural networks do capture the defining characteristics of the Madden-Julian Oscillation, and so we then extend the interpreta-

tions to make new statements regarding its nonlinearities and seasonality. As of the writing of this dissertation, this manuscript is under review for publication in the European Geophysical Union journal Geoscientific Model Development.

Within the fourth chapter of the dissertation, we use the interpretable neural network framework to identify modes of predictability on decadal timescales within a climate model. While we apply the method to identifying predictable modes on decadal timescales, it can be applied to any timescale of interest. This chapter is the first attempt within the geosciences to use the interpretable neural network framework to advance our understanding of subseasonal-to-decadal earth-system variability. This study has not yet been submitted for review in a peer-reviewed journal, but will be submitted within a few weeks of the writing of this dissertation.

The final chapter ties these three independent studies together and provides some closing thoughts about the future of neural networks in geoscience. The work towards this dissertation began two years ago when the geosciences had not yet fully embraced the potential of neural networks. This research has occurred during a rather dramatic increase in interest in applying such methods to geoscientific problems, and I hope has contributed and will continue to contribute to their usage for geoscientific discovery. The following text describes this journey and summarizes the findings from my PhD.

Before venturing into the scientific findings of my PhD, I will note that the following chapters are only one part of my contributions to the geoscientific community throughout my four years in graduate school. I have had the opportunity to teach within and lead conference workshops, present my work at conferences and various institutions across the country, and serve as a coauthor on a few studies for which the lead authors were generous to include me [33, 34]. These additional roles have offered me at least as much personal and scientific growth as the research detailed in this dissertation. With that said, I hope that the following research has advanced the geoscientific community's understanding of how neural networks can be gainfully applied within our domain.

Chapter 2

Physically Interpretable Neural Networks for the Geosciences

2.1 Introduction

Machine learning methods are emerging as a powerful tool in scientific applications across all areas of geoscience [13, 14], including marine science [15], solid earth science [16], and atmospheric science [17–20]. This revolution in machine learning within the geosciences has been spurred by the coincident introduction of novel algorithms, an influx of large quantities of high-quality data, and an increase in computational power for processing immense quantities of data simultaneously. There have been limitations to the application of machine learning methods within geoscience, however, as their interpretation is commonly deemed difficult, if not impossible. Here, we show that two recent techniques from computer science for interpreting one of the most common forms of machine learning methods – neural networks – have the potential to transform how geoscientists use machine learning within their research. More specifically, these methods enable the usage of neural networks for the discovery of physically meaningful relationships within geoscientific data.

Neural networks, also occasionally dubbed “deep learning" [35], are one of the most versatile types of machine learning methods and can be used for a broad range of applications within the geosciences. Such models have been used for time-series prediction [36,37], identifying patterns of weather and climate phenomena within observations and simulations [38–41], and parameterizing sub-grid scale physics within numerical models [42–47]. The structure of the neural networks employed within these applications can vary substantially, although the general concept is the same: given a set of input variables, the neural network is tasked with identifying the desired output as accurately as possible.

Neural networks consist of consecutive layers of nonlinear transformations and adjustable weights and biases [48]. The mathematics of how these layer-to-layer transformations are applied to the data are well understood since the individual transformations themselves are mathematically simple [49]. However, once a neural network has been trained, the reasoning of how and why it combines information across its weights and biases and from each transformation to the next to arrive at its ultimate output is not easily deduced, due to the potentially high complexity of the network architecture and the increasing level of abstraction in later layers of the network [50]. Thus, in practice, neural networks are often used - including in geoscience - without a detailed understanding of the reasoning they employ to arrive at their output.

Even for applications where the network's output is all that is desired, a lack of understanding of a network's reasoning can lead to many problems. For example, the neural network can overfit to the data and attempt to explain noise rather than capturing the meaningful connections between the input and output. Additionally, within the geosciences sample sizes are typically limited, which means that the available samples might not capture the full range of possible outcomes and thereby might also not be representative of the true underlying physics driving the relationship between the inputs and outputs. In this scenario, the network may fail to model the relationship correctly from a physical perspective, even if it accurately captures a relationship between the inputs and outputs given the provided training data. Thus, the ability to interpret neural networks is important for ensuring that the reasoning for a network's outputs are consistent with our physical understanding of the earth system.

The various applications of neural networks within the geosciences commonly rely on indirect scientific inference. In many cases, the primary objective of the neural networks has been to maximize the accuracy of the networks' outputs, from which indirect inferences have been made about the earth system. For example, by using neural networks to predict the likelihood that a convective storm would produce hail, Gagne et al. (2019) showed that the neural networks made accurate predictions by identifying known types of storm structures. In another case, Ham et al. (2019) used a neural network to predict the evolution of the El Niño Southern Oscillation

(ENSO), and then used interpretation techniques to show that ENSO precursors exist within the South Pacific and Indian Oceans. However, even in these cases, the primary objective was to construct a neural network that most accurately predicted its output, with the interpretation being used to ensure the network attained high accuracy using reasoning consistent with physical theory. This theme is common throughout geoscientific applications of neural networks: the network's output is the ultimate objective, and interpretation techniques are used to ensure the network is making decisions according to our current understanding of how the earth system evolves. There have also been recent efforts within the geoscience community to compile methods for improving machine-learning model interpretability, including those by McGovern et al. (2019).

We propose an additional use for neural networks, whereby the ultimate scientific objective of using a neural network is its interpretation rather than its output. From this perspective, we show how neural networks can be used to directly advance our understanding of the earth system. To do so, we focus on two methods – backwards optimization and layerwise relevance propagation – which trace the decision of a neural network back onto the original dimensions of the input image, and thereby permit the understanding of which input variables are most important for the neural network's decisions. These methods are particularly well-suited for scientific inference when a physical understanding of relationships is important, such as within geoscience. We find that layerwise relevance propagation is particularly well suited for geoscientific applications, and has yet to be introduced to the geoscience community to the best of our knowledge.

We first discuss the theory and logic behind the two interpretation methods, then provide two examples of how these methods can be used to explore physically meaningful patterns of earth system variability. The objective of this paper is to showcase the utility of using neural network interpretations for scientific inference. So, we analyze two commonly studied climate phenomena, the El Niño Southern Oscillation and its relationship to seasonal prediction, so that we can first ensure the interpretation methods capture known patterns of geophysical variability before extending into the unknown.

2.2 Neural Network Architecture

In this work, we use separately trained fully-connected neural networks of identical design (detailed in Figure 2.1). A fully-connected neural network is the most basic form of neural network. Each neural network that we use has an input layer which receives the input sample, two intermediate “hidden" layers of nodes with eight nodes each, and an output layer with two nodes that classifies which of two categories the input is associated with. This type of network is commonly known as a classifier. The inputs for our examples are vectorized maps (i.e. images) of geospatial phenomena and are labeled with a two-unit vector that describes which of two categories, or classes, the image is associated with. Within the two-unit labeling vector, a 1 is placed in the index that the sample is associated with and a 0 is placed in the other. The output of the neural network is also a two-unit vector which represents the neural network's estimation of the likelihood that the input sample belongs in each class such that the output vector always sums to 1, and is calculated using a softmax operator (see appendix for more details). If the neural network is more confident that a sample belongs in a particular class, then the output for the corresponding unit of the output vector will be closer to 1. The objective of the neural network is to output a two-unit vector that is as similar to the label vector as possible, which means it is tasked with maximizing its confidence that each input sample belongs in its labeled category. More extensive details of the neural network architecture and training procedure are provided in the appendix.

It is worth noting that we use a basic form of a neural network for our examples, but could have chosen more advanced architectures such as convolutional neural networks (CNNs, e.g. Kriszhevsky et al., 2012). The neural networks we employ are relatively shallow in that they have few layers, whereas it is becoming more common to use “deep" neural networks with many layers. However, the intent of this paper is to present the usage of the interpretation of neural networks as a tool for scientific inference and not to showcase the utility of various neural network architectures. We therefore opt to keep the networks as simple as possible. In addition, we will show that this basic network architecture is sufficient to capture the known relationships between the inputs and outputs of our examples. The interpretation methods we use also place some restrictions on the structures

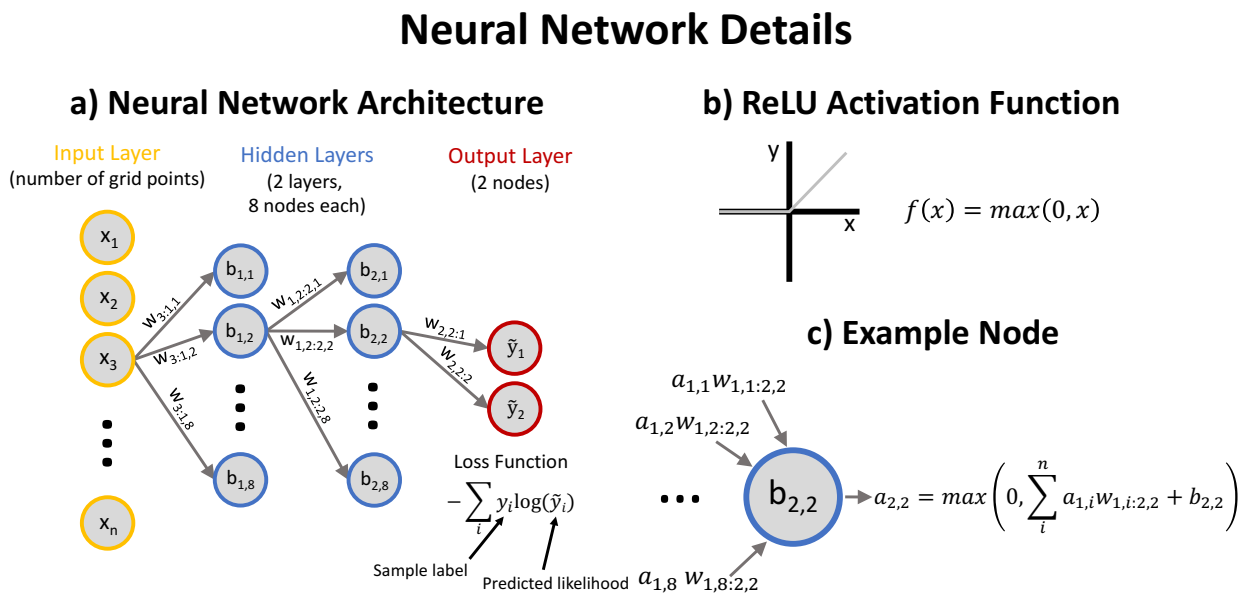

Figure 2.1: Illustration of the neural network architecture used in this study.

of the neural networks, the details of which are discussed in the subsequent sections, and so our neural networks abide by these requirements. With that said, the interpretation methods we discuss here are also applicable to a variety of other neural network architectures.

2.3 Neural Network Interpretation Methods

2.3.1 Backwards Optimization (Optimal Input)

The technique called backwards optimization calculates the input that maximizes a neural network's confidence in its output, and we therefore refer to the generated pattern as the "optimal input" [51–53]. This method offers insights into which patterns the neural network thinks are most associated with a particular output by using the weights and biases of a trained neural network to iteratively update an input sample until it is most closely associated with a user-specified output of the network.

Once a neural network is trained, the weights and biases can be frozen, which means that they are no longer updated as the neural network sees new samples. So, in turn, the backwards opti-

mization method takes the reverse approach to how a neural network is trained, and rather than updating the weights and biases of the network itself, an input sample is iteratively updated given a trained neural network with frozen weights and biases. The fact that the optimized input has the same dimensions as the samples used to train the network is particularly useful and is helpful for determining which patterns within the input vector are most important for describing any relationships between the input and output variables. The optimized input can also be interpreted in the same units as the input samples used to train the network.

The backwards optimization method is illustrated in Figure 2.2, detailed in code in the supporting information, and proceeds as follows:

Method Input: User-defined output of a trained neural network
Method Output: An optimized input that shows the input pattern most closely associated with the user-defined output according to the trained neural network

Procedure:

1. A neural network is trained, and the weights and biases are frozen, which means that they are not updated when a sample is input into the neural network.

2. A desired output from the neural network is defined. For example, if the network is trained to identify whether a sample belongs in one of two categories, the desired output could be when the neural network is 100% confident that the input belongs in one of the two categories.

3. A sample is generated of the same shape as the samples used to train the neural network, but the sample is initialized as all zeros.

4. This all-zero sample is passed through the network, and the output is gathered. The output is then compared to the desired output, and the loss (i.e. error) of the all-zero sample is calculated with respect to the desired output. The loss function is the same function used to train the network.

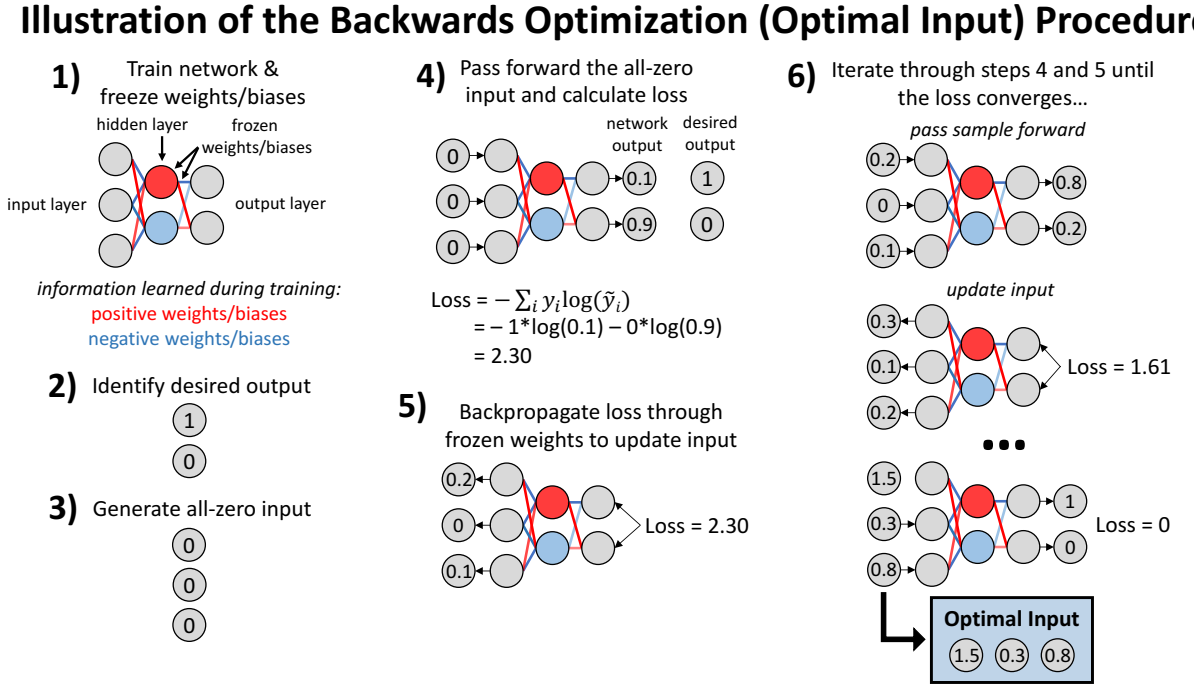

Figure 2.2: Illustration of the backwards optimization procedure used in this study for interpreting neural networks. The steps illustrated here correspond to the steps listed in Section 2.3.1. The neural network within this schematic has already been trained, and the training procedure is not illustrated.

5. The loss is translated backwards through the neural network to the input layer using back-propagation. But, rather than updating the weights and biases of the network along the way, the input sample itself is updated in a manner which reduces the loss using an increment of the information, or gradient, that was translated back to the input layer.

6. Iterate over steps 4 and 5 until the input is optimized such that iterations no longer reduce the error of the neural network's output.

Gagne et al. (2019) and McGovern et al. (2019) provide other examples of how the backwards optimization technique has been used in geoscience, and more specifically meteorology. We note that other techniques for the initialization of the unoptimized input sample have been suggested, such as using Gaussian noise rather than all zeros, but we have found that the optimized patterns are not sensitive to these initialization techniques for our examples.

As will be discussed throughout the remainder of this paper, the backwards optimization technique offers valuable insights into a neural network's decision-making process, but it is not without

its limitations. Briefly, the optimized input offers one composite perspective of the patterns the network looks for within the input data. This composite perspective introduces problems when applied to domains where, for example, multiple modes of variability may lead to the same outcome. In these cases, the optimal input may contain a combination of each mode, but will not elucidate how these modes may evolve either independently or in tandem with each other. There are ways that the backwards optimization method can be used for some of these applications too, however, such as by optimizing an actual input sample rather than an all-zero sample toward a target output from the neural network. We do not discuss this application here, but McGovern et al. (2019) briefly discuss such a technique.

Because of the complications of optimizing for a single optimal pattern, it is useful to also understand what information within each input sample is important for the neural network's associated output. Fortunately, there are methods for interpreting a neural network in this manner, one of which is called layerwise relevance propagation, which we discuss next.

2.3.2 Layer-Wise Relevance Propagation (LRP)

While backwards optimization has previously been used by the geoscience community, we are unaware of any published applications of layerwise relevance propagation to geoscientific problems, and so we go into additional detail describing this method. In contrast to the optimal input technique which generates a single optimized input given a desired output, layerwise relevance propagation (LRP) considers one input sample at a time. The form of LRP that we use was introduced to the computer science community by Bach et al. (2015). This form of LRP is also referred to as a "deep Taylor decomposition" of the neural network because of its relationship to Taylor series expansion [55], although the more general class of methods is referred to as LRP and we will therefore refer to the method as such.

For each input sample, LRP identifies the relevance of each input feature for the network's output, and therefore helps isolate which input features are important for a network's output on a sample-by-sample basis. For example, if the input is an image, the resulting output from LRP is

a heatmap in the dimensions of the original image that shows the regions of the image which are most important for generating the network's output for that particular sample. It bears repeating that the heatmap is specific to the input sample and so different inputs yield different heatmaps, the patterns of which depend on how the information from that input is transferred through the network as it makes its decision. LRP can be applied to any sample that is of the same dimensions as those used to train the network, even if the neural network did not see the sample during training.

Next, we generally describe how LRP traces the reasoning of a neural network's decision-making process, although we refer the reader to the manuscripts of Bach et al. (2015) and Montavon et al. (2017) for more details. We note that while the LRP methods presented by Bach et al. (2015) and Montavon et al. (2017) are one formulation of LRP, new formulations can be developed according to the more general guidelines posed within Bach et al. (2015).

The algorithm of LRP is illustrated in Figure 2.3 and proceeds as follows:

Method Input: An input sample
Method Output: The relevance of each feature within the input sample for the associated output of the neural network

Procedure:

1. A neural network is trained, and the weights and biases are frozen, which means that they are not updated when a sample is input into the neural network.

2. A sample is then input into the frozen neural network, and the output values are retained. If the neural network has categorical output and uses a softmax operator following the output nodes, then the output values prior to the softmax operator are retained. A single node of the output layer is identified as the node for which the relevance should be calculated. For cases of categorical output, this node is typically the one with the highest output likelihood for the given sample.

3. The output value of the single node is then propagated backwards through the network using information about the weights and biases of each node of the neural network. The propagation is done according to a particular set of propagation rules, which are discussed below. These rules depend on the types of the neural network and input data, and what type of information is to be inferred from the network.

4. This backwards propagation through the network is done until reaching the input layer. The resulting values have the same dimensions as the input and correspond to the relevance of each input feature for the neural network's decision of its output.

5. This process is completed for each sample of interest, from which the relevances for each sample can be studied independently or through composites or clusters of similar patterns of relevance.

An important aspect of LRP is the rules by which the relevance is translated backwards from the output layer toward the input layer. For our purposes, we only show the relevance propagation rules that are most fundamental to the theory of LRP. The rules that we use here, and which were introduced by Bach et al. (2015), have been constructed such that the total summed relevance after propagation back to the input layer is equal to the value of the output. For these rules, only information that $positively$ contributes to the output is propagated backwards, and negative weights and biases are therefore ignored. That is, only information that makes the network more confident in its categorical output is propagated backwards, and information that makes the network less confident is ignored. However, there are variants of LRP that permit the inclusion of information that reduces the network's confidence which are also useful for network interpretability, but extend beyond the scope of this paper [55].

We note again that LRP traces information for a single output node [54]. So, in the case of categorical output as we present within this paper, the relevance is propagated backwards for one of the categorical output nodes – typically the node with the maximum output likelihood for the sample of interest. If the neural network uses a softmax operator in its output layer, then during

the relevance calculations the softmax operator is ignored and the relevance is calculated for the network's output prior to the softmax. The softmax operator is helpful to ensure the network converges on a solution during training, but the pre-softmax output is more useful for interpretability purposes since it is an unscaled representation of the network's confidence in its output.

Once a sample has been input, passed forward through the network, and the output has been collected, the first step in LRP is to use the following propagation rule to pass the information backwards from the output layer to the previous layer of nodes:

$$R_i = \sum_j \frac{a_i w_{ij}^+ + max(0, b_j)}{\sum_i a_i w_{ij}^+ + max(0, b_j)} R_j. \tag{2.1}$$

Within Equation A.1, the i subscript represents the i-th node in the layer of the network to which the relevance is being translated backwards, the j subscript represents the j-th node in the layer of the network from which the relevance is being translated, R_i is the relevance translated backwards to the i-th node, R_j is the relevance of the j-th node, a_i is the output from the i-th node after the non-linearity has been applied when the sample is passed forward through the network, w_{ij}^+ is the weight of the connection between the i-th and j-th nodes where the $+$ signifies that only positive weights are considered, and b_j is the bias of the j-th node. The terms within this equation are illustrated schematically within Figure 2.3. As previously mentioned, the form of LRP that we use neglects all negative weights and biases and only traces information backwards through positive weights and biases. This rule in Equation A.1 is used to propagate the relevance backwards through the network from one layer to the next, starting with the output layer and extending backwards to the first hidden layer.

There are separate rules for translating information to the input layer from the first layer of hidden nodes, the rules of which depend on whether the values of the input features are bounded or unbounded. A case where the values are unbounded is when the data is standardized and so has zero mean and unit variance, but is not necessarily restricted from varying across all real numbers. A case where the values are bounded, on the other hand, is when all the input values are normalized

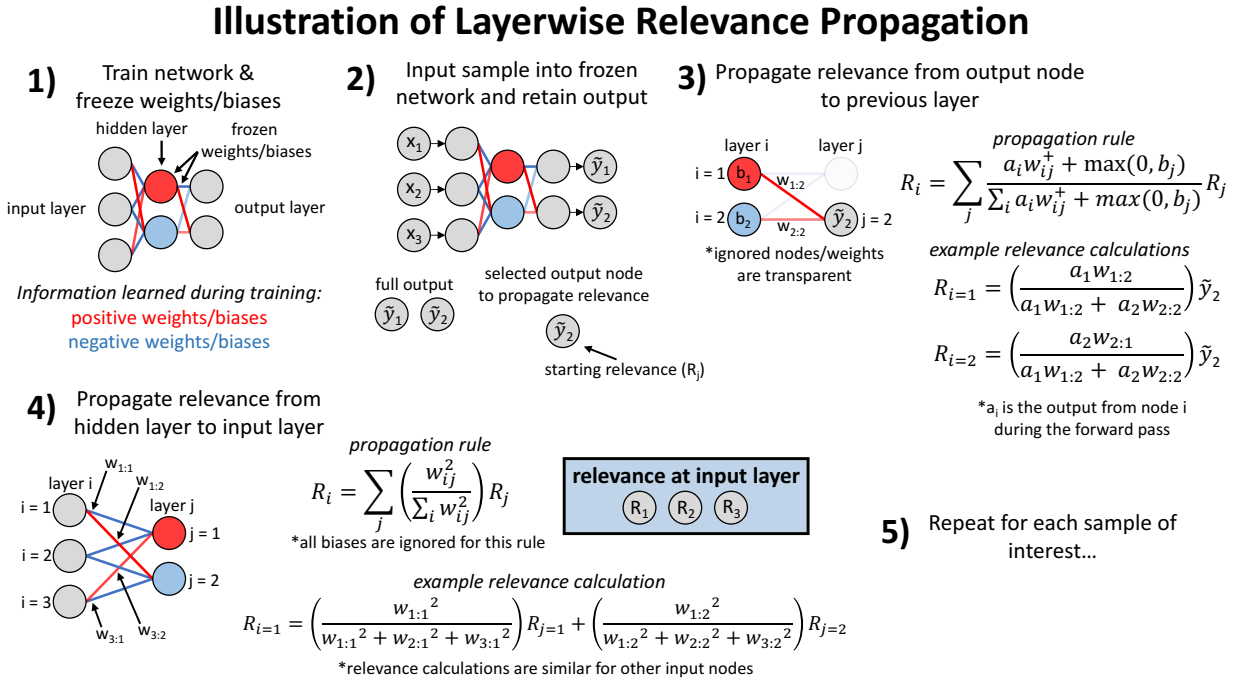

Figure 2.3: Illustration of the layerwise relevance propagation (LRP) procedure used in this study for interpreting neural networks. The steps illustrated here correspond to the steps listed in Section 2.3.2. The neural network within this schematic has already been trained, and the training procedure is not illustrated. While the illustration does not show the propagation from one hidden layer to another hidden layer, the associated propagation method is identical to the propagation from the output node to the final hidden layer shown in step 3.

between 0 and 1. For the case where the input values are unbounded, the rule for translating the relevance from the first hidden layer to the input layer is:

$$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j \tag{2.2}$$

where all terms are as previously discussed for Equation A.1. We use unbounded input data within our examples, and so we provide the propagation rule for the case of bounded data within the supporting information. Additional information about other propagation rules is available within Samek et al. (2019).

The rules for LRP presented within the literature have thus far been formulated for a specific subset of activation functions, types of neural networks, and neural network tasks. The rules that we present have been developed to work best with the Rectified Linear Unit (ReLU) activation function, since they test whether a node has been "activated" or not [54, 55]. Neurons that use the ReLU activation function are activated in the sense that their output is equal to the input if the input is greater than zero, but is zero if the input is less than zero (see Figure 2.1b for an illustration of the ReLU function). So, the formulation of LRP that we use ensures that it only traces information back through the network if the nodes are activated and therefore pass information forward when the neural network is making its decision for a particular sample. If the i-th node is not activated during the forward pass through the network, then the a_i term is zero in Equation A.1, the relevance for the unactivated neuron i is zero, and the relevance is distributed to the other activated neurons within that layer of nodes.

As we have discussed, we use a form of LRP that only propagates information that positively contributes to the output node, which means that the relevance heatmaps show regions that contribute to increases of the output likelihood that a sample belongs to a particular category. This interpretation is helpful for classification tasks, when increasing the likelihood that an input belongs in a particular category is of interest. There are limitations to this approach for regression problems, however, where it is desirable to understand which inputs cause an increase *or* decrease

in the final output. For this reason, we have found that the formulations described by Bach et al. (2015) are not well suited for interpreting neural networks tasked with regression, and we therefore suggest that an LRP formulation needs to be developed specifically for regression problems. However, there have been examples of using LRP for regression problems in other fields [57], and so while LRP may similarly be a viable approach for regression problems in geoscience, care should be taken in how the interpretations are used.

In addition, this formulation of LRP works well for fully-connected neural networks (as we use in this study) and convolutional neural networks, for which the propagation rules are similar [58]. There have been efforts to expand LRP to more complicated neural network architectures, but in these cases other propagation rules need to be used [59]. It is therefore critical that the neural network architecture be carefully considered prior to training if LRP is to be used.

Additional propagation rules for other cases, such as when negative relevances are to be considered, can be found in the supporting information of this paper or within Montavon et al. (2017) and Samek et al. (2019). We use an implementation of LRP from the authors of the method, which is described in detail within Alber et al. (2019), although an abundance of similar implementations also exist. The implementation we use is available as the $innvestigate$ package within Python, which has been written to work with the $Keras$ neural network package. Tutorials covering how to implement LRP within other programming languages are available at $heatmapping.org$, and a list of other resources for LRP in $Keras$ and other Python packages is offered within the supporting information.

While there are limitations to LRP, neural networks can be thoughtfully constructed to mediate some of these limitations. For example, many problems of regression can be reformulated as categorical problems by discretizing a continuous output into a number of categories. Additionally, many tasks in geoscience do not seem to require exceedingly complex neural network architectures [27, 33, 40], and in many cases a basic form of neural network is sufficient to attain high accuracy. Therefore, while the current formulations of LRP do not solve all the limitations of interpreting neural networks for geoscience, we show throughout the remainder of this paper that it still offers

opportunities for interpreting neural networks that are thoughtfully constructed with the ultimate objective of interpretation in mind.

2.4 Applications to Earth System Variability

To illustrate how the interpretation of neural networks can be used to advance scientific knowledge, we apply the backwards optimization and LRP methods to two well-known patterns of climate variability within the earth system. We intentionally choose patterns that have been extensively researched by the earth system/climate community, because our intent is to demonstrate the usage of neural networks for scientific inference by first showing that the techniques can replicate what we already know before extending into the unknown. Our aim is to provide readers with the intuition and confidence to use the techniques for their own research questions.

For our examples, the inputs to the neural networks are vectorized geospatial fields, the domains of which are discussed in their respective subsections. The neural network is tasked with identifying which of two categories the input geospatial fields are associated with, and what the categories represent depends on the example. It is worth noting that backwards optimization and LRP can be applied to neural networks with any number of output categories, but we limit the output to two categories for the sake of illustration. Additional details about the neural network architectures we use are discussed in Section 2.2 and the appendix.

2.4.1 The El Niño-Southern Oscillaton (ENSO) Pattern

The first example we use is the simpler of the two, and shows how the backwards optimization and LRP methods can be used to interpret a neural network's understanding of the spatial structure of a well-known climate pattern. We show that backwards optimization is useful for gaining a composite interpretation of the neural network's understanding of the climate pattern, and that LRP extends beyond this composite and also allows the interpretation of what information is useful to the neural network within each individual sample. This example is intentionally simple so we can

test the abilities of the interpretation techniques, rather than gain new knowledge about the climate pattern itself.

A neural network is tasked with identifying whether a sea-surface temperature (SST) pattern is characteristic of a positive (El Niño) or negative (La Niña) phase of the El Niño Southern Oscillation (ENSO). ENSO is a dominant mode of earth-system variability that acts on an interannual timescale and manifests as sea-surface temperature anomalies within the tropical Pacific, although its indirect influences on weather and climate are global [61, 62]. We define the state of ENSO using the conventional Niño3.4 index, which is a spatial average of the sea-surface temperature anomalies within the equatorial Pacific Ocean (between 5°S to 5°N and 170°W to 120°W). We calculate the spatial average using the 1° by 1° Cobe V2 dataset [63]. According to this index, negative sea-surface temperature anomalies within the east-central tropical Pacific are characteristic of La Niña, while positive sea-surface temperature anomalies are characteristic of El Niño. Composite sea-surface temperature anomalies for each phase are shown in Figure 2.4.

For the neural network setup (shown in Figure 2.5), the first index of the label vector corresponds to La Niña samples and the second index to El Niño samples. An example vector label for a La Niña case is therefore $[1, 0]$, and the output of the neural network is of similar form with the output value in each index corresponding to the network's estimated likelihood that the sample belongs in each category. The input dataset is monthly sea-surface temperature anomalies for the years 1880 through 2017 from the 1° by 1° Cobe V2 dataset [63]. We calculate the anomalies separately for each grid point by removing the mean for the years 1980 through 2009 and thereafter removing the linear trend. Samples from the years 1880 through 1990 are used to train the network and those from 1990 through 2017 are used to test the network, and we only test and train on months during which the Niño3.4 index magnitude was greater than 0.5. The network does not see the 1990 through 2017 samples during training, and those samples are only used to test whether what the network learns during training generalizes to samples on which the network was not trained. We vectorize the global images of sea-surface temperature anomalies before inputting them into the neural network.

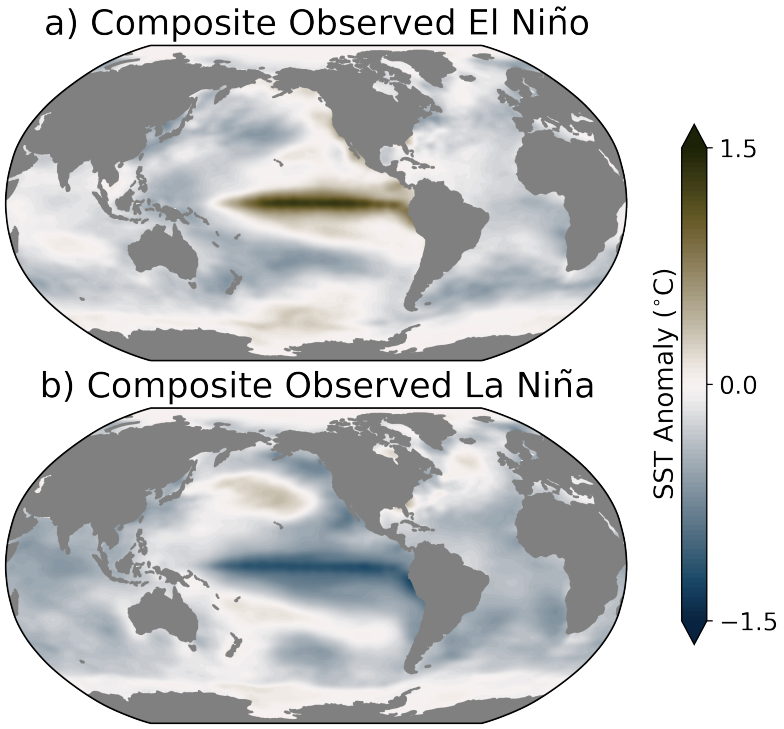

Figure 2.4: Composites of the monthly sea-surface temperature anomalies during (a) El Niño (337 samples) and (b) La Niña (485 samples). The composites include all events with a Niño3.4 index magnitude of greater than 0.5.

We also compare the results to linear regression to verify that the neural network is capturing physically reasonable patterns, since the sea-surface temperature signal of ENSO is predominantly linear although does exhibit nonlinearities [64, 65]. For this approach, we first obtain a map of regression coefficients by regressing the time series of global sea-surface temperature anomaly maps onto the Niño3.4 index time series. We then project this map of regression coefficients onto the observed sea-surface temperature anomalies to identify the ENSO phase.

The trained neural network identifies the ENSO phase with 100% accuracy on both the training (654 samples) and testing (168 samples) datasets. It is expected that the neural network would have nearly perfect accuracy given the intended simplicity of this example, which we use to illustrate the usefulness of the interpretation techniques. Regardless, in order to achieve this accuracy, the weights and biases of the neural network must contain information about the spatial patterns of sea-surface temperature variability characteristic of ENSO. The linear regression approach is

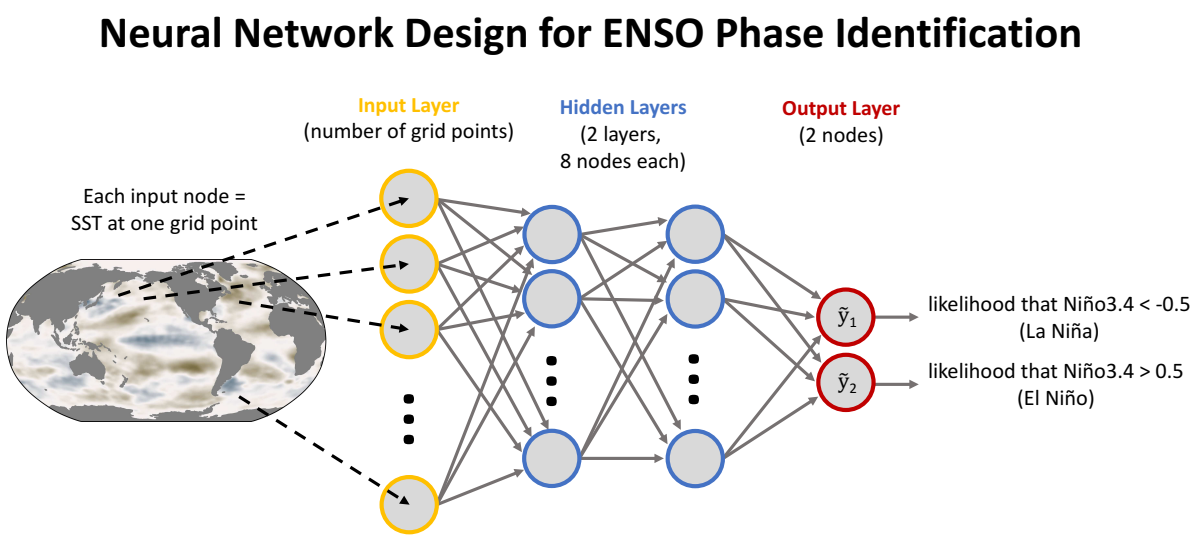

Figure 2.5: Illustration of the neural network design for ENSO phase identification.

accurate for only 82.5% of samples, and this lower accuracy is likely caused by noise within the global inputs. That is, with enough input samples, the linear regression should determine that the bounding box used to define the Niño3.4 index (between 5°S to 5°N and 170°W to 120°W) is the most useful region and ignore the remainder of the globe. To support this idea, the linear regression approach is 100% accurate when only sea-surface temperatures from this box are used as inputs.

We focus on the interpretation of the neural network's understanding of El Niño, although the interpretation for La Niña is similar and provided in the supporting information (Figure S2). We first generate the optimal input to identify the composite spatial pattern of sea-surface temperature anomalies that maximizes the network's confidence that the sample is an El Niño event (Figure 2.6a) and the composite relevance heatmaps for all of the El Niño samples (Figure 2.6b). Then, we use LRP to identify the regions on which the network focuses its attention for El Niño events on a sample-by-sample basis (Figure 2.7). The relevance values output from LRP for each sample are normalized to range from 0 to 1 by dividing each heatmap by its own maximum relevance value. We do this so that the relevances for each sample are weighted equally when compositing the relevance across samples.

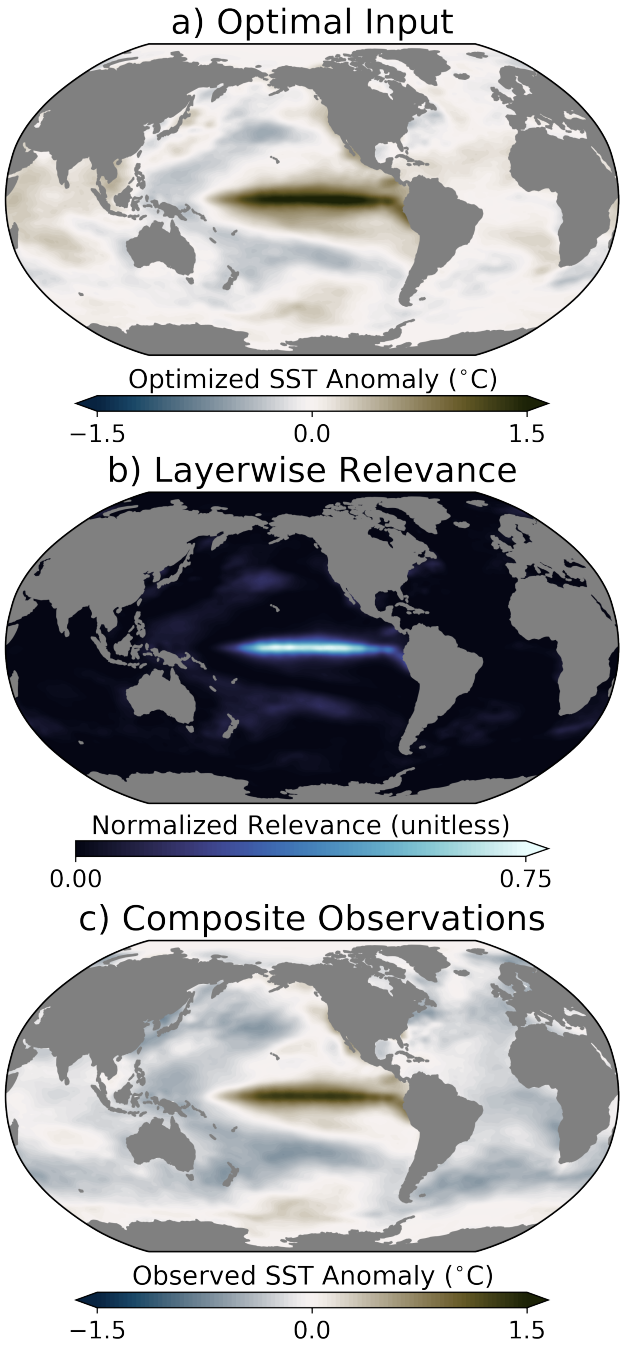

Figure 2.6: Interpretation of the neural network's understanding of the spatial structure of El Niño based on 337 total El Niño samples (including both training and testing data). a) The optimal input field that shows the input image that maximizes the confidence of the network that the sample is an El Niño event. b) The LRP composite for all El Niño events, where higher values denote greater relevance for the network's decision. Relevance values are normalized between 0 and 1 for each sample, such that 1 denotes the highest relevance in each individual sample and 0 denotes the lowest relevance. c) Composite observed monthly sea-surface temperature anomalies for all El Niño samples (Niño3.4 > 0.5), identical to what is shown in Figure 2.4.

Backwards optimization recovers a map of sea-surface temperature anomalies that is similar to the observed ENSO pattern in both spatial structure and magnitude, particularly within the tropical Pacific (Figure 2.6a,c). There are some differences in the sign and magnitude of the anomalies outside of the tropical Pacific, such as in the Atlantic Ocean, although these regions are not conventionally considered to be a part of the predominant ENSO pattern and are also not highlighted to be important to ENSO by the LRP relevance composites (Figure 2.6b) [61]. The composite relevance for the El Niño samples also shows that the neural network mainly focuses its attention on the tropical Pacific (Figure 2.6b). A region of non-zero relevance exists within the North Pacific (Figure 2.6b), which may be associated with a well-known correlation between oceanic variability within this region and the tropical signal of ENSO [66]. The linear regression coefficients are spatially similar to the optimal input pattern, which increases confidence in the robustness of the neural network visualization methods (Figure S1).

The utility of LRP is further highlighted by analyzing relevance heatmaps for individual samples. Figure 2.7a shows examples of eastern Pacific and central Pacific (i.e. Modoki; Ashok et al., 2007) ENSO events in 1998 and 1987 respectively, and highlights that the network refocuses its attention on different regions of the tropical Pacific to identify an El Niño event depending on the input. Furthermore, the neural network focuses its attention on the regions of sea-surface temperature anomalies that are most commonly associated with the two types of El Niño, and learns to ignore other anomalies of similar magnitude within the western Pacific that are distinct from ENSO. We only show the spatial relevance patterns for these two examples, although the relevance time series for the central and eastern Pacific show that the network correctly refocuses its attention for all of the input samples depending on the type of El Niño event (Figure 2.7c). Samples associated with central Pacific El Niño events have higher relevance within the central Pacific than within the eastern Pacific, and vice versa for samples associated with eastern Pacific El Niño events (Figure 2.7c).

We have shown that the neural network learns the physical structures of the various modes of ENSO, which lends confidence that backwards optimization and LRP can be used to better

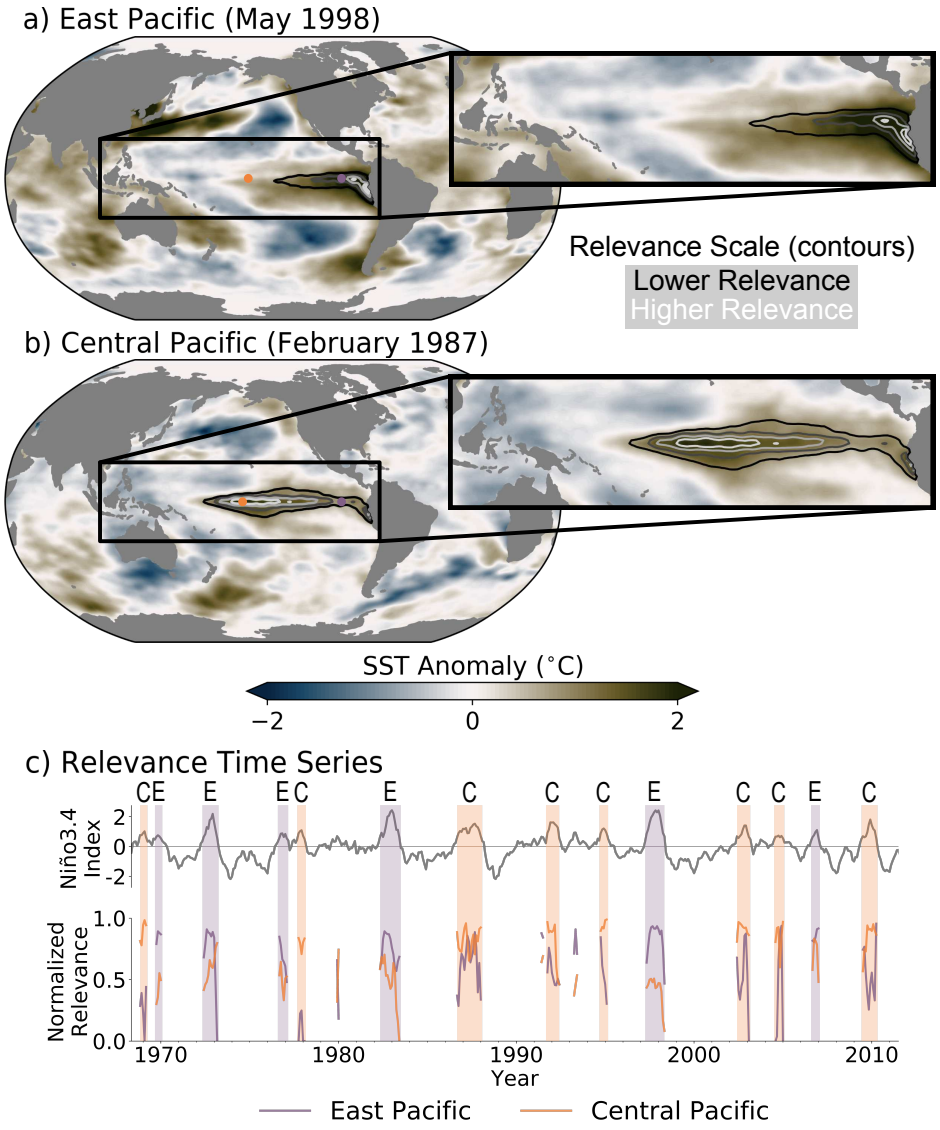

Figure 2.7: An illustration of how the neural network focuses on different regions of sea-surface temperature anomalies for different types of El Niño: a) an eastern Pacific El Niño event and b) a central Pacific (Modoki) El Niño event. The observed sea-surface temperature anomalies for each case are shown in fill and the LRP relevance is contoured. The relevance has been normalized to lie on a scale from 0 to 1, and the contours range in value from 0.2 to 1.0 in increments of 0.2. Relevance values less than 0.2 have been omitted. c) (top) The Niño3.4 index time series from 1968 to 2011; (bottom) Time series of the normalized relevance values for locations within the central Pacific and eastern Pacific from 1968 through 2011. Relevance values are only shown for months during which the Niño3.4 index was greater than 0.5. The central (eastern) Pacific location is denoted by the orange (purple) dot in panels a and b, and is located on the equator at a longitude of 200° (250°). The types of each El Niño event during the 1968 through 2011 period are as labeled in Ashok et al. (2007), Lee and McPhaden (2010), and Wang and Wang (2014), and are denoted above the time series as either central ("C") or eastern Pacific ("E") events. If an event was not determined to be separable into a central or eastern Pacific event by Ashok et al. (2007), Lee and McPhaden (2010), or Wang and Wang (2014), then it is not labeled.

our understanding of other patterns of earth system variability. This example also highlights the capability of LRP to identify what information a neural network uses in its decision-making process for each individual sample. The earth system rarely behaves according to a composite, and so the ability to analyze which aspects of each individual sample are important for the neural network's associated output is particularly useful for gaining new insights into earth-system variability.

2.4.2 Seasonal Prediction Using the Ocean

To further illustrate the usefulness of the backwards optimization and LRP methods, we next extend their usage to a slightly more complex example in which we train a neural network to predict a surface temperature response to sea-surface temperature anomalies months in advance. We focus on seasonal prediction, for which the ocean is a predominant source of atmospheric predictability [67–69]. Specifically, while it is well known that ENSO is a dominant contributor to atmospheric seasonal predictability [70, 71], there are other regions of oceanic variability that offer extended atmospheric predictability. One such region is the North Pacific, which can impact surface temperature and precipitation across North America [72–74]. We therefore predict continental surface temperature anomalies along the west coast of North America, which is more complicated than predicting the phase of ENSO since the neural network must identify the numerous coincident patterns of sea-surface temperature anomalies across different spatial and temporal scales that can contribute to seasonal temperature predictability.

As shown in Figure 2.8, we train the neural network to predict the sign (above or below zero) of surface temperature anomalies at a location along the west coast of North America (50°N, 240°E) using maps of sea-surface temperature anomalies within the tropics and Northern Hemisphere (north of 20°S). Surface temperatures at the chosen location, which is denoted by the red dot in subsequent figures, have previously been shown to have extended predictability due to sea-surface temperature forcing on seasonal to annual timescales [72, 75]. We input sea-surface temperature anomalies from the 1° by 1° Cobe V2 monthly sea-surface temperature anomaly dataset that is linearly interpolated onto a daily basis [63], and we use the years 1950 to present day. The corre-

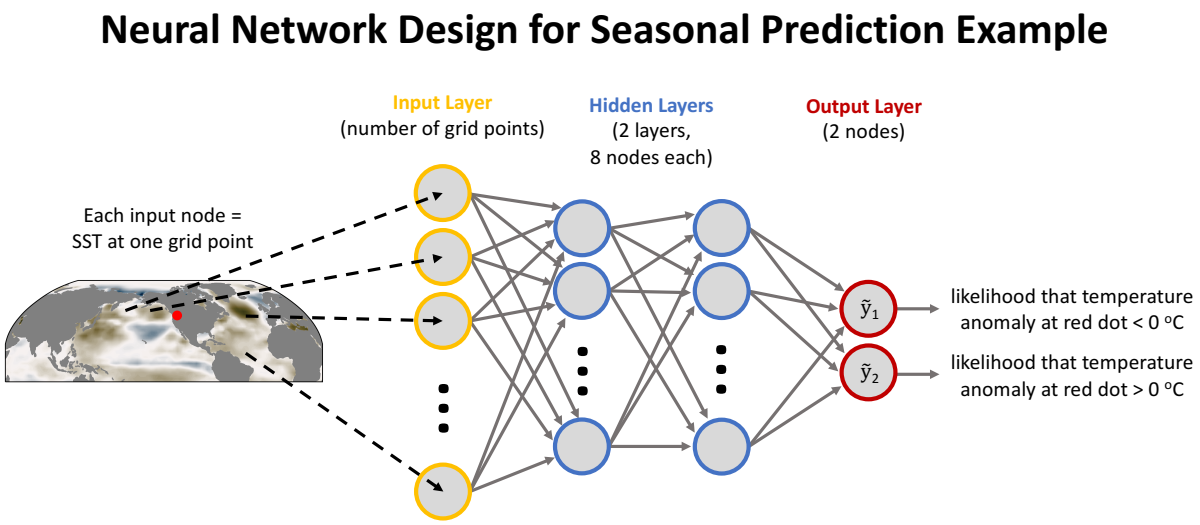

Figure 2.8: Illustration of the neural network design for the seasonal prediction example.

sponding daily surface temperature anomaly labels are gathered from the Berkeley Earth Surface Temperatures (BEST; Rohde et al., 2013) dataset, also spanning from 1950 to present day. For both the sea-surface and continental surface temperatures, we calculate the anomalies separately for each grid point by subtracting the mean values for the years 1980 through 2009 and thereafter removing the linear trend. The training dataset spans from 1950 through 2000 ($\sim$18,000 samples), and the testing dataset spans from 2000 through 2018 ($\sim$7,000 samples). The surface temperature anomalies are averaged over a 60-day period to ensure the predictions are capturing longer-term surface temperature variability, and the averages are centered such that a prediction with a lead time of 60 days implies a prediction of the average 30- to 90-day surface temperature anomalies.

We use interpretations of the neural network to identify which sea-surface temperature patterns are useful for making extended surface temperature predictions at various prediction lead times. We first train a neural network to predict the sign of the 30- to 90-day average surface temperature anomalies (i.e. a 60-day lead time using our definition), for which the network has 67% accuracy. We then focus on interpreting the neural network for cases when the surface temperature anomalies are positive, although the interpretation for the cases with negative anomalies is similar and provided within the supporting information (Figure S3). For this lead time, the optimal

input and LRP composite identify similar regions of SST patterns that lend predictability across the tropical Pacific and North Pacific (Figure 2.9a,b). Both of these regions have been identified by previous studies as sources of seasonal temperature predictability for the west coast of North America [71, 72, 75].

We next test the fidelity of the neural network interpretations by varying the prediction lead time of the continental surface temperature anomalies from 180 days prior to 60 days following their occurrence. We compare the neural network interpretations with that of linear regression to test whether the interpretations are reliable and if they offer any unique insight compared to more conventional approaches. Our linear regression approach is similar to the approach used for the ENSO example. We first obtain a map of regression coefficients by regressing the time series of global sea-surface temperature anomaly maps onto the time series of surface temperature anomalies over the west coast of North America. We then project the regression coefficient map onto the global maps sea-surface temperature anomalies to predict the sign of the surface temperature anomaly. The resulting accuracies of both prediction methods and the associated sea-surface temperature patterns that lend predictability are shown in Figure 2.10.

At extended leads, the spatial patterns of sea-surface temperature anomalies identified by backwards optimization and LRP are similar to those identified by regression (Figure 10). Particularly, the tropical Pacific stands out as being a predominant source of surface temperature predictability across the 180-day, 120-day, and 60-day prediction lead times for both the neural network interpretation and the regression maps (Figure 2.10a,b,c). For the 60-day prediction lead time, within the neural network interpretations the importance of the North Pacific begins to increase relative to the ENSO region, and the North Pacific becomes the dominant source of predictability for the concurrent and 60-day lagged sea-surface temperature anomalies (Figure 2.10c,d,e). Unlike the neural network, the regression approach continues to highlight the tropical Pacific Ocean as important for identifying the concurrent and 60-day lagged surface temperature anomalies.

The neural network is more accurate than the regression approach for all prediction ranges, which suggests that the neural network interpretations likely capture the sea-surface temperature

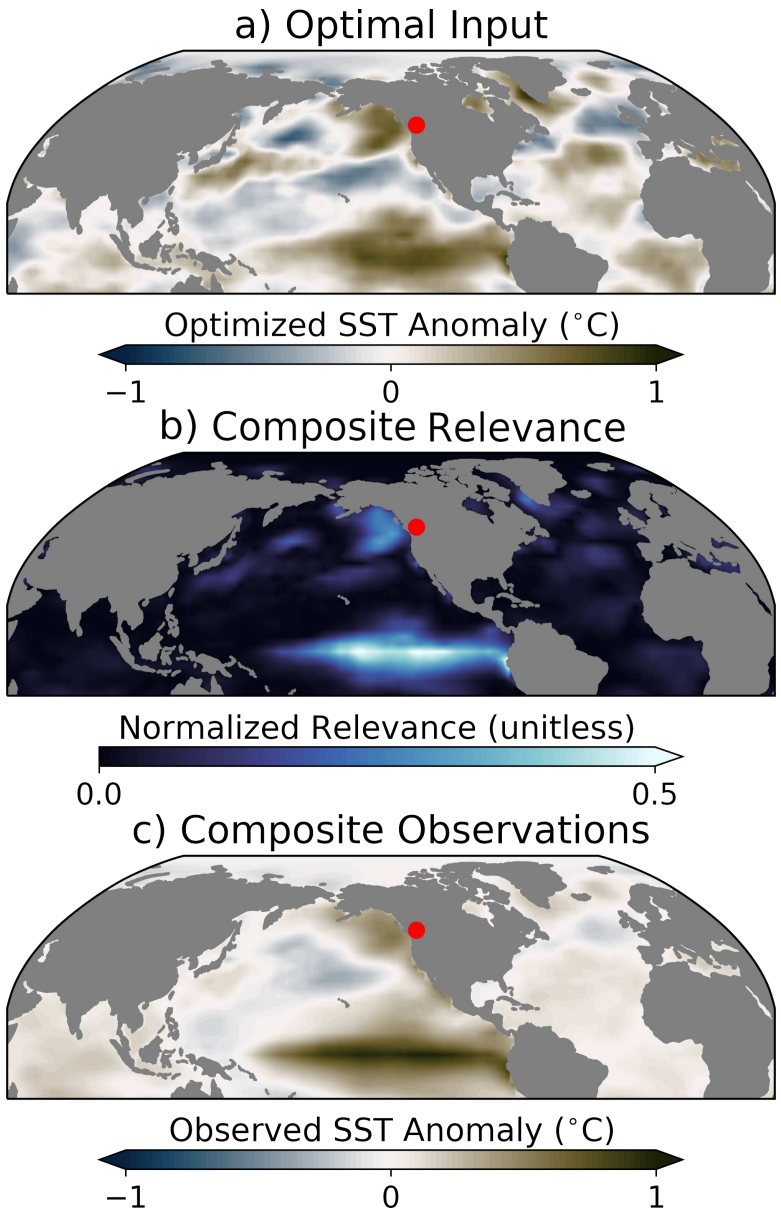

Figure 2.9: Interpretation of the neural network tasked with predicting 30- to 90-day average surface temperature anomalies at the red dot based on ∼12,000 total samples (including both training and testing data). Only the interpretation for positive surface temperature anomalies is shown, and the interpretation for negative anomalies is shown in Figure S3. a) The optimal input field that maximizes the network's confidence that the input sample is associated with positive temperature anomalies at the red dot. b) The LRP composite for all correctly categorized samples of positive temperature anomalies, where higher values denote greater relevance. Relevance values are normalized between 0 and 1 for each sample, such that 1 denotes the highest relevance in each individual sample and 0 denotes the lowest relevance. c) Composite observed sea-surface temperature anomalies for all cases where the neural network accurately predicts positive surface temperature anomalies.

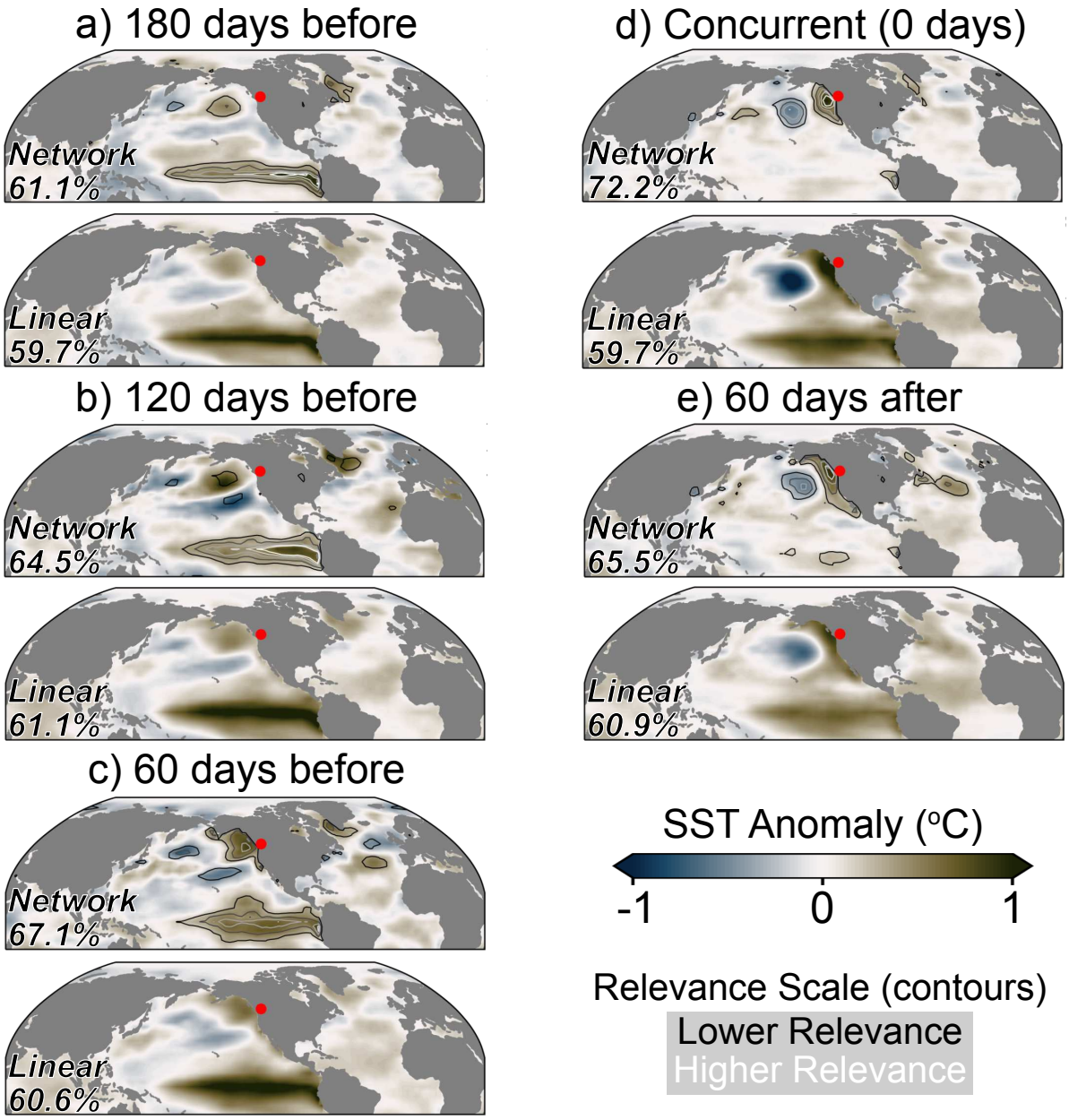

Figure 2.10: A comparison of the spatial patterns of sea-surface temperature deemed important for predicting surface temperature at the red dot using neural networks and linear regression. An evolution of the sea-surface temperature patterns at various lead times is shown, ranging from 180 days prior to the surface temperature anomalies to 60 days afterwards. The prediction is made for surface temperatures averaged across a 60-day window, and the prediction lead time listed above the sub-figures is the center of this window. So, for example, the 180-day lead time prediction is actually a prediction of the 150- to 210-day average surface temperature. For each lead/lag, the top panel shows the neural network optimal input in fill and LRP relevance in open contours, and the bottom panel shows the regression coefficients for the linear regression approach. The open contours denote LRP relevance values ranging from 0.1 to 0.3 in increments of 0.05.

patterns more closely associated with the seasonal surface temperature anomalies. Specifically, the neural network interpretations suggest that the North Pacific is the predominant modulator of concurrent surface temperature anomalies along the west coast of North America, while the tropical Pacific offers extended lead predictability (Figure 2.10). This idea is corroborated by previous research that found the North Pacific modulates temperatures across western North America separately from the tropical Pacific [72]. So, while the neural network is only slightly more accurate than the linear regression model, the increase in accuracy is caused by an improved understanding of the most relevant sea-surface temperature patterns. Either nonlinearities or the increased pathways for information to flow through the neural network likely contribute to this improved understanding.

2.5 Discussion and Conclusions

The recent surge in the popularity of neural networks within the geosciences has inspired the need for techniques to interpret their decisions. Neural networks are conventionally thought of as “black boxes" within the geosciences with limited tools for the interpretation of the reasoning behind their decision-making process. We have shown that the usage of two separate techniques enables physically meaningful inference from thoughtfully designed neural networks. This ability to reliably interpret neural networks opens the door to using the interpretation of how and why the network makes its decisions as the ultimate science outcome.

The backwards optimization method can be used to quantify the patterns within the input data that maximize a neural network’s confidence that an input is associated with a particular output. For the case of categorical output as we present within this paper, backwards optimization iteratively changes an input to maximize the neural network’s confidence that it belongs in a particular category. The optimized input has the same dimensions and can be interpreted in the same units as the input samples used to train the network, but provides no direct indication as to which characteristics of the optimized input are most important. In general, however, backwards optimization is useful for identifying the dominant pattern of variability the neural network looks for when making

its decisions. In our examples of ENSO phase identification and seasonal prediction, backwards optimization was able to extract the dominant modes of variability known to be associated with each problem (Figure 2.6; Figure 2.10).

Layerwise relevance propagation (LRP), on the other hand, considers each sample individually, and provides information about the characteristics of each sample that are most important, or relevant, for the network's associated output. LRP can thereby provide insights into how relationships between the inputs and outputs of a neural network vary on a case-by-case basis. The usefulness of this quality is exemplified by comparing the relevance heatmaps for two types of El Niño events – the eastern Pacific and central Pacific, or Modoki, patterns (Figure 2.7). Although the optimal input pattern does not distinguish between these two modes of El Niño variability because it offers a composite interpretation (Figure 2.6a), LRP shows that the network does redirect its focus depending on where the sea-surface temperature anomalies occur (Figure 2.7). While we do not examine this capability within this paper, it is possible to cluster the LRP relevance heatmaps to identify secondary modes variability within each input category if there is no a-priori knowledge of their existence [77]. The fact that the neural network learns the variable spatial structures of ENSO, and that LRP can elucidate this understanding, suggests that LRP can be used to identify physically meaningful patterns within other geoscientific datasets, as well.

There are particular requirements of the backwards optimization and LRP techniques that constrain how a neural network is constructed, the details of which are discussed in Section 2.3. We therefore emphasize that neural networks must be constructed thoughtfully so as to maximize the scientific value of their interpretation. The network architecture must be complex enough to capture any existing relationships between the input and output data, but not so complex that interpretation methods are no longer usable, the balance of which depends on the use-case. The relative value of the accuracy and interpretability of a neural network is of critical importance to scientific analyses, and should be assessed carefully prior to training. For example, first training a simple neural network and building towards a more complex model enables an understanding of whether more complex and thereby less interpretable networks are necessary. If a network is

too simple to accurately capture the relationships between the input and output, then its accuracy will be low and any interpretations of its understanding will be limited in scientific value. On the other hand, if a network is too complex and interpretation is impossible, then its value is limited solely to its output. A balance between network complexity and interpretability must be struck if the interpretation of what a network has learned is to be scientifically useful.

We have shown that techniques for interpreting neural networks have the potential to extend their usage to the discovery of unknown patterns within geoscientific data, a concept which will be further explored in future research. The ultimate scientific outcome of a neural network can now also be the interpretation of what the neural network has learned, rather than only the output of the network itself. Regardless of the specific application, it is now apparent that neural networks offer scientists a useful new way to discover and understand connections within geoscientific data.

Chapter 3

Testing the Reliability of Interpretable Neural Networks in Geoscience Using the Madden-Julian Oscillation

3.1 Introduction

Neural networks have the potential to improve our understanding of the earth system in ways that are unique from other statistical and machine learning methods. Recent research within the geosciences has shown that neural networks can be used to accelerate climate model parameterizations [44, 45], discover patterns of earth-system variability [78], and make accurate global weather predictions [79], among numerous other applications in weather and climate [38, 80]. These advances have been rooted in the theory that neural networks are universal function mappers – that is, given a sufficient level of neural network complexity and quality of input data, a neural network can map any relationship between two datasets [81].

Neural networks may be particularly useful within the geosciences if the relationships contained within their learned parameters can be understood and interpreted. Numerous methods have been proposed for such interpretation within the computer science community, and have even been shown to be applicable to improving the understanding of geoscientific phenomena such as ENSO, sources of seasonal predictability, and severe convective storms [32, 40, 78, 80]. The critical caveat of using interpretable neural networks within geoscience is that the interpretations must accurately portray the relationships captured by the neural network and not mislead the scientist toward incorrect conclusions. Therefore, any interpretability methods should first be tested on topics that are well understood so that trust can be lent to studies that use the methods to discover entirely new patterns.

The Madden-Julian Oscillation (MJO) [5, 29] has been a focus of hundreds of publications across numerous decades, and although it is not fully understood, a few of its characteristics are commonly accepted by the scientific community. For example, its core characteristic is an anomaly in deep convection and associated cloud cover within the tropics that forms within the western tropical Indian Ocean and propagates eastward toward the tropical eastern Pacific ocean over the course of 3 0 to 60 days [29, 82, 83]. While we will focus on the tropical characteristics of the MJO, it is also generally accepted that the atmospheric response to deep convective heating within the MJO can generate teleconnection patterns across the globe [84–86]. The formation and propagation of the MJO are not as well understood, although numerous theories have been put forth [87], one of which suggests that the MJO propagates in response to gradients of tropical water vapor anomalies [88, 89]. Another theory suggests that the MJO could be a large-scale envelope of eastward and westward propagating gravity waves, and that its eastward propagation occurs because the eastward waves travel faster than the westward waves [90, 91]. Anomalies in atmospheric state variables that coincide with the MJO are also well documented [83, 92–94], although their relationship with the seasonality of the MJO is less clear, particularly given remaining uncertainties in mechanisms driving the seasonality of the MJO itself [95, 96].

We use the MJO as an opportunity to test whether interpretable neural networks can capture known patterns of variability within complex geoscientific data, and we then extend our analysis into inferring new information about the MJO itself. We also provide a new definition of MJO seasonality, for both the conventional outgoing longwave radiation definition and across atmospheric state variables. The aim of this paper is threefold: 1) to highlight the ability of neural networks to capture complex relationships within geoscientific data; 2) to test neural network interpretation methods to ensure they can reliably infer the relationships captured by neural networks; and 3) use the interpretations to gain new insights into the MJO. This paper thereby offers a conceptual guideline for how a geoscientist might go about using a neural network to discover new patterns within geoscientific data. Those interested in the MJO itself will also find new insights into its spatial structures and seasonality.

3.2 Data and Methods

We first discuss the data we use to define the MJO and then detail how we design a neural network to infer information about its spatial structure and seasonality.

3.2.1 Data

We define the MJO according to the Outgoing Longwave Radiation MJO Index (OMI) [97], which tracks the state of the MJO using anomalies in top-of-atmosphere outgoing longwave radiation (OLR) [98]. Increased cloud-cover inhibits the upwards ventilation of longwave radiation to space, so outgoing longwave radiation is generally used as a proxy for cloud cover in studies of the MJO. Some of the details of OMI are listed below, and it can generally be defined as a linear representation of the MJO based on outgoing longwave radiation anomalies with periods of 20 to 96 days. An important advancement of OMI beyond other MJO indices is that the structure of the MJO is calculated for each day of the year across a 121-day rolling window, and thereby accounts for seasonality. The index is constructed by calculating the two leading principal components in tropical (20°S to 20°N) outgoing longwave radiation anomalies, following the removal of the seasonal cycle and filtering the outgoing longwave radiation field to contain only eastward propagating waves with a periodicity of 30 to 96 days. The MJO also exhibits higher frequency modes of variability and occasional westward propagation [99, 100], so outgoing longwave radiation anomalies that include both eastward and westward propagating waves with periods of 20 to 96 days are then projected onto the 30- to 96-day principal components. This projection results in OMI including all eastward and westward propagating components of the MJO with periods of 20 to 96 days, with the caveat that they must coincide with the dominant, eastward propagating, 30- to 96-day mode of the MJO.

While the process of calculating OMI is complicated, the resultant phase-space and spatial perspectives of the MJO are relatively simple, as shown in Figure 3.1. A two-dimensional phase space is commonly used to define the phase and amplitude of the MJO, with each axis representing the two OMI principal components. As the MJO progresses, it completes a circle about its

two-dimensional phase space, which represents the eastward propagation of a spatially coherent dipole in outgoing longwave radiation anomalies (Figure 3.1a). The phase space is conventionally separated into eight octants for convenience, so the MJO is commonly studied according to its evolution across eight discrete phases. The phase of the MJO is based on the azimuth of the linear combination of the two principal components, and its magnitude is determined based on the distance of this point from the origin. An MJO event is generally considered to be “active" once the principal component magnitude is greater than 1, which is delineated by the red dots in Figure 3.1b. Because the principal components are standardized to have zero-mean and unit variance, MJO events of increasing amplitude become increasingly rare, such that most events have low amplitude and are clustered about the origin.

We test whether a neural network can identify the phase of the MJO given inputs of cloud characteristics and atmospheric state variables. The inputs to the neural network are tropical (30°S to 30°N), 20- to 96-day filtered fields of outgoing longwave radiation and 850-hPa, 500-hPa, and 200-hPa zonal wind, meridional wind, temperature, water-vapor mixing ratio, and geopotential (Figure 3.2a), and the outputs are the eight discrete phases of the MJO according to OMI. Only days during which the MJO was active are used (i.e. its principal component magnitude was greater than one). We use atmospheric state variables from the NASA MERRA-2 reanalysis [101] and outgoing longwave radiation from the NOAA once-daily outgoing longwave radiation climate data record [102], both spanning from January 1, 1980 through December 31, 2016. We remove the seasonal cycle, defined as the annual-mean cycle from all 37 years of input data, before applying a 20- to 96-day Lanczos bandpass filter with 121 weights and interpolating each variable onto a homogeneous 2° grid. The training data spans from January 1, 1980 through December 31, 2009, and the validation data span from January 1, 2010 through December 31, 2016. The training and validation data generally capture similar phase and amplitude distributions across each MJO phase (Figure 3.2b).

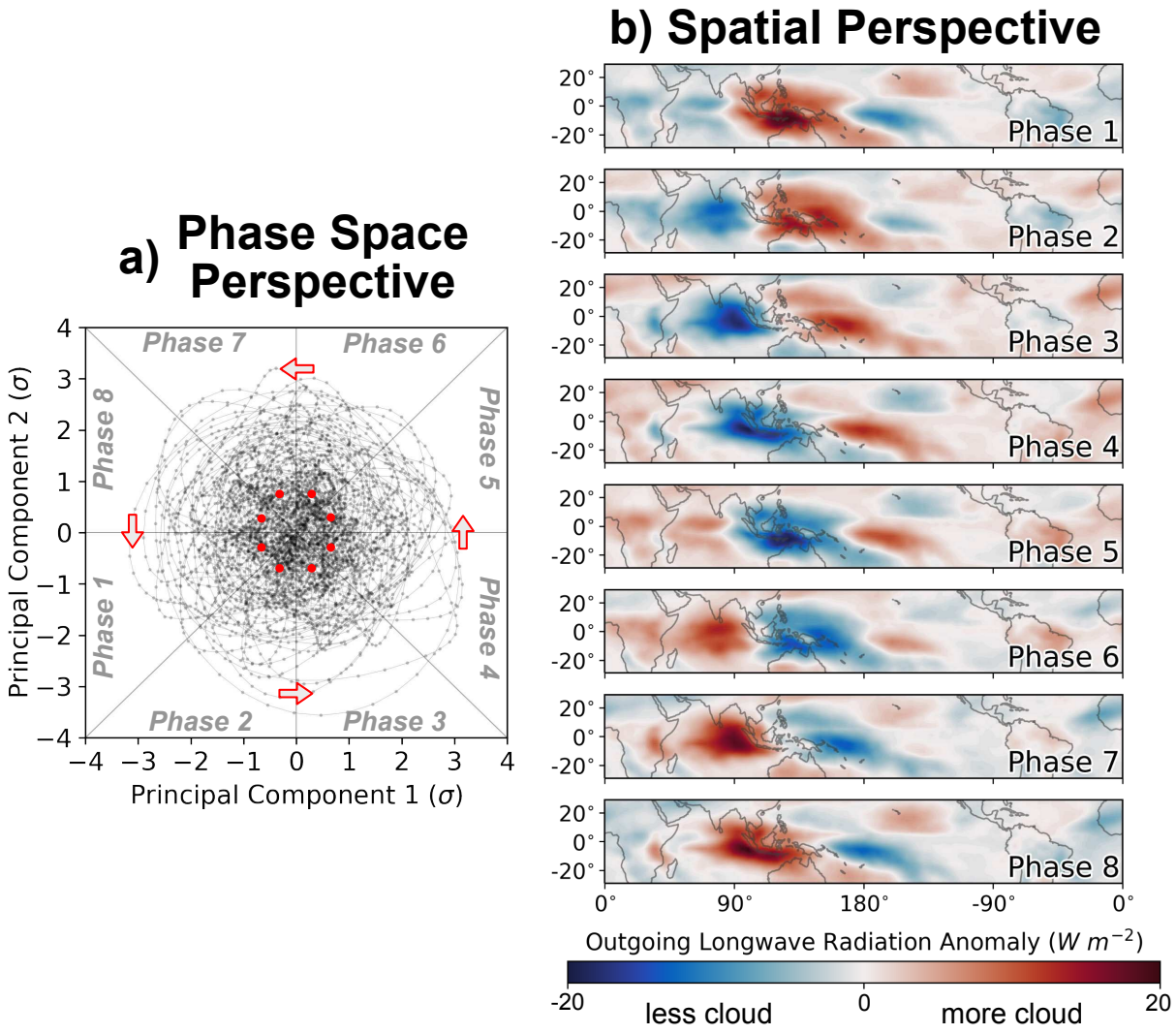

Figure 3.1: Spatial and phase-space perspectives of the Madden-Julian Oscillation. a) The phase space depiction of the MJO, again according to OMI for all MJO cases from January 1, 1980 through December 31, 2016.; (b) The spatial evolution of the MJO through its eight-phase phase space according to the outgoing longwave radiation MJO index (OMI)

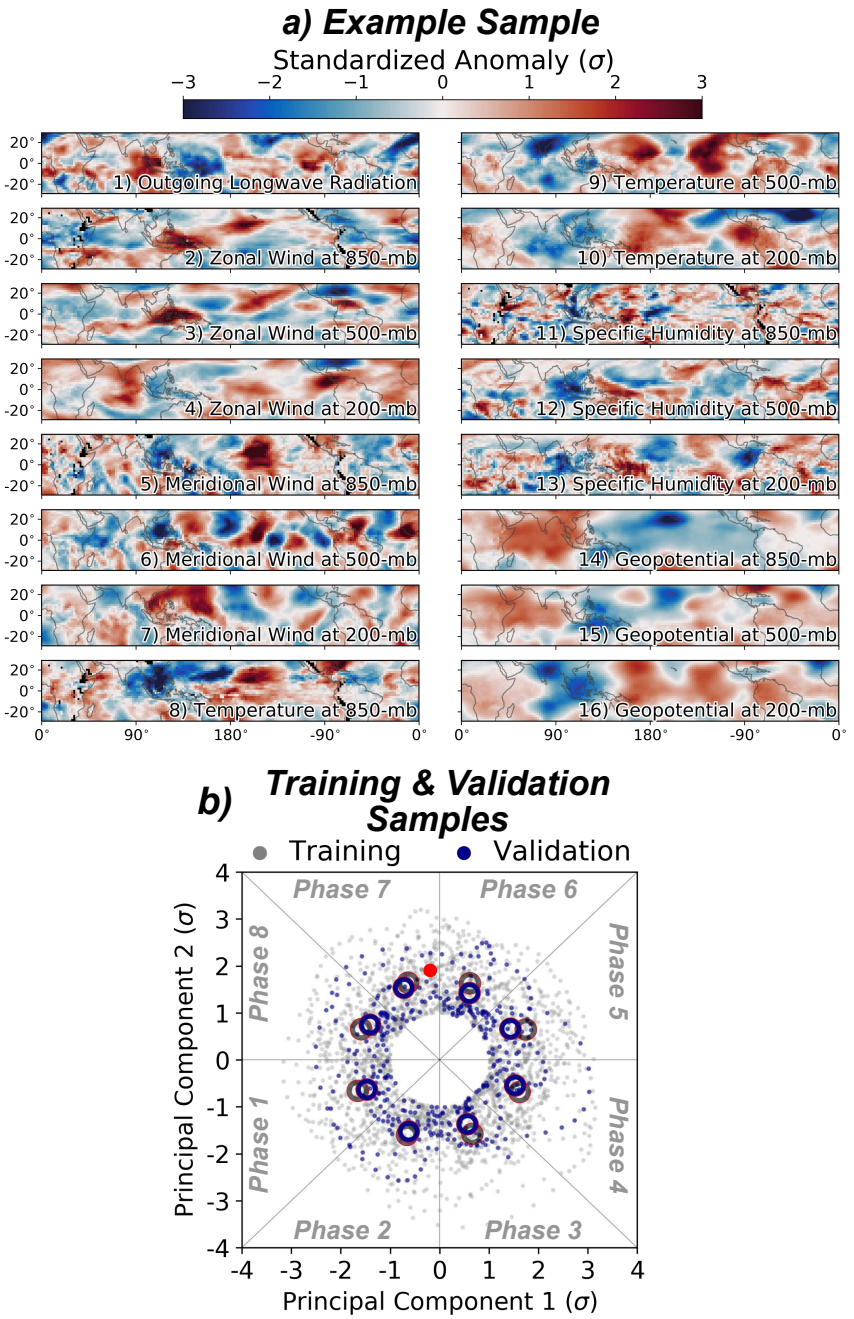

Figure 3.2: (a) An example input sample, which corresponds to a Phase 7 MJO day. Each variable was standardized for each grid point to have zero mean and unit variance across all samples from January 1, 1980 through December 31, 2016. (b) A visualization of how the samples are split between the training and validation datasets. The red dot corresponds to the sample shown in (a), the gray denotes denote the training samples, and the purple dots denote the validation samples. The gray rings denote the training sample mean phase and amplitude for each phase, and the blue rings denote the same but for the validation data.

3.2.2 Neural Network Design

We design a neural network to be as simple as possible, while still ensuring it can capture any relationships between the input atmospheric state variables and the phase of the MJO. We use fully-connected networks, which can be thought of as a chain of nonlinear regression functions that map the relationships between input and outputs datasets. The neural network has one input layer, two hidden layers with 64 and 128 nodes each, and one output layer with eight nodes, each of which represent a phase of the MJO (Figure 3.3). The hidden nodes all use the ReLu activation function, which applies the $max(0, x)$ operator to the output of each node. A softmax operator is applied to the output layer, which normalizes the output of the neural network such that the sum across all output nodes is equal to one. The outputs can therefore be thought of as a likelihood, with higher values for each node corresponding to a higher likelihood that the input sample belongs in that particular phase of the MJO. During labeling, each MJO event is labeled using an eight unit vector, and each unit represents one phase of the MJO. An input associated with phase three of the MJO would therefore have an output label of [0,0,1,0,0,0,0,0], which in the perspective of the neural network implies a 100% likelihood that the sample is associated with phase 3 of the MJO.

We train separate neural networks on data from 121-day bins centered on each calendar week of the year in order to study the seasonality of the MJO. Each neural network is therefore tasked with identifying the phase of the MJO according to the outgoing longwave radiation and state variable patterns during the period of the year to which it is assigned. Comparisons between interpretations of each neural network offer insights into the seasonality of the MJO, as discussed in subsequent sections.

Neural network interpretability generally becomes more challenging with increasing network complexity [58]. The neural network design we use is simple enough to enable robust interpretations, but complex enough to capture useful relationships between the input state variables and MJO phase. We find that decreasing the number of internal nodes reduces the accuracy, presumably because the network is then not complex enough to model the relationships between the atmospheric state variables and MJO. On the other hand, increasing the number of nodes also re-

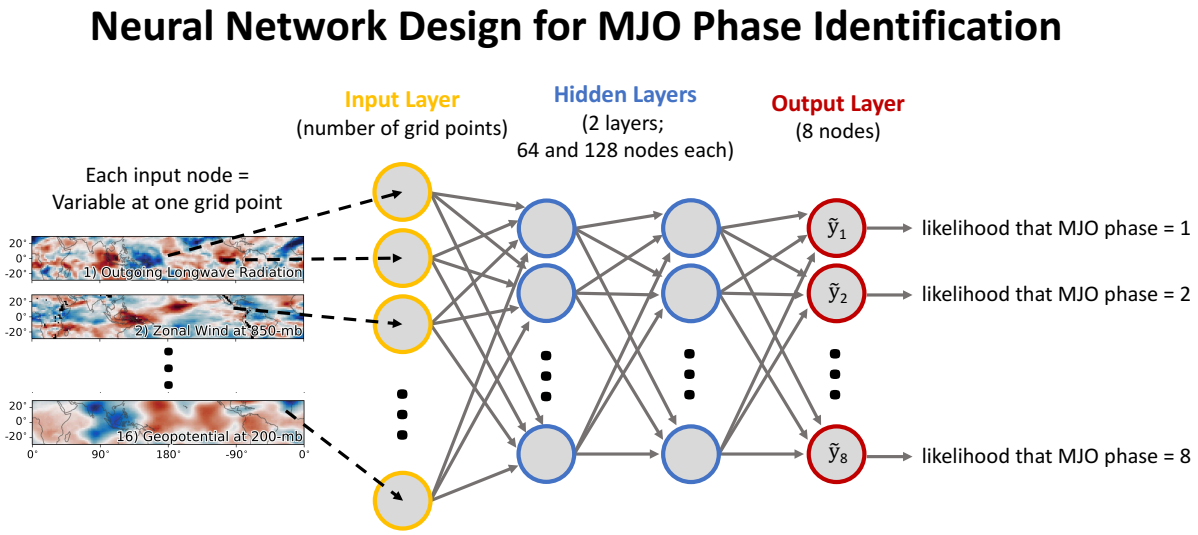

Figure 3.3: Schematic for the neural network used in this study. The first layer ingests vectorized input images, with two subsequent hidden layers the first with 64 nodes and the second with 128 nodes, and an output layer of 8 nodes that correspond to the eight phases of the MJO. A separate neural network is trained for each calendar week of the year.

duces the accuracy of the neural network on the validation dataset, because it is able to overfit on meaningless noise within the inputs using the additional weights and biases. We address any overfitting by applying L2-regularization to the weights connecting the input layer to the first layer of hidden nodes, which forces the network to focus its attention on broader spatial patterns within the inputs. We thoroughly tested the accuracy of convolutional neural networks (CNNs) for our particular problem, and found that the fully connected networks were both more accurate and interpretable than their CNN counterparts when L2 regularization is applied to the first layer of hidden nodes.

3.2.3 Neural Network Interpretability

The novelty of this paper is the demonstrated ability to interpret what the neural networks have learned, and to then gather scientific value from the interpretations. We use two interpretation methods which we briefly discuss here and are explained in more extensive detail in the context of geoscience within Toms et al. (2020). The two methods we use are called backward optimization

and layerwise relevance propagation, both of which map the decision-making process of the neural network onto the original input dimensions.

Backward optimization uses the same method that is used to train a neural network (i.e. back-propagation) to instead interpret what a trained network has learned [51–53]. Rather than updating the weights and biases of the network, the input itself is updated to minimize the difference between the network's associated output and a user-defined output. This process generates a single optimized pattern associated with a particular output, and thereby offers a composite interpretation of patterns contained within a neural network. In our case, we input blank (i.e. all-zero) maps, and optimize them to be most closely associated with a particular phase of the MJO. In doing so, we can identify the optimal patterns in each state variable for each phase of the MJO.

Layerwise relevance propagation (LRP) interprets the neural network's decision-making process for each individual input sample [54, 55, 58]. Given a trained neural network, an input sample is passed forward, the associated output is collected, and the unique pathways through which information flows from the input to the output for that specific sample are analyzed. The pathways are traced by propagating information backwards from the output layer to the input layer using rules specific to LRP. By tracing these pathways, the "relevance" of each input variable to the network's associated output can be quantified for each individual input example. The resultant relevance is unique to each input sample, because the pathways through which information flows through a neural network is similarly unique for each sample. A particularly important aspect of LRP is that the formulation of neural network that we use (i.e. fully connected networks with ReLu activation functions) conserves the relevance from the output layer to the input layer, meaning that all information important to the network's decision is included within the final LRP interpretation. LRP traces the information that $positively$ contributes to the output of the neural network, so is well suited to categorical output. So, in our case, LRP shows which regions of atmospheric state variables are most relevant to increases in the neural network's confidence that the sample belongs to a particular phase of the MJO.

3.3 Results

3.3.1 Neural Network Accuracy

We first ensure the neural networks are accurate enough to offer scientifically valuable interpretations. As a reminder, we train separate neural networks on data from 121-day windows centered on each calendar week of the year. The accuracy of the neural network for the window centered on January 10th is presented from both a deterministic and probabilistic perspective in Figure 3.4. The deterministic accuracy is assessed by counting the number of input samples the neural network assigns to the correct phase of the MJO. The most common error of the neural network is to assign an input sample to a phase that is one phase prior to or after the correct phase, which is likely caused by the MJO being a continuous phenomenon that we have discretized for the sake of interpretation. So, another useful accuracy metric is how often the neural network correctly assigns the input samples into either the correct phase or one phase before or after the correct phase. For the neural network centered on January 10th, the deterministic accuracy without a one-phase buffer is 74% and the accuracy with a one-phase buffer is 92%. From the composite probabilistic perspective (Figure 3.4b), the neural network assigns the highest likelihoods to the correct phase, although the phases immediately before and after the correct phase also have appreciably high likelihoods.

An important question regarding the usage of neural networks is whether they out-perform conventional methods, such as regression. If regression performs similarly to a neural network, then the increased complexity and nonlinearity of a neural network is not required. We therefore similarly use linear regression to identify the phase of the MJO using the input state variables and outgoing longwave radiation across 121-day windows centered on each calendar week. The linear regression models have no hidden nodes and no nonlinearities, but are otherwise identical to the neural networks in that the regression model assigns a normalized likelihood that the input is associated with a particular MJO phase by using a softmax operator before the final output. We regularize the regression models using L2-regularization to ensure they are not overfit to the training data, similar to the neural networks. The accuracies of the neural network and linear regression approaches are compared in Figure 3.5. The neural networks are nearly twice as accurate as lin-

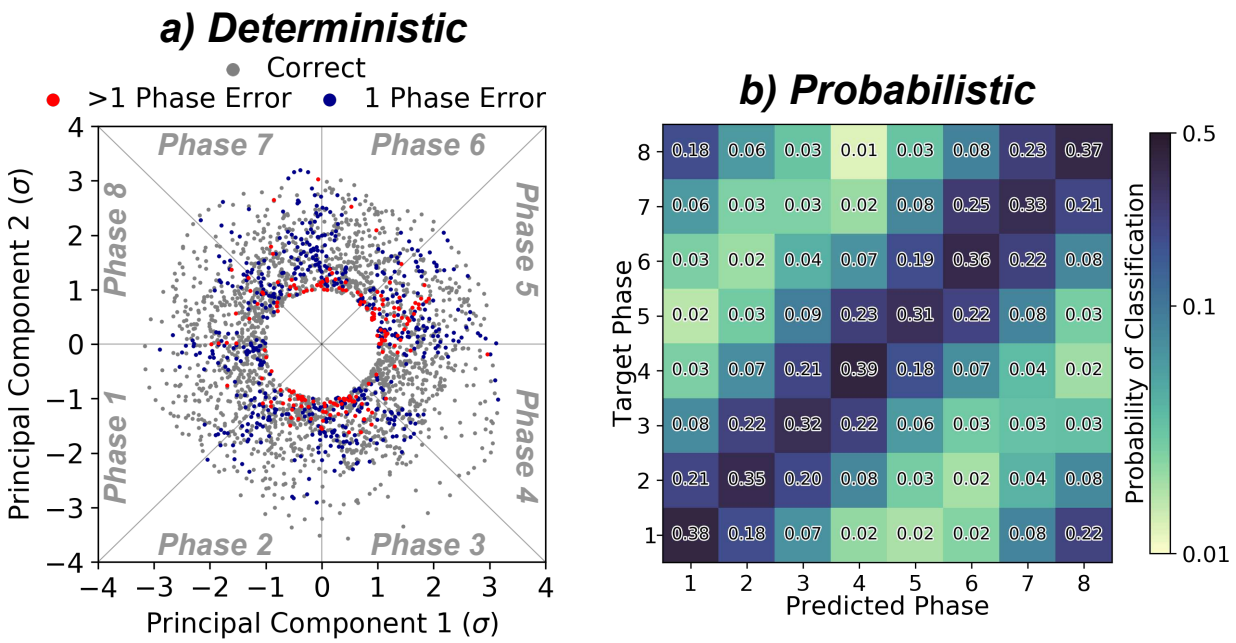

Figure 3.4: Example visualizations of the accuracy of the neural networks, in this case for the neural network centered on January 10. (a) Deterministic accuracy, where samples that are correctly classified are colored grey, those assigned to a phase one before or after the true phaes are colored blue, and those assigned to a phase two or more different from the true phase are colored red. (b) Probabilistic accuracy, where the average probabilities assigned to each sample within the validation dataset is shown for each target phase. The probabilities summed across each row sum to one.

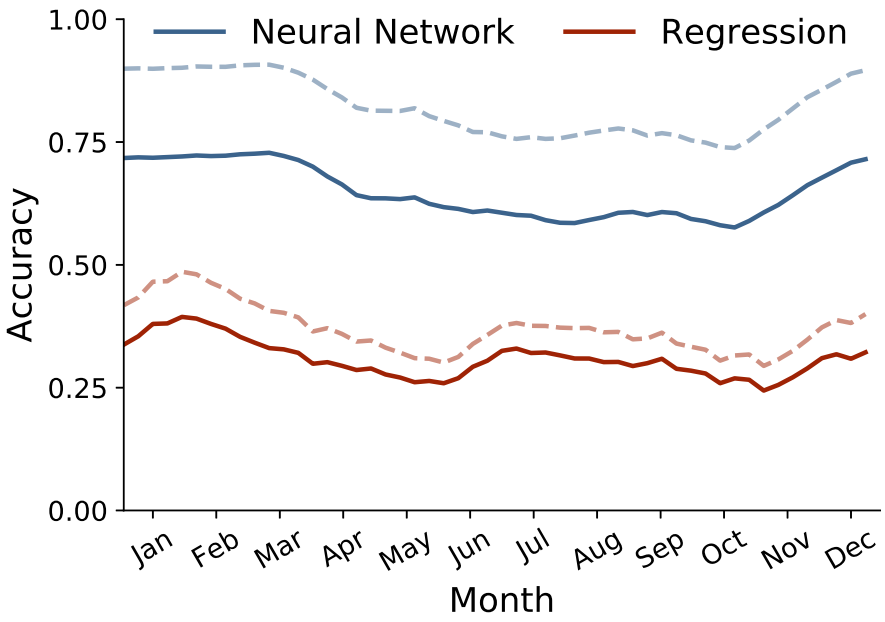

Figure 3.5: The accuracy of the neural network and linear regression approaches for each calendar week throughout the year. The neural network accuracy is plotted in blue, and the regression accuracy is plotted in red. The solid lines show the accuracy for all input samples, and the dashed lines show the accuracy if a one-phase error is permitted.

ear regression for all weeks of the year, which means the nonlinearities and increased number of pathways for information to flow through the neural network are essential to modeling the spatial structures of the MJO. We therefore conclude that interpretations of the neural networks can offer insights into relationships between the MJO and atmospheric state variables that conventional linear methods can not.

3.3.2 Interpreting the Neural Network

Identifying the Spatial Structures of the MJO

We use backward optimization and layerwise relevance propagation (LRP) to infer the spatial structure of the MJO and its seasonality according to the neural networks. Examples of LRP applied to inputs for the neural network trained on the 121-day window centered on January 10th are shown in Figure 3.6. We use four examples of MJO phase 7 for which the neural network correctly identifies the phase of the MJO, and for simplicity we only show the LRP maps for out-

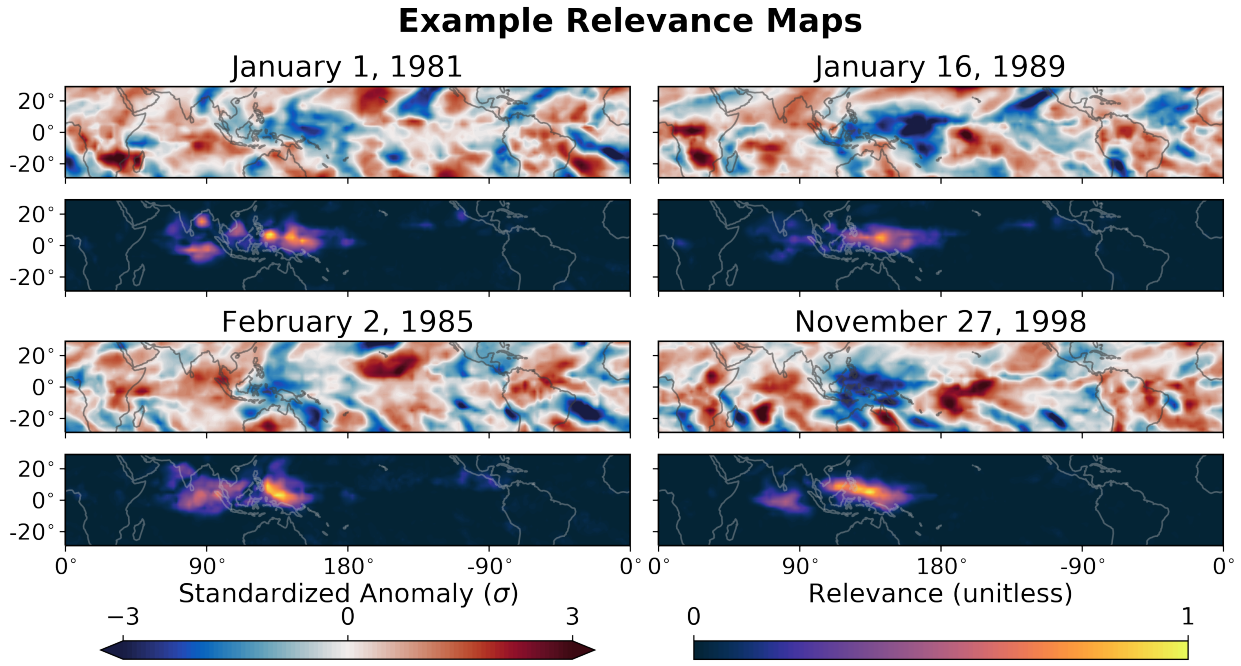

Figure 3.6: Example relevance heatmaps from the layerwise relevance propagation interpretation technique. The outgoing longwave radiation field from four example inputs into the neural network are shown, each corresponding to a separate Phase 7 MJO day. The corresponding relevance heatmaps are shown below each example outgoing longwave radiation field, and shows where the neural network focuses its attention to determine that the examples are associated with a Phase 7 MJO day.

going longwave radiation although similar maps are generated for each input variable. The LRP maps show that the neural network focuses its attention on outgoing longwave radiation anomalies across the Maritime Continent, particularly within its eastern extent, which is consistent with previous research on the regions of convection associated with phase 7 of the MJO [29, 97]. The LRP heatmaps also highlight the spatial uniqueness of each phase 7 MJO event, which can not be inferred by a linear regression model. It is likely that the increased accuracy of the neural networks compared to the linear regression models is caused by this ability of the neural network to capture the spatial uniqueness of each event.

We next test the neural networks more rigorously, and challenge them to identify the most common spatial structures of the MJO across its eight phases. To do so, we use backward optimization, and optimize inputs such that the spatial patterns within the inputs make the neural networks most confident that the inputs are associated with a particular phase of the MJO. Numerically, this means

that the outputs associated with the optimized inputs have a likelihood of approximately 1 in the phase for which they are optimized, and likelihoods of 0 for all other phases. We again only show the optimized outgoing longwave radiation fields for simplicity, although the optimization also identifies the characteristic patterns in the 15 other state variables.

The spatial pattern of the MJO during boreal winter (January 10th) and boreal summer (August 1st) according to both OMI and the neural networks are shown in Figure 3.7. The neural networks capture similar features to OMI during both seasons, in particular the prominent eastward propagation during boreal winter and the transition to northeastward propagation during boreal summer. The neural network focuses on a core of outgoing longwave radiation anomalies across the Indo-Pacific region and within the eastern Pacific, while OMI includes a greater magnitude of anomalies within the central Pacific. Given the similarities between OMI and the composite neural network interpretations, we conclude that nonlinearities of the MJO are primarily manifested through the uniqueness of each event as highlighted in Figure 3.6. The composites of LRP relevance similarly capture the dominant structures of the MJO during both seasons, and agree rather well with the optimized inputs (Figure 3.7).

Testing the Seasonality of the MJO

Because the neural network so accurately captures the seasonal evolution of the MJO within the outgoing longwave radiation composites, we now extend the interpretations to study the seasonality of the MJO. We first test how the spatial structure of the MJO changes across seasons using LRP. To do so, we calculate the composite relevance for each variable for each calendar week of the year, and present the annual evolution of the relevance in Figure 3.8. The relevances of each variable exhibit unique seasonal cycles, aside from outgoing longwave radiation which is similarly relevant throughout all periods of the year. For example, the seasonal cycle of lower-tropospheric zonal wind (U850) reaches a maximum in relevance during boreal summer, whereas upper-tropospheric zonal wind (U200) is most relevant during the spring and fall. Some variables exhibit a unimodal seasonal cycle (e.g. U850, T200), whereas other variables exhibit a bimodal seasonal cycle

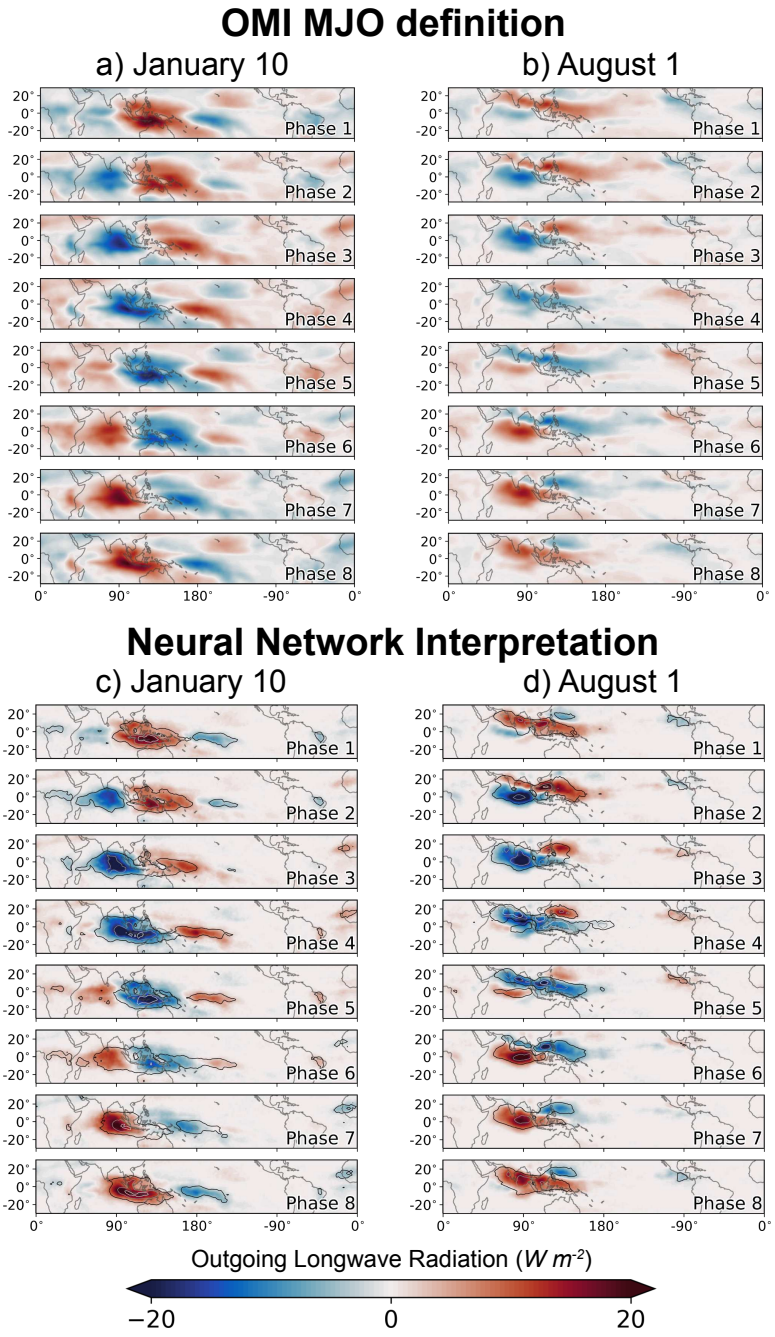

Figure 3.7: (a, b) The outgoing longwave radiation fields for each MJO phase according to the OMI for the boreal winter (January 10) and boreal summer (August 1) examples and those identified by the neural network. (c, d) The outgoing longwave radiation fields for each MJO phase according to the neural network based on the backward optimization and layerwise relevance propagation interpretation methods. The fill value shows the optimized outgoing longwave radiation patterns for each phase of the MJO, and the open contours show the composited relevance from LRP for all samples within each phase.

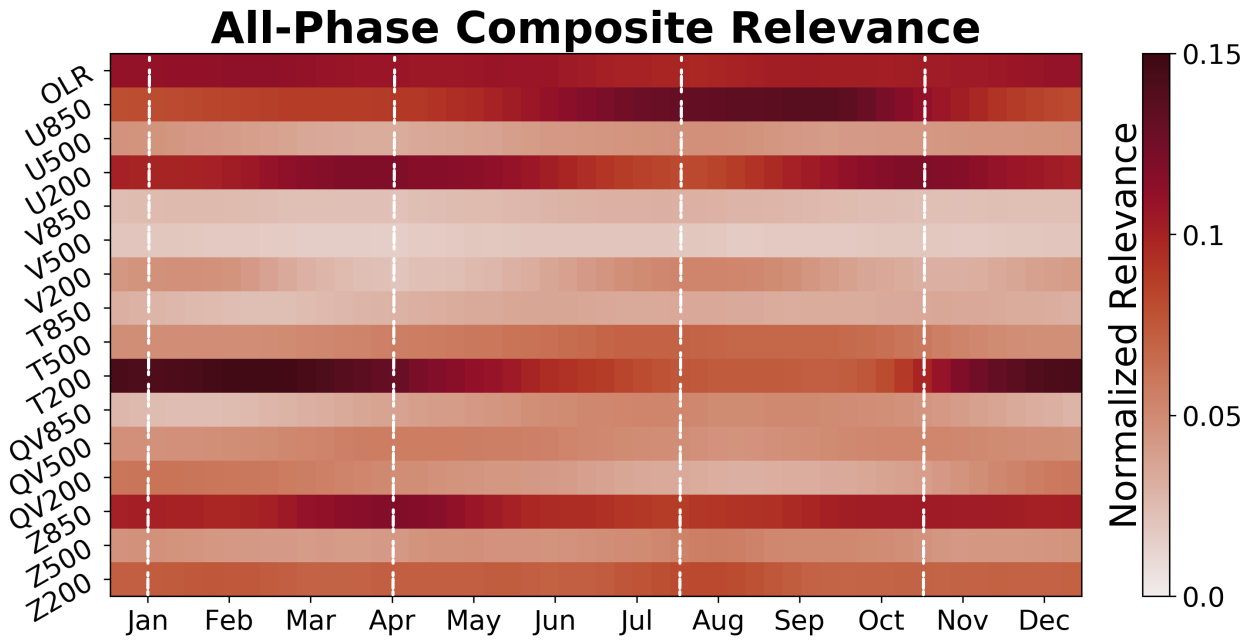

Figure 3.8: Composite normalized LRP relevance across all variables for each calendar week throughout the year. The relevance is normalized to sum to one across all variables for each calendar week (i.e. along the vertical axis).

(e.g. U200, V200, Z850) In general, upper-tropospheric anomalies are most relevant during boreal winter, while lower-tropospheric anomalies are most relevant during boreal summer.

The fact that upper-tropospheric anomalies are most relevant to the MJO during boreal winter may explain the seasonality in coupling between the MJO and the stratosphere [86, 103, 104]. Previous research has hypothesized that the MJO can be modulated by sources of stratospheric variability such as the quasi-biennial oscillation through a downward influence of upper-tropospheric temperature anomalies [105, 106]. So, because upper-tropospheric thermodynamic anomalies are particularly relevant to the MJO during boreal winter (Figure 3.8), then any influences on the thermodynamic structure of the upper troposphere by the stratosphere may have an increased impact on the MJO. This discussion highlights the capability of neural network interpretations to guide and test proposed hypotheses, although a direct test of this hypothesis is beyond the scope of this paper.

We now examine the optimal spatial patterns of the MJO throughout the year to provide some spatial context to the seasonality of the relevances shown in Figure 3.8. Figure 3.9 shows the opti-

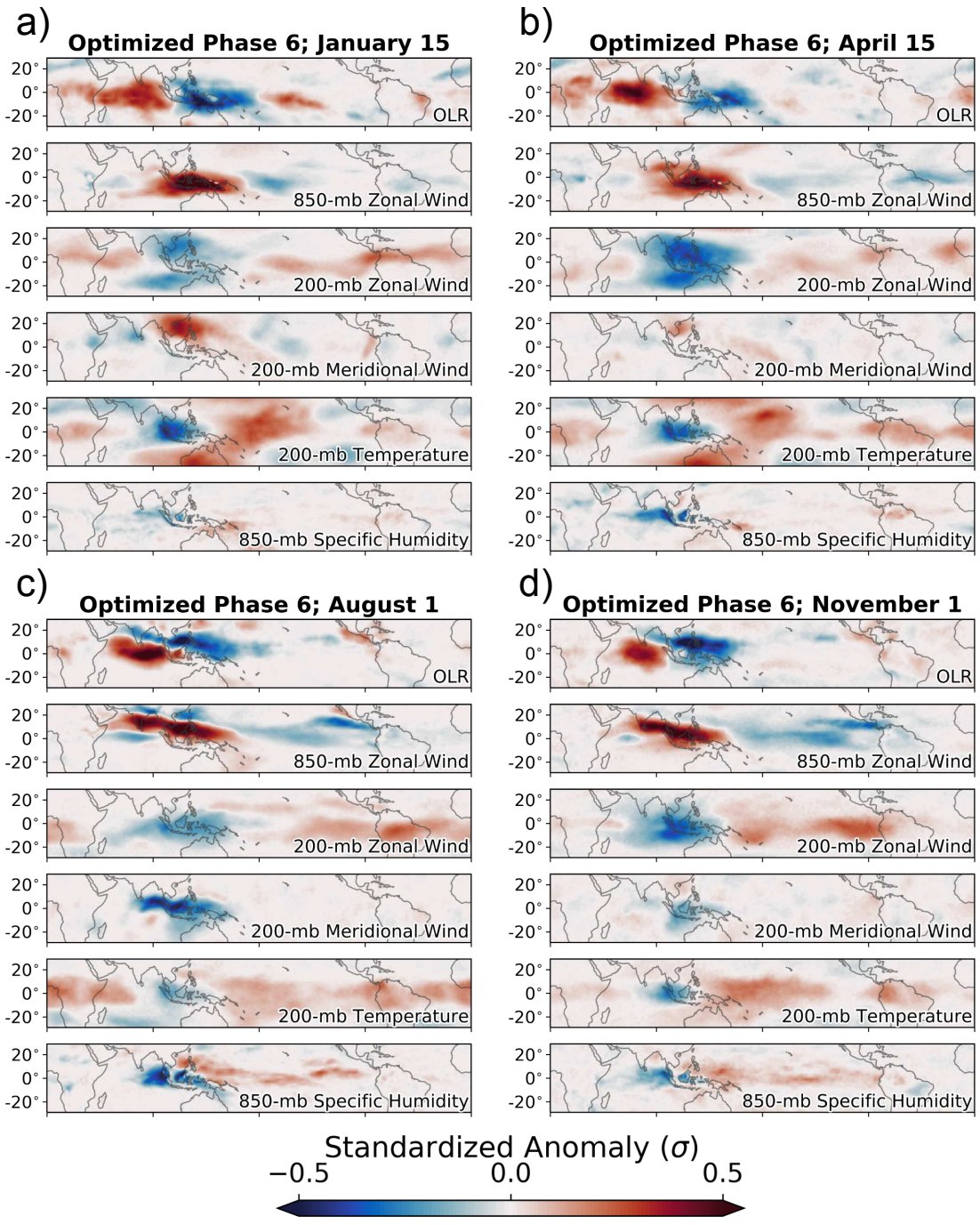

Figure 3.9: Optimized patterns for phase 6 of the MJO for different periods of the year. The central date on which the neural network is trained for each optimization is shown in the title of each subfigure. Each subfigure shows outgoing longwave radiation, 850-mb zonal wind, 200-mb zonal wind, 200-mb meridional wind, 200-mb temperature, and 850-mb specific humidity.

mal spatial patterns associated with phase 6 of the MJO at four different times of the year, the times of which are denoted by the dashed white lines in Figure 3.8. In general, the spring structure of the MJO is more similar to the winter structure than the summer structure. The April 15 optimal patterns are nearly identical to the January 15 optimal patterns, aside from lower-tropospheric moisture anomalies which are more similar between April 15 and August 1. The upper-tropospheric anomalies during boreal winter are more representative of a Matsuno-Gill type response to convective heating [107], whereas during boreal summer the signature is more diffuse and elongated across the equator. Figure 3.9 is generally supportive of the idea that lower-tropospheric anomalies are most relevant during boreal summer whereas upper-tropospheric anomalies are most relevant during boreal winter.

Mechanistic studies of the MJO commonly depend on accurate definitions of when each MJO seasonal mode occurs, since the spatial structures of the winter and summer modes differ so substantially (Figure 3.9). Should the seasonal definitions of the MJO be inaccurate, then there is a risk that the mechanistic studies themselves are not targeting processes specific to each season. We therefore use the backward optimization interpretations of the neural networks to define the MJO seasonal modes. To do so, we spatially correlate the optimal patterns for each state variable to the optimal patterns on January 10th and August 1st, which are generally considered to be the peak of the boreal winter and boreal summer modes. We then define the boreal winter mode to exist during periods for which the optimized MJO patterns have a correlation of greater than 0.75 with the optimized pattern for January 10th, and similarly define the boreal summer mode to exist when the correlation is greater than 0.75 with the optimized pattern for August 1st. Using our definition, the seasonality differs across atmospheric state variables, although the boreal winter and summer modes generally span from late November through early March and early June through early October, respectively (dark colors in Figure 3.10). Lower-tropospheric variables generally lead the transition from the boreal winter mode to the equinoctial transition toward the boreal summer mode, although a less clear relationship exists during the transition back to the boreal winter mode.

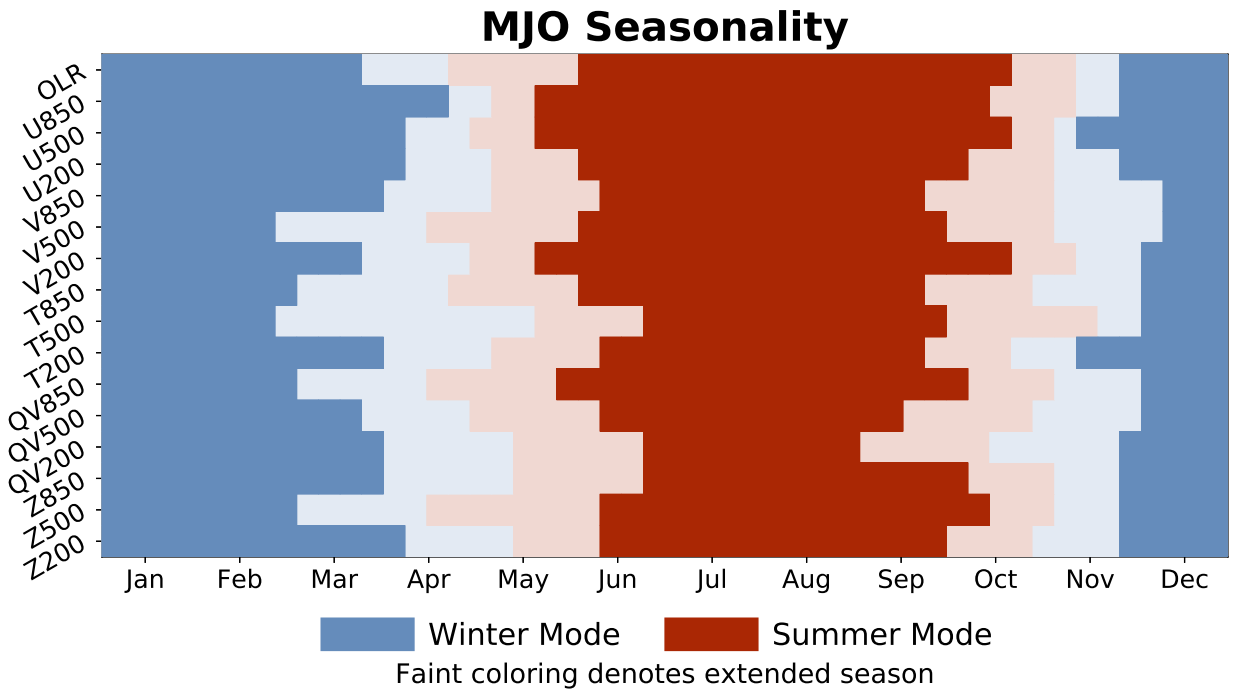

Figure 3.10: Seasonality of the Madden-Julian Oscillation according to interpretations of the neural networks. The extended boreal summer and winter modes are shown in red and blue, respectively, and periods of transition are denoted by the lighter red and blue colors. The winter (summer) mode is defined as periods during which the correlation between the optimized MJO pattern on January 10th (August 1st) and the optimized pattern for each respective calendar week is greater than 0.75. and the transition periods extend between these two modes. The extended boreal winter mode is defined as periods during which the optimized pattern for each respective calendar week is more highly correlated with the January 10th optimized pattern than the August 1st optimized pattern, and visa versa for the extended boreal summer mode.

Finally, we define extended boreal winter as the period during which the correlation between each weekly optimal pattern and the January 10th optimal pattern is greater than that between the weekly optimal patterns and the August 1st optimal pattern. Extended boreal summer spans the rest of the year. Using this definition, extended boreal winter MJO extends from early November through late April across most state variables, and from mid-November through late April for outgoing longwave radiation in particular (dark and light colors in Figure 3.10). Many studies of the MJO have previously used extended winter and summer seasons which span the months of October through March and April through September, respectively [108, 109]. Our results suggest that these extended seasons should, at a minimum, be shifted one month later in the year.

3.4 Discussion and Conclusions

We have tested the ability of interpretable neural networks to identify complex, multi-scale geophysical phenomena via their application to the Madden-Julian Oscillation (MJO). We first evaluated whether neural networks can identify the MJO, and then used neural network interpretability methods to study the seasonality and spatial structure of the MJO and its relationship to atmospheric state variables. Our study therefore contributes both to the general usage of neural networks within geoscience and to knowledge of the MJO itself, so we separate our discussion of the implications for both communities below.

3.4.1 Implications for Neural Networks in Earth Science

We have shown that neural networks are highly interpretable, even for complex, multi-scale geophysical phenomena. Two methods proposed by the computer science community – backward optimization and layerwise relevance propagation – provide particularly useful interpretations of neural networks [78]. Namely, backward optimization offers composite interpretations, while layerwise relevance propagation enables interpretations on either a composite or case-by-case basis. Both methods project the decision-making process of a neural network back onto the original di-

mensions of the input, which is particularly useful for geoscientific applications where each input variable may have unique physical importance to the problem being studied.

The capability of neural networks to include nonlinearities and simultaneously model different input patterns that lead to similar outputs proved useful for studying the seasonality of the MJO. The neural networks identified the phase of the MJO twice as accurately as linear regression, which implies that interpretations of the neural network characterize the MJO more accurately than linear regression. We hypothesized that the increase in accuracy was caused by the neural networks' ability to model the uniqueness of each MJO event, which is not feasible using conventional linear approaches such as regression. The amount of neural network complexity required for tasks across the geosciences will vary greatly, so the benefits of interpretable neural networks are also likely to vary across sub-disciplines. We have found that a baseline approach of comparing the accuracy of neural networks to more simple methods such as linear regression is useful in determining the necessity of a neural network.

Based on this study and other supporting work [78], the interpretations of what a neural network learns can be used to advance geoscientific knowledge. Even for cases where interpretability is not the main objective, neural network interpretations can offer insights into how and why neural networks are making their decisions, and can be used to ensure that neural networks are making decisions using reasoning consistent with physics. While we use a relatively simple type of neural network, the proposed methods are applicable to other types of neural networks as well, such as convolutional neural networks and long short term memory (LSTM) networks. We found fully connected networks to be particularly useful for our application and more accurate than convolutional neural networks, which, in light of the surging popularity of convolutional neural networks within geoscience, suggests that fully-connected networks also have utility for geospatial problems.

3.4.2 Implications for the Madden-Julian Oscillation

We also used neural networks as an approach to better understand the spatial structure and seasonality of the MJO. Our results are generally consistent with the thorough body of literature

on the MJO, which supports the reliability and robustness of interpretable neural networks within geoscience.

Consistent with previous studies, we find that the spatial structure of the MJO generally exhibits two dominant modes of variability distinguished between the boreal summer and winter. We find that the extended boreal winter mode of the MJO occurs between early November and late April, with the boreal summer mode occurring throughout the remainder of the year. This definition of the extended seasons is delayed one month compared to conventional definitions, which use an extended boreal winter of October through March. Furthermore, the seasonality of the relationship between the MJO and atmospheric state variables is more complex, with each variable exhibiting a unique seasonality. Some state variables such as lower-tropospheric zonal winds exhibit a uni-modal seasonality, whereas others such as upper-tropospheric zonal winds exhibit a bi-modal seasonality. We also find that upper-tropospheric thermodynamic anomalies are particularly relevant to the MJO during boreal winter, which may relate to the enhanced coupling between the MJO and stratospheric processes during this season.

Consistent with previous studies, we find that the spatial structure of the MJO generally exhibits two dominant modes of variability distinguished between the boreal summer and winter. We also extend our analysis to test numerous aspects of the MJO, from its nonlinearities to its relationships with atmospheric state variables. The key points of this analysis are as follows:

1. The neural networks identify the phase of the MJO twice as accurately as linear regression, which suggests that nonlinearities are important to the structure of the MJO. These nonlinearities are reflected in the spatial uniqueness of each MJO event, given that the composite structure of the MJO identified by the neural networks and linear methods are remarkably similar (Figure 3.5; Figure 3.6).

2. Each state variable exhibits a unique seasonality in its relationship with the MJO. For example, some state variables such as lower-tropospheric zonal winds exhibit a uni-modal seasonality, whereas others such as upper-tropospheric zonal winds exhibit a bi-modal seasonality (Figure 3.8; Figure 3.9).

3. Upper-tropospheric thermodynamic anomalies are particularly relevant to the MJO during boreal winter, which may relate to the enhanced coupling between the MJO and stratospheric processes during this season (Figure 3.8).

4. We find that the extended boreal winter mode occurs between early November and late April, while the boreal summer mode occurs throughout the remainder of the year. This definition of the extended seasons is delayed one month compared to the conventional definition, which uses an extended boreal winter of October through March (Figure 3.10).

Our results show that neural networks are highly interpretable, even for spatially complex geoscientific applications. Because of the high reliability of the interpretations, neural networks are viable tools for testing hypotheses related to the MJO and other spatially complex geophysical phenomena. More complex hypotheses can now be tested: for example, does horizontal advection of the lower-tropospheric mean moisture by the MJO circulation govern the propagation of the MJO [96]? Or, a neural network can be used to identify whether an MJO event will initiate given spatial inputs of atmospheric variables, from which interpretability methods can identify the most relevant patterns for MJO initiation. A critical requirement for using neural networks in such studies is the proven ability to reliably interpret what the networks have learned, which is now possible.

Chapter 4

Using Neural Networks to Identify Predictable Modes of Earth-System Variability

4.1 Introduction

Interpretable neural networks have opened new doorways in Earth science research [78], with applications ranging from the identification of climate change indicators [33], hail detection within severe thunderstorms [40], and the improvement of numerical model parameterizations [110], among other applications [41]. The specific usage of neural network interpretation techniques ranges substantially across such studies, however, as the interpretations can be used as either direct or indirect tools for scientific discovery. For example, interpretation efforts can be either a secondary objective by ensuring a network's reasoning is consistent with existing physical theory (e.g. Brenowitz et al., 2020; Ebert-Uphoff and Hilburn, 2020; Toms et al., 2020), or the primary objective, with their usage focused on discovering new patterns of Earth system variability (e.g. Barnes et al., 2020; Toms et al., 2020). Here, we focus on the latter application, whereby we use neural networks to identify predictable modes of Earth system variability on decadal timescales in a fully coupled Earth system model.

An extensive body of literature exists on theoretical and observed sources of decadal predictability, and, more recently, on the development of operational decadal prediction systems. Modes of regional and global-scale decadal variability within the ocean are well documented (e.g. Barnett et al., 1999; Kirtman and Schopf, 1998; Xie and Tanimoto, 1998), and these patterns have been found to contribute to atmospheric anomalies on decadal timescales via ocean-atmosphere feedbacks (e.g. Newman et al., 2016; Schneider et al., 2002; Wen et al., 2016). The discovery of this coupling has led to the usage of oceanic variability to make decadal predictions of atmospheric anomalies relevant to society. Recently, oceanic observations have been assimilated

into Earth system models to generate large ensembles of global decadal predictions [117–119], which have a reasonable amount of prediction skill for variables such as continental temperature and precipitation [120] and ocean acidification [121]. Additional efforts have created statistical decadal prediction models based on knowledge of specific modes of oceanic decadal variability (e.g. Simpson et al., 2019).

There are, however, limitations to decadal predictions that use dynamical Earth system models, including how to initialize the observational fields [123, 124] and long-standing model biases in simulating known ocean-atmosphere and land-atmosphere interactions [122, 125, 126]. It is therefore not clear whether regions that lack predictability in decadal prediction ensembles have limited predictability in the observed world, or whether model limitations preclude accurate predictions. This uncertainty also exists for other timescales of Earth system prediction, such as subseasonal-to-seasonal timescales, for which it is possible that numerical models have not yet realized their maximum predictive skill due to inaccurate simulations of the necessary modes of atmospheric and oceanic variability [86, 127–130]. For statistical models, a complete knowledge of which modes of oceanic variability offer predictability is important for the correct selection of model inputs and thereby a maximization of statistical prediction skill (e.g. DelSole and Banerjee, 2017; Simpson et al., 2019; Wilks, 2008).

Because of these uncertainties, it is useful to identify predictable modes of Earth system variability within both models and observations. Knowledge of such modes may, for example, help guide efforts to improve the robustness of observational assimilation within dynamical decadal prediction systems, or inform which variables and regions to include within statistical models. To this end, we use a new method, namely interpretable neural networks, to identify sources of decadal predictability within a fully coupled Earth system model. We take a purely methodological approach and test whether the proposed method is viable for identifying such modes of predictability, which opens opportunities for its application to a broader range of predictability problems in future studies.

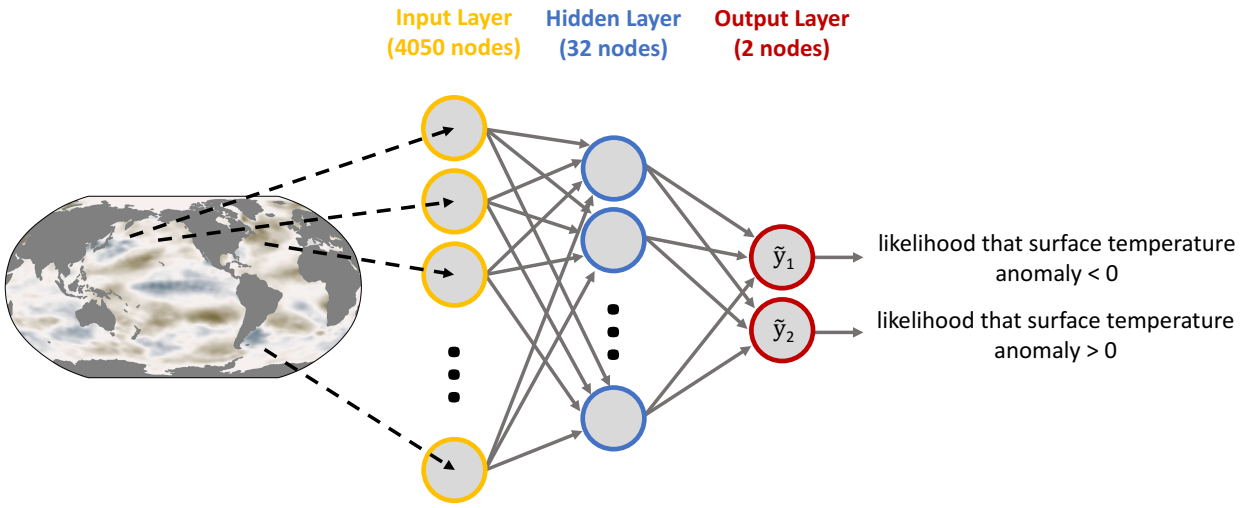

Figure 4.1: Schematic of the neural network design. The neural network receives a vectorized sea-surface temperature field as input, passes the input forward to a single hidden layer of 32 nodes, and finally outputs a likelihood that the input is associated with surface temperature anomalies of a particular sign for a specified location.

4.2 Data and Methods

Our neural network architecture is designed to receive inputs of oceanic fields from an Earth system model and output the predicted sign of a continental temperature anomaly at a given location. Figure 4.1 describes this neural network design, and the appendices contain additional information about the training procedure. It is important to note that we have opted to keep the neural network as simple as possible to both maximize interpretability and to ensure our approach is valid before venturing into more complex networks in future studies. The neural network has one hidden layer of 32 nodes which is connected to two output nodes, both of which represent a different outcome associated with the input oceanic field. We use the rectified linear unit (ReLu; $max(0, x)$) activation function and apply a softmax operator to the output layer. The softmax operator transforms the neural network outputs into relative likelihoods of the two output climate states.

For our particular application, we input vectorized maps of global sea-surface temperature (SST) and the neural network is trained to output the associated likelihood that future continental surface temperatures across locations of North America will be anomalously warm or cold.

The SST and continental surface temperature data are gathered from the Community Earth System Model Version 2 (CESM2; Danabasoglu et al., 2020) pre-industrial control simulation of the Coupled Model Intercomparison Project, Phase 6 (CMIP6; Eyring et al., 2016). We remove the seasonal cycle from both fields and re-grid the SST field onto a 4° by 4° grid to reduce the number of inputs into the neural network. This grid spacing still permits the resolution of dominant patterns of oceanic variability, as we will show in Section 4.3. We also apply a 24-month running average to the SST anomalies and a 60-month running average to the continental surface temperature anomalies, such that for any time the corresponding SST field represents the precedent 24-month mean and the continental surface temperature represents the future 60-month mean. We use these input and output smoothing durations to demonstrate the utility of the proposed methodology, and they can be changed for particular timescales or seasons of interest. The CMIP6 CESM2 pre-industrial control simulation offers 1,400 years of monthly data, and so we train the neural networks on all 16,800 samples aside from the beginning and end of the time-series which are contaminated by the temporal smoothing. We note that because we train the neural networks using a pre-industrial control simulation, all estimates of predictability provided by the neural networks are for internal variability only and do not include information about any predictable response due to anthropogenic forcing.

After training the neural network, we use an interpretation method called layerwise relevance propagation (LRP; Montavon et al., 2018) to assess what the network has learned. In brief, LRP traces the decision-making process of a neural network for each individual input sample. For each input sample, the network pathways through which information flows to arrive at the associated output is traced backwards and projected back onto the dimensions of the input. This projection enables an interpretation of which inputs are most important for making predictions on a case-by-case basis. Our usage of LRP therefore offers insights into which patterns of SST variability lend predictability of decadal surface temperature anomalies over continental North America. A more detailed discussion of LRP and its applicability to Earth system research is discussed in Toms et al.

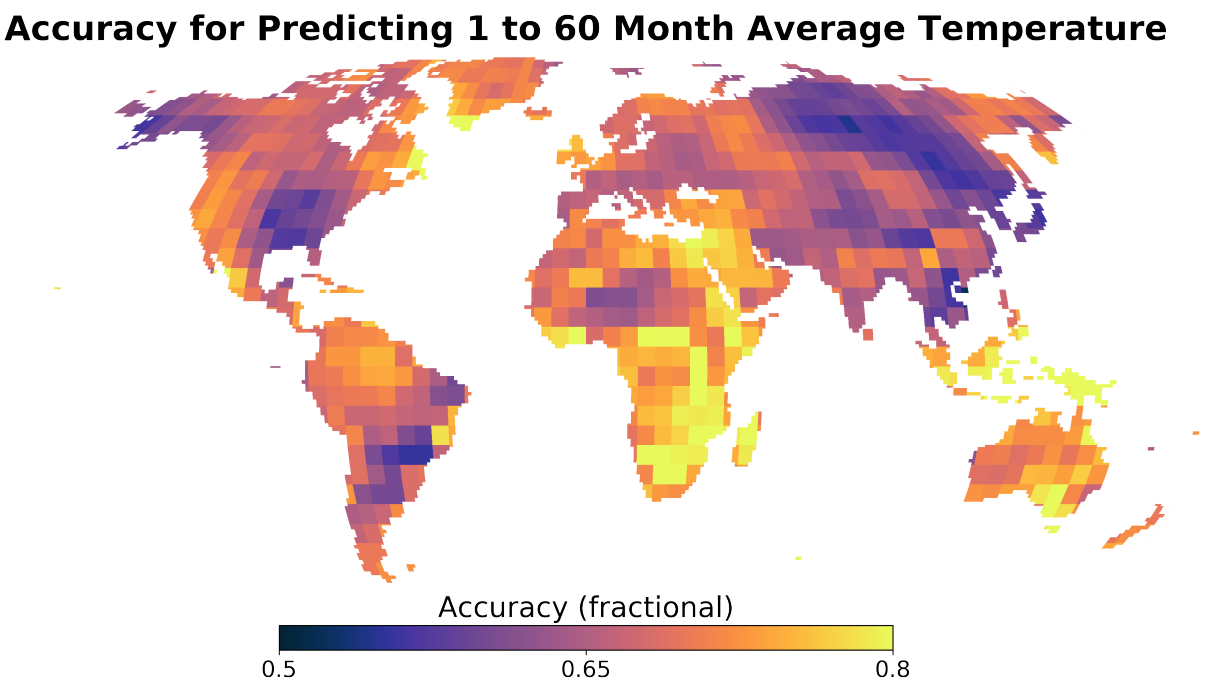

Figure 4.2: Accuracy for the neural network approach. The accuracy is defined in a Boolean sense, and the output node with the highest likelihood is taken as the networks' prediction. The accuracy values therefore represent the fraction of predictions for which the neural networks predict the correct sign of continental surface temperature anomalies.

(2020), and additional applications are available in Barnes et al. (2020), Ebert-Uphoff and Hilburn (2020), and Toms et al. (2020).

4.3 Assessment of Decadal Predictability

We train a separate neural network for each location on a 5° by 5° grid across the globe. We choose this resolution due to the computational expense of training a neural network for every location across the globe. Each neural network can then identify patterns of SST that lend predictability unique to each location, which is helpful for understanding if the predictability across different regions of the globe is sourced from different oceanic modes. Figure 4.2 shows the resultant accuracy for each of these neural networks in predicting the 1-to-60 month average surface temperature using a global map of the prior 24-month mean SST within the CESM2 pre-industrial control simulation. The accuracy varies across the globe, with southern Africa, southern Australia, the Maritime Continent, and parts of northeastern North America exhibiting the highest accuracy.

We then use LRP to assess which modes of oceanic variability contribute to the predictability within the CESM2 pre-industrial control simulation, as shown in Figure 4.2. The following analysis is applicable to any region of the globe, although we choose North America as an example. We only assess the LRP interpretations for cases when the neural networks make accurate predictions. We further separate the interpretations into accurate predictions of positive and negative temperature anomalies and only show the results for the positive anomalies, although the analysis for the negative anomalies is similar (see supplementary information). The composite LRP patterns for four regions across North America suggest that predictability is sourced from different oceanic patterns for different regions (Figure 4.3). Perhaps surprisingly, continental temperature anomalies within Central America are most associated with SST anomalies off the east coast of Japan (Figure 4.3a), likely within the Kuroshio Extension [135]. SST anomalies within the North-Central Pacific Ocean are associated with continental temperature anomalies along the west coast (Figure 4.3b), while those within the tropical Pacific Ocean contribute to predictability across central North America (Figure 4.3c). The North Atlantic Ocean contributes predictability to the four locations, although its impacts are particularly prominent across the northeast portions of the continent (Figure 4.3d).

These regions of oceanic anomalies that lend predictability within CESM2 are spatially similar to known modes of decadal oceanic variability. The SST linkages within the tropical Pacific may be associated with the El Niño-Southern Oscillation [111, 136, 137], while those in the North Pacific may be associated with the Pacific Decadal Oscillation [115, 138], and anomalies within the North Atlantic may be associated with the Atlantic Meridional Overturning Circulation [139, 140]. A mechanistic study would need to be conducted to confirm that these modes of variability are indeed the origins of the predictability, although it is highly likely that they are the primary contributors given their strong similarities in spatial structure to the patterns identified by the neural networks.

A unique aspect of using LRP to identify modes of predictability is that LRP highlights which input patterns contribute to predictability on a case-by-case basis. So, we further analyze which patterns of oceanic variability lend continental temperature predictability by using k-means clus-

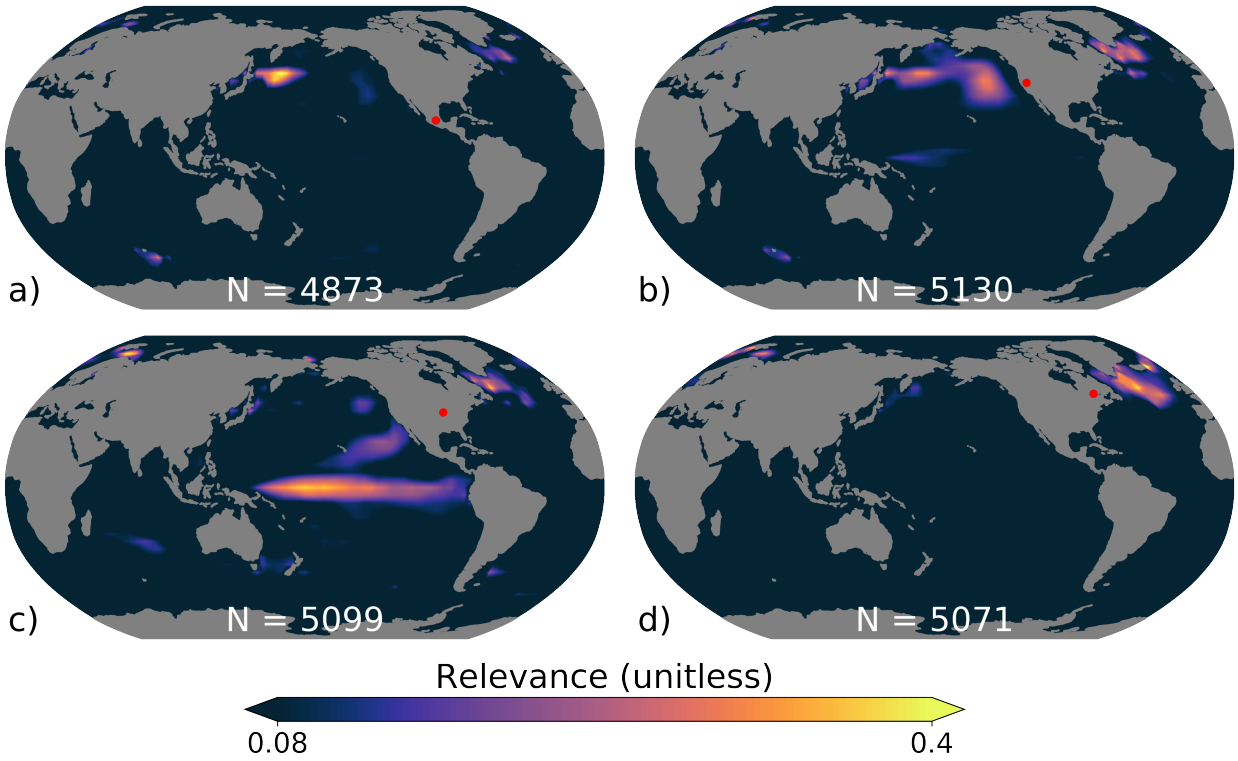

Figure 4.3: Composite layerwise relevance propagation interpretations for accurate predictions of positive surface temperature anomalies at four locations across North America. The continental locations associated with the composites are denoted by the red dots in each panel. The LRP interpretation for each sample is normalized between a value of 0 and 1 before compositing to ensure each prediction carries the same weight in the composite. Relevance values below the 95th percentile confidence bounds (0.08) are not shown. Confidence bounds were determined using a null hypothesis of no predictability by randomly shuffling the order of the input sea-surface temperature maps, and calculating the 95th percentile values of the associated LRP composites.

tering. The composite interpretation in Figure 4.3 risks averaging together temporally distinct modes of predictability, and so the clustering approach allows us to analyze these potentially distinct modes separately. We focus in particular on the west coast of North America in a region that exhibits high continental surface temperature predictability (according to Figure 4.1). We determine the optimal number of clusters by plotting the number of clusters against the mean Euclidian distance between each cluster, and selecting the number of clusters which falls in the inflection point of this curve (not shown). The inflection point denotes the number of clusters after which the addition of new clusters offers substantially less new information than the previous clusters. This technique is colloquially called the “elbow" technique (e.g. Dimitriadou et al., 2002).

Using this approach, we find three dominant modes of oceanic variability within CESM2 that lend predictability at the chosen location along the west coast of North America (Figure 4.4). The three modes are spatially similar to separate, known modes of oceanic decadal variability. Namely, the most common mode is similar to the known SST structure of the the Atlantic Meridional Overturning Circulation [139, 142], while the second and third clusters are spatially similar to the Kuroshio Extension [135] and Pacific Decadal Oscillation [115], respectively (Figure 4.4a, b, c). This finding suggests that the predictability at the chosen location can be sourced independently from any of these three oceanic modes of variability. Furthermore, the clustering analysis is forced to identify the most spatially distinct modes of oceanic variability that lend predictability. So, it is likely that there are also situations in which the identified modes lend predictability in tandem.

Along with its prediction, the neural networks output likelihoods that the input SST field will lead to positive or negative continental temperature anomalies. We therefore use the likelihoods to assess the oceanic state for highly confident (i.e. high likelihood) accurate predictions, and compare those cases to accurate predictions with lower confidence. In doing so, we find that higher confidence predictions for the west coast of North America are made when SST anomalies are of greater magnitude within the northern Atlantic and Pacific oceans (Figure 4.5). Regions associated with the Pacific Decadal Oscillation and Atlantic Meridional Overturning Circulation

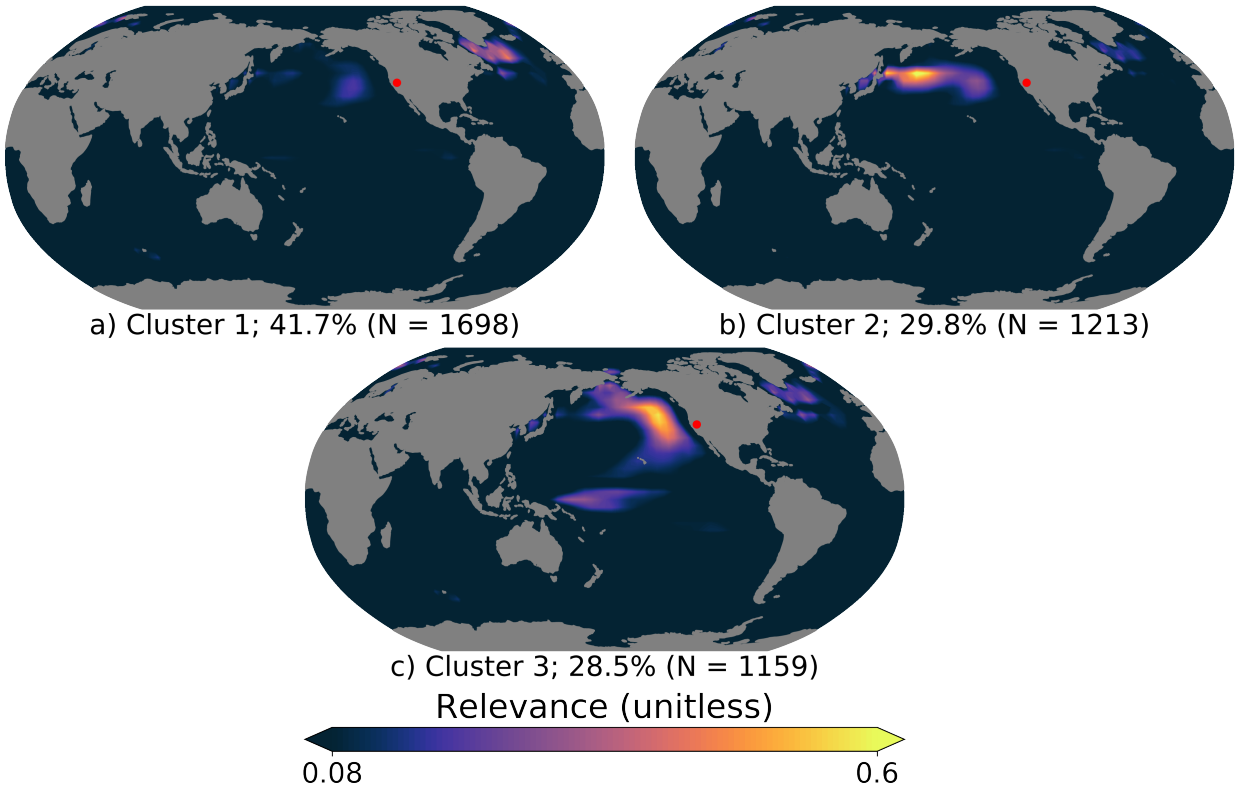

Figure 4.4: K-means clusters of the layerwise relevance propagation interpretations for accurate predictions of positive surface temperature anomalies at the red dot. The percentage of cases corresponding to each cluster is listed in the bottom left of each sub-panel. The LRP interpretation for each sample is normalized between a value of 0 and 1 before compositing to ensure each prediction carries the same weight in the composite. Relevance values below the 95th percentile confidence bounds (0.08) are not shown. Confidence bounds were determined using a null hypothesis of no predictability by randomly shuffling the order of the input sea-surface temperature maps, and calculating the 95th percentile values of the associated LRP composites.

are most magnified in the high confidence predictions. According to LRP, the SST anomalies within the North Pacific Ocean are particularly relevant for the high confidence scenarios.

4.4 Discussion

We demonstrate that neural networks can identify modes of oceanic variability that lend predictability on decadal timescales within Earth system models. In particular, the neural networks identify known patterns of decadal oceanic variability as sources of predictability for continental surface temperature anomalies across North America within the CMIP6 CESM2 pre-industrial control simulation. The identified modes of oceanic variability each offer distinct sources of pre-

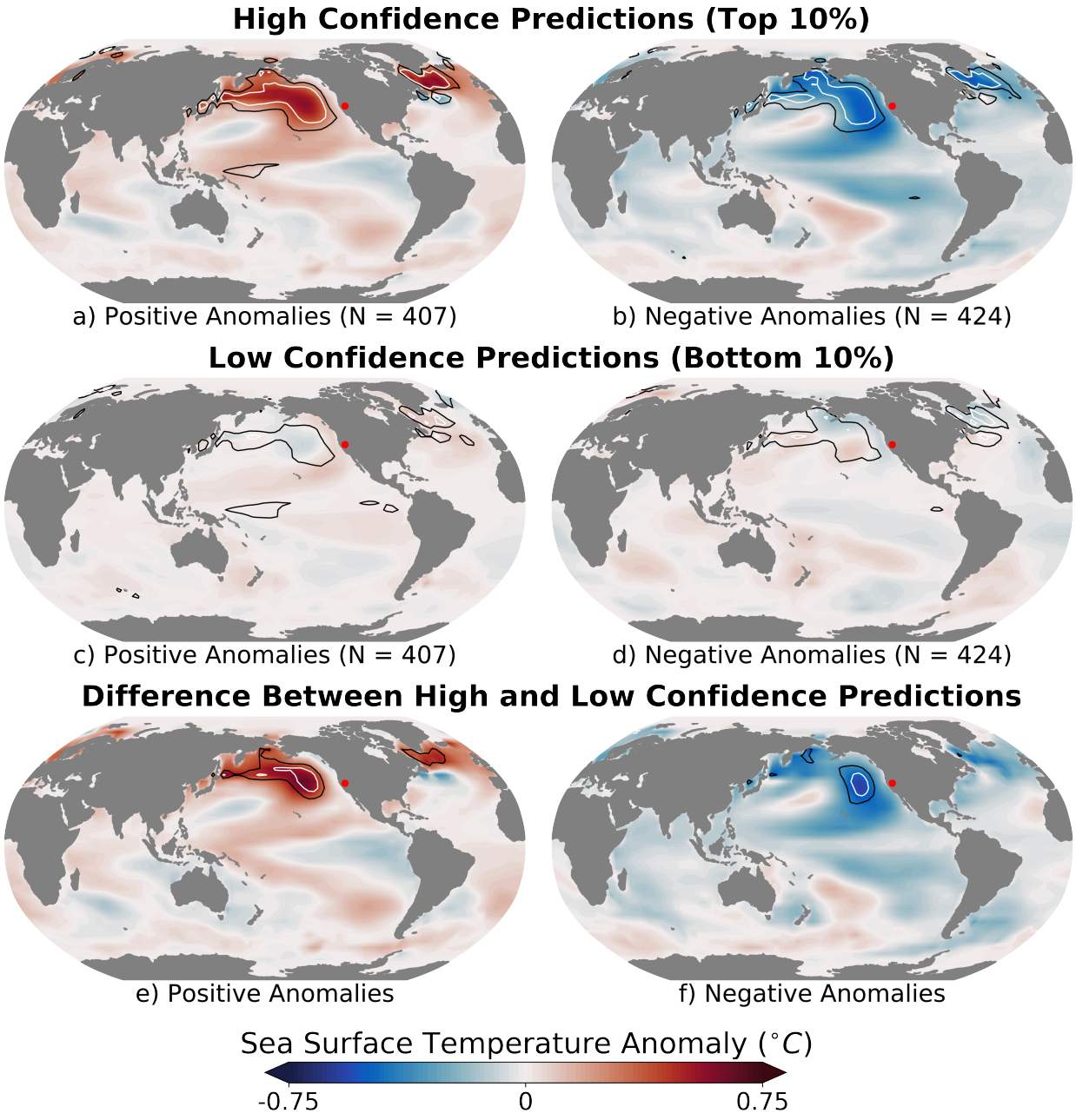

Figure 4.5: Differences in sea-surface temperature anomalies and LRP relevance for the top 10% and bottom 10% confident predictions for (a) positive surface temperature anomalies and (b) negative surface temperature anomalies at the red dot. The difference in sea-surface temperature anomalies is shown in fill, and the difference in LRP relevance is shown in open contours. The black (white) contour denotes an LRP difference of +0.1 (+0.2). Negative LRP relevance differences are also allowed to be shown, although none exist with magnitudes of -0.1 or greater.

dictability, at least across the west coast of North America where the useful oceanic regimes may be associated with the Atlantic Meridional Overturning Circulation, Pacific Decadal Oscillation, and Kuroshio Extension.

While we propose the methodology in this paper through its application to a single Earth system model (CESM2), the method can be applied to a collection of climate models to assess the similarities of predictable climate modes across different models. Additionally, the methods are likely viable for other timescales of predictability, such as the subseasonal-to-seasonal range. These timescales lie at the intersection of predictable processes in the atmosphere, land, and ocean [130, 143, 144], and interpretable neural networks may therefore be useful in determining coincident modes of predictability within each domain.

The complexity of the proposed method can be varied as necessary, although we introduce it here with intentional simplicity. For example, the neural networks can be made more nonlinear through the addition of more nodes and hidden layers, temporal information can be included within the inputs and outputs, and numerous Earth-system variables can be input rather than sea-surface temperature alone. The method may also be applicable to observational data, particularly cases for which an extensive observational record exists (e.g. subseasonal-to-seasonal prediction). Our formulation also only tasks the neural network with predicting positive or negative temperature anomalies without regard to magnitude, so the addition of more categories of output temperature anomalies can help separate anomalies of different magnitudes. From a broader perspective, this study contributes to the growing body of evidence that interpretable neural networks can be used to advance geoscientific knowledge.

Chapter 5

Summarizing Thoughts

Interpretable neural networks have the potential to drive a new wave of geoscientific discovery. Within this dissertation, I introduced the usage of neural network interpretations for such a purpose by first providing a framework for neural network-driven geoscientific discovery and then applying the framework to two unique aspects of the earth-system. The intention of these studies has been to provide practical, objective methods to advance our science. Below, I provide a brief overview of each chapter before my parting thoughts on the future of neural networks within the earth sciences.

The first study of my dissertation proposed a new framework for using neural network to discover patterns of earth-system variability. The framework is relatively simple and generalizable to numerous types of neural networks, which should enable its usage across a broad range of geoscientific sub-disciplines. We focused on using interpretation methods that project the decision-making process onto the original input dimensions, since the geoscientist that designs the neural network is likely familiar with the physical meaning of each input. In this sense, we propose a framework for physically interpetable neural networks, whereby the geoscientist is responsible for discovering the physical meaning expressed by the interpretations. We applied the proposed framework to two relatively simple problems to illustrate its utility (phase identification of ENSO and its contributions to seasonal predictability), with hopes that the community will embrace the methods and apply them to problems of greater scientific interest. We found that the proposed framework captures the expected physical patterns and provides reliable insights into how oceanic variability is connected to both the El Niño-Southern Oscillation and seasonal temperature variability across the west coast of North America.

The framework outlined in the first study is tested more rigorously in the second, wherein we task neural networks with describing the seasonality of a spatiotemporally complex tropical disturbance called the Madden-Julian Oscillation (MJO). In order to accurately describe the seasonal evolution of the MJO, the neural networks must filter through meaningless noise within the input

fields, differentiate between eight phases of the MJO's evolution, and accurately describe the relationship between the MJO and sixteen atmospheric state variables. A relatively simple neural network architecture with only two hidden layers is capable of such a feat, and the interpetability techniques prove that the neural networks do, indeed, capture the physical structures of the MJO. We then extended our analysis to study the nonlinearity of the MJO, and found that the it is indeed nonlinear, and expresses this nonlinearity through the uniqueness of each event. These findings are unique in that our proposed methodology enables an assessment of the nonlinearity of the MJO, and any physical interpretations of the relationship between the MJO and atmospheric state variables are therefore more complete than conventional linear analyses.

The final study extends the interpretable neural network framework into earth-system predictability, on timescales for which knowledge of predictability is relatively limited. Decadal prediction has been a topic of interest over the past decades, and a growing body of evidence suggests that the ocean provides a majority of the predictability on these timescales. We therefore used interpretable neural networks to identify modes of oceanic variability within a climate model that lend predictability for decadal-timescale continental surface temperature anomalies across North America. We found that the neural networks identified patterns of decadal oceanic variability previously known to the geoscientific community, including the Kuroshio Extension, the Pacific Decadal Oscillation, and the Atlantic Meridional Overturning Circulation. The proposed methodology is generalizable to any timescale of predictability, and so will likely contribute to the discovery of predictable patterns on other timescales, as well.

Perhaps the most important contribution of this dissertation is the newfound ability to clearly interpret how and why neural networks make their decisions within a geoscientific setting. Neural networks have historically been treated as “black boxes" within the earth sciences, whose interior workings can not be understood. With the assistance of methods originally invented by the computer science community, we have shown that thoughtfully constructed neural networks are highly interpretable within geoscientific applications. This newfound confidence will hopefully engage those doubtful of the utility of neural networks in geoscience and encourage those already pursuing

their usage to understand how and why their models make decisions. My research has been most interested in whether these interpretations can be used to discover geoscientific patterns, but they can also be used for numerous other applications. For example, my colleagues have found these methods useful in understanding why neural networks make decisions in order to properly optimize the network architectures [80]. There is also interest from a growing community of scientists using neural networks for numerical parameterizations to understand the mechanisms driving the parameterizations prior to their placement within climate models [110]. Regardless of the ultimate objective, the proposed framework has the potential to improve the applicability of neural networks within the geosciences.

Although this dissertation makes strides towards the usage of neural networks for discovering patterns of geoscientific variability, there are limitations that should be addressed. Each of these limitations are listed below, in no particular order.

First, the framework we developed required neural networks of a particular formulation. Namely, layerwise relevance propagation has thus far been formulated for simpler types of neural networks, discluding those with skip connections, batch normalization, or other similar more advanced techniques that help neural networks converge more efficiently. In informal discussions with the group who developed LRP, there are efforts to generalize LRP to these types of networks, however, and so there will likely be a time when the type of neural network used in our framework will be less limited. With that said, there are methods for neural network interpretation that can be applied to more complex networks. Saliency is one such example (e.g. Gagne et al., 2019; Mcgovern et al., 2019). We did not discuss the usage of saliency within our framework because LRP addresses scientific understanding more directly. LRP tells the user which inputs lead to an increased confidence in the neural network for a particular outcome, which can be inferred as the inputs that are important, or most relevant, for a neural network's decision. Saliency, on the other hand, tells the user what changes to the inputs would lead to a maximal change in a selected output. These answer very different scientific questions, and saliency requires more careful interpretation. So, there is a trade-off between a more generalizable interpretation method (saliency) and more directly ap-

plicable interpretations (LRP) – at least in the opinion of the writer of this dissertation. Saliency does have a useful place in geoscientific discover, as well, and so can be used similarly as LRP within the framework proposed in Chapter 2. Future studies can incorporate saliency to assess its usefulness.

A second limitation is the necessity of caution in claiming scientific discovery. It is possible for a scientist to fool themselves into thinking they have made a discovery by forcing physical meaning onto a pattern discovered by statistical or other automated pattern discovery methods (such as neural networks). The framework we propose is therefore not automated and, if anything, requires an even more careful assessment from the scientist. It may be easy for a scientist to be fooled by patterns identified by a neural network, since the process of pattern identification itself becomes automated and requires less procedural thought. If care is taken, this automation can enable a scientist to allocate their time and intellectual efforts towards interpreting patterns in a more thorough manner to ensure they do, indeed, have physical meaning, rather than spending this time on developing the methods used to identify the patterns in the first place. An example of how the proposed framework could be used to discover new patterns of earth-system variability in a comprehensive manner is as follows. First, the neural network interpretations can be used to identify meaningful patterns within a dataset. Then, the scientist can study these patterns and use any background knowledge available to them to assess whether the patterns may hold physical meaning. If the patterns do hold promise, the scientist can then impose the patterns within a model constrained by physics to test more targeted hypotheses related to the patterns. An example of such a physically constrained model may be a simplified dynamical model, wherein the components of the earth system are reduced to study simplified hypotheses. In this sense, the framework can be used to generate scientific hypotheses, which must then be tested with scientific rigor typical of research before the invention of the framework.

Finally, neural networks can be difficult to optimize, which means the framework depends on the ability of the user to carefully and correctly tune a neural network. Readers familiar with neural networks will know the multitude of parameters that can be tuned to ensure a neural net-

work is accurate – training speeds, the type of error function used to optimize the network, how many nonlinearities are included, and so on. A potential solution for this issue is to create teams of computational scientists and physical scientists that can work together to propose interesting physical problems and model them in appropriately tuned neural networks. Applications of neural networks to geoscientific discovery are also at an advantage over users of neural networks for prediction, in that prediction typically requires the networks be tuned for maximum accuracy. Geoscientific discovery may not depend on maximizing accuracy; rather, discovery depends on maximizing interpretability. A perfectly optimized solution is therefore not as important for a purpose of geoscientific discovery, so long as the network is not overfit to the training data or otherwise improperly tuned.

As with any scientific study, care must be taken when applying the proposed framework for geoscientific discovery or any other purpose. This idea is not new to science – scientists are trained to be skeptical, careful, and to question the details. We are confident that scientists will continue to use such care when advancing the earth sciences using neural networks.

Now that we have addressed potential limitations of the proposed framework, it is useful to acknowledge future work that may advance upon the work in this dissertation. The title of this dissertation uses the word “towards" intentionally – I anticipate future work will make the framework more useful, and perhaps another scientist will propose an even more useful framework. The framework can also be applied to a broader range of problems, or applied to the problems within this dissertation with more rigor.

A proposed avenue for future research from a methodological perspective is to generalize the framework from fully connected networks to more sophisticated convolutional networks. Convolutional networks may be particularly promising, in that a single network can be used to make predictions for not one location, but for every location across the globe. Such an application could make use of the popular U-Net neural network framework [145]. Then, this single neural network can be used to make interpretations for every location across the globe, and the patterns that lend predictability for each location will be intertwined throughout the weights and biases of a single

network. This extension of the framework would be particularly useful for the decadal predictability example. If a separate neural network is trained to make prediction for each location across the globe – as is done in this dissertation – then there is a possibility that the interpretations for neural networks trained for adjacent locations will differ drastically. We found this not to be the case for decadal timescales, and that the neural networks converged on similar minima, such that interpretations for adjacent locations were similar, if not identical. However, by using a single neural network to make predictions and interpretations for the entire globe, comparisons of patterns of predictability across different locations would be more reliable.

An extension of the proposed framework could be one that considers causality. A major limitation of the framework as it exists in this dissertation is its focus on correlated patterns, rather than causal patterns. If a framework can be developed that identifies causal patterns, the framework would further enhance a scientist's ability to generate exciting scientific hypotheses. I am optimistic that such a framework can be developed, and given the recent increased interest in the atmospheric science community towards methods of automated causality discovery, it is possible that such a framework will be developed soon after the publication of this dissertation.

5.1 The Future of Neural Networks in Geoscience

Neural networks have been used in geoscience for more than two decades [146], and it is likely that their usage will continue throughout at least the next two. The computer science community has recently experienced record numbers of abstracts submitted to top-tier artificial intelligence and neural network-focused conferences, and so the surge of knowledge originating from the computer science community will likely also continue for some time. There has also been increased interest in high-resolution climate simulations to ensure mesoscale processes are appropriately modeled, which will thereby continue to increase the quantity of high-quality climate data [147, 148]. A greater number of realizations of possible weather and climate patterns are also being observed as global observation networks evolve and time progresses. The so-called nexus of computing

capabilities, data availability, and algorithmic advances that led to the recent uptick in geoscientific interest in neural networks will therefore only continue to expand.

As algorithmic advances are made in the computer science community, the importance of ensuring neural networks are viable tools for geoscience will continue. While a notable scientific advancement, this dissertation merely scratches the surface of the immense number of possible geoscientific applications of neural networks, the applications of which will continue to grow in complexity as the geoscientific community becomes more comfortable with the method. Efforts to understand how and why neural networks make decisions will therefore be a persistent necessity. Of course, I am not the only one to recognize this importance. Numerous studies have been published on interpretation methods for neural networks and other machine learning methods in geoscience [32, 40], and the National Science Foundation recently invested $20 million into an AI institute focused on “Trustworthy AI in Weather, Climate, and Coastal Oceanography". The future is bright for interpretable neural networks in geoscience and those who wish to apply them.

Neural networks are but one aspect of machine learning, which is also but one aspect of artificial intelligence. While neural networks have proven to be useful tools for geoscience, there will likely be new tools in the future of artificial intelligence and machine learning that will make similar - if not greater - advancements. It is therefore imperative that the geoscientific community continues to collaborate with the computer science community. It is likely that the geoscience community’s long history of dependence on new computational methods will continue, and so remaining open-minded to novel algorithms will be important to advance the field as efficiently as possible.

Bibliography

[1] Karumuri Ashok, Swadhin K Behera, Suryachandra A Rao, Hengyi Weng, and Toshio Yamagata. El Niño Modoki and its possible teleconnection. *Journal of Geophysical Research: Oceans*, 112(C11), 2007.

[2] Xin Wang and Chunzai Wang. Different impacts of various El Niño events on the Indian Ocean Dipole. *Climate dynamics*, 42(3-4):991–1005, 2014.

[3] Tong Lee and Michael J. McPhaden. Increasing intensity of El Niño in the central-equatorial Pacific. *Geophysical Research Letters*, 37(14), 2010.

[4] David WJ Thompson and John M Wallace. The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical research letters*, 25(9):1297–1300, 1998.

[5] Roland A. Madden and Paul R. Julian. Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *Journal of the atmospheric sciences*, 28(5):702–708, 1971.

[6] Chris Bretherton, V Balaji, T Delworth, RE Dickinson, JA Edmonds, JS Famiglietti, and LL Smarr. A national strategy for advancing climate modeling, 2012.

[7] D Menemenlis, C Hill, A Adcrocft, J-M Campin, B Cheng, B Ciotti, I Fukumori, P Heimbach, C Henze, A Köhl, et al. NASA supercomputer improves prospects for ocean climate research. *Eos, Transactions American Geophysical Union*, 86(9):89–96, 2005.

[8] Warren M Washington, Lawrence Buja, and Anthony Craig. The computational future for climate and Earth system models: on the path to petaflop and beyond. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890):833–846, 2009.

[9] Peter M. Atkinson and Adrian R. L. Tatnall. Introduction neural networks in remote sensing. *International Journal of remote sensing*, 18(4):699–709, 1997.

[10] James R Campbell, Dennis L Hlavka, Ellsworth J Welton, Connor J Flynn, David D Turner, James D Spinhirne, V Stanley Scott III, and IH Hwang. Full-time, eye-safe cloud and aerosol lidar observation at atmospheric radiation measurement program sites: Instruments and data processing. *Journal of Atmospheric and Oceanic Technology*, 19(4):431–442, 2002.

[11] David D Turner, RA Ferrare, LA Heilman Brasseur, WF Feltz, and TP Tooman. Automated retrievals of water vapor and aerosol profiles from an operational Raman lidar. *Journal of Atmospheric and Oceanic Technology*, 19(1):37–50, 2002.

[12] David D. Turner and Ulrich Löhnert. Information content and uncertainties in thermodynamic profiles and liquid cloud properties retrieved from the ground-based Atmospheric Emitted Radiance Interferometer (AERI). *Journal of Applied Meteorology and Climatology*, 53(3):752–771, 2014.

[13] Yolanda Gil, Suzanne A Pierce, Hassan Babaie, Arindam Banerjee, Kirk Borne, Gary Bust, Michelle Cheatham, Imme Ebert-Uphoff, Carla Gomes, Mary Hill, et al. Intelligent systems for geosciences: An essential research agenda. *Communications of the ACM*, 62(1):76–84, 2018.

[14] Anuj Karpatne, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[15] Ketil Malde, Nils Olav Handegard, Line Eikvil, and Arnt-Børre Salberg. Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, 2019.

[16] Karianne J Bergen, Paul A Johnson, V Maarten, and Gregory C Beroza. Machine learning for data-driven discovery in solid Earth geoscience. *Science*, 363(6433):eaau0323, 2019.

[17] Sid-Ahmed Boukabara, Vladimir Krasnopolsky, Jebb Q. Stewart, Stephen G. Penny, Ross N. Hoffman, and Eric Maddy. Artificial intelligence may be key to better weather forecasts. *EOS*, August 2019.

[18] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195, 2019.

[19] Alex Lopatka. Meteorologists predict better weather forecasting with AI. *Physics Today*, 72(5):32–34, May 2019.

[20] Amy McGovern, Kimberly L Elmore, David John Gagne, Sue Ellen Haupt, Christopher D Karstens, Ryan Lagerquist, Travis Smith, and John K Williams. Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, 98(10):2073–2090, 2017.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[24] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B I . Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[25] Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, 2017.

[26] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):645–657, 2016.

[27] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year ENSO forecasts. *Nature*, pages 1–5, 2019.

[28] Clara Deser, Michael A. Alexander, Shang-Ping Xie, and Adam S. Phillips. Sea surface temperature variability: Patterns and mechanisms. 2009.

[29] Matthew C. Wheeler and Harry H. Hendon. An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly weather review*, 132(8):1917–1932, 2004.

[30] Sheng Chen and Stephen A. Billings. Neural networks for nonlinear dynamic system modelling and identification. *International journal of control*, 56(2):319–346, 1992.

[31] Davide Castelvecchi. Can we open the black box of AI? *Nature News*, 538(7623):20, 2016.

[32] 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. *Bulletin of the American Meteorological Society*, (2019), 2019.

[33] Elizabeth A. Barnes, Benjamin Toms, James W. Hurrell, Imme Ebert-Uphoff, Chuck Anderson, and David Anderson. Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, page e2020MS002195, 2020.

[34] Karthik Kashinath, Mayur Mudigonda, Sol Kim, Lukas Kapp-Schwoerer, Andre Graubner, Ege Karaismailoglu, Leo von Kleist, Thorsten Kurth, Annette Greiner, Kevin Yang, et al. ClimateNet: an expert-labelled open dataset and Deep Learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development Discussions*, pages 1–28, 2020.

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[36] MW Gardner and SR Dorling. Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. *Atmospheric Environment*, 33(5):709–719, 1999.

[37] Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, and Jingjie Wang. Artificial neural networks forecasting of PM2. 5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment*, 107:118–128, 2015.

[38] Elizabeth A. Barnes, James W. Hurrell, Imme Ebert-Uphoff, Chuck Anderson, and David Anderson. Viewing forced climate patterns through an AI Lens. *Geophysical Research Letters*, 46(22):13389–13398, 2019.

[39] Ryan Lagerquist, Amy McGovern, and David John Gagne. Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting*, 34(4):1137–1160, 2019.

[40] David John Gagne, Sue Ellen Haupt, Douglas W Nychka, and Gregory Thompson. Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, 147(8):2827–2845, 2019.

[41] Benjamin A. Toms, Karthik Kashinath, Prabhat, and Da Yang. Testing the reliability of interpretable neural networks in geoscience using the Madden-Julian oscillation. *Geoscientific Model Development Discussions*, 2020:1–22, 2020.

[42] Thomas Bolton and Laure Zanna. Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, 11(1):376–399, 2019.

[43] Noah D. Brenowitz and Christopher S. Bretherton. Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, 11(8):2728–2744, 2019.

[44] Stephan Rasp, Michael S Pritchard, and Pierre Gentine. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39):9684–9689, 2018.

[45] Noah D Brenowitz and Christopher S Bretherton. Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, 45(12):6289–6298, 2018.

[46] F. Chevallier, F. Chéruy, N. A. Scott, and A. Chédin. A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of Applied Meteorology*, 37(11):1385–1397, 1998.

[47] Vladimir M. Krasnopolsky, Michael S. Fox-Rabinovitz, and Dmitry V. Chalikov. New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, 133(5):1370–1383, 2005.

[48] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[49] P. Sibi, S. Allwyn Jones, and P. Siddarth. Analysis of different activation functions using back propagation neural networks. *Journal of Theoretical and Applied Information Technology*, 47(3):1264–1268, 2013.

[50] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv preprint arXiv:2003.07631*, 2020.

[51] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

[52] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[53] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[54] 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, 2015.

[55] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[56] Wojciech Samek. *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer Nature, 2019.

[57] Andrei Dobrescu, Mario Valerio Giuffrida, and Sotirios A. Tsaftaris. Understanding deep neural networks for regression in leaf counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[58] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[59] Leila Arras, José Arjona-Medina, Michael Widrich, Grégoire Montavon, Michael Gillhofer, Klaus-Robert Müller, Sepp Hochreiter, and Wojciech Samek. Explaining and interpreting LSTMs. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 211–238. Springer, 2019.

[60] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. iNNvestigate Neural Networks! *Journal of Machine Learning Research*, 20(93):1–8, 2019.

[61] S George H Philander. El Nino southern oscillation phenomena. *Nature*, 302(5906):295, 1983.

[62] Eugene M Rasmusson and John M Wallace. Meteorological aspects of the El Nino/southern oscillation. *Science*, 222(4629):1195–1202, 1983.

[63] Shoji Hirahara, Masayoshi Ishii, and Yoshikazu Fukuda. Centennial-scale sea surface temperature analysis and its uncertainty. *Journal of Climate*, 30(20), 2017.

[64] Dietmar Dommenget, Tobias Bayr, and Claudia Frauen. Analysis of the non-linearity in the pattern and time evolution of El Niño southern oscillation. *Climate dynamics*, 40(11-12):2825–2847, 2013.

[65] Adam Hugh Monahan. Nonlinear principal component analysis: Tropical Indo–Pacific sea surface temperature and sea level pressure. *Journal of Climate*, 14(2):219–233, 2001.

[66] Yuan Zhang, John M. Wallace, and Naoto Iwasaka. Is climate variability over the North Pacific a linear response to ENSO? *Journal of Climate*, 9(7):1468–1478, 1996.

[67] Matthew Collins. Climate predictability on interannual to decadal time scales: The initial value problem. *Climate Dynamics*, 19(8):671–692, 2002.

[68] Francisco J Doblas-Reyes, Javier García-Serrano, Fabian Lienert, Aida Pintó Biescas, and Luis RL Rodrigues. Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, 4(4):245–268, 2013.

[69] NJ Dunstone, DM Smith, and R Eade. Multi-year predictability of the tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean. *Geophysical Research Letters*, 38(14), 2011.

[70] Chester F. Ropelewski and Michael S. Halpert. North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Monthly Weather Review*, 114(12):2352–2362, 1986.

[71] Klaus Wolter, Randall M Dole, and Catherine A Smith. Short-term climate extremes over the continental United States and ENSO. Part I: Seasonal temperatures. *Journal of Climate*, 12(11):3255–3272, 1999.

[72] Antonietta Capotondi, Prashant D Sardeshmukh, Emanuele Di Lorenzo, Aneesh C Subramanian, and Arthur J Miller. Predictability of US West Coast Ocean Temperatures is not solely due to ENSO. *Scientific reports*, 9(1):10993, 2019.

[73] Hui Wang and Mingfang Ting. Covariabilities of winter US precipitation and pacific sea surface temperatures. *Journal of Climate*, 13(20):3711–3719, 2000.

[74] Karen Aline McKinnon, Andrew Rhines, M. P. Tingley, and Peter Huybers. Long-lead predictions of eastern United States hot days from Pacific sea surface temperatures. *Nature Geoscience*, 9(5):389–394, 2016.

[75] Alexander Gershunov. ENSO influence on intraseasonal extreme rainfall and temperature frequencies in the contiguous United States: Implications for long-range predictability. *Journal of Climate*, 11(12):3192–3203, 1998.

[76] R Rohde, RA Muller, R Jacobsen, E Muller, S Perlmutter, A Rosenfeld, J Wurtele, D Groom, and C Wickham. A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinformats and Geostatistics: An Overview*, 7:2, 2013.

[77] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

[78] Benjamin A. Toms, Elizabeth A. Barnes, and Imme Ebert-Uphoff. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002002, 2020.

[79] Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. Can machines learn to predict weather? Using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, 11(8):2680–2693, 2019.

[80] Imme Ebert-Uphoff and Kyle A. Hilburn. Evaluation, tuning and interpretation of neural networks for meteorological applications. *arXiv preprint arXiv:2005.03126*, 2020.

[81] Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.

[82] Harry H. Hendon and Brant Liebmann. Organization of convection within the Madden-Julian oscillation. *Journal of Geophysical Research: Atmospheres*, 99(D4):8073–8083, 1994.

[83] George N. Kiladis, Katherine H. Straub, and Patrick T. Haertel. Zonal and vertical structure of the Madden-Julian oscillation. *Journal of the atmospheric sciences*, 62(8):2790–2809, 2005.

[84] 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. *Journal of Climate*, 23(15):4045–4059, 2010.

[85] Kai-Chih Tseng, Eric Maloney, and Elizabeth Barnes. The consistency of MJO teleconnection patterns: An explanation using linear Rossby wave theory. *Journal of Climate*, 32(2):531–548, 2019.

[86] Benjamin A. Toms, Elizabeth A. Barnes, Eric D. Maloney, and Susan C. van den Heever. The global teleconnection signature of the Madden-Julian oscillation and its modulation by the quasi-biennial oscillation. *Journal of Geophysical Research: Atmospheres*, page e2020JD032653, 2020.

[87] C. Zhang, ÁF Adames, B. Khouider, B. Wang, and D Yang. Four theories of the Madden-Julian oscillation. *Reviews of Geophysics*, page e2019RG000685, 2020.

[88] Adam Sobel and Eric Maloney. Moisture modes and the eastward propagation of the MJO. *Journal of the Atmospheric Sciences*, 70(1):187–192, 2013.

[89] Ángel F. Adames and Daehyun Kim. The MJO as a dispersive, convectively coupled moisture wave: Theory and observations. *Journal of the Atmospheric Sciences*, 73(3):913–941, 2016.

[90] Da Yang and Andrew P. Ingersoll. Testing the hypothesis that the MJO is a mixed Rossby–gravity wave packet. *Journal of the atmospheric sciences*, 68(2):226–239, 2011.

[91] Da Yang and Andrew P. Ingersoll. Triggered convection, gravity waves, and the MJO: A shallow-water model. *Journal of the atmospheric sciences*, 70(8):2476–2486, 2013.

[92] Ángel F. Adames and John M. Wallace. Three-dimensional structure and evolution of the MJO and its relation to the mean flow. *Journal of the Atmospheric Sciences*, 71(6):2007–2026, 2014.

[93] Joy M. Monteiro, Ángel F. Adames, John M. Wallace, and Jai S. Sukhatme. Interpreting the upper level structure of the Madden-Julian oscillation. *Geophysical Research Letters*, 41(24):9158–9165, 2014.

[94] Ángel F. Adames and John M. Wallace. Three-dimensional structure and evolution of the moisture field in the MJO. *Journal of the Atmospheric Sciences*, 72(10):3733–3754, 2015.

[95] Chidong Zhang and Min Dong. Seasonality in the Madden–Julian Oscillation. *Journal of Climate*, 17(16):3169–3180, 2004.

[96] Xianan Jiang, Ángel F. Adames, Ming Zhao, Duane Waliser, and Eric Maloney. A unified moisture mode framework for seasonality of the Madden-Julian oscillation. *Journal of Climate*, 31(11):4215–4224, 2018.

[97] George N. Kiladis, Juliana Dias, Katherine H. Straub, Matthew C. Wheeler, Stefan N. Tulich, Kazuyoshi Kikuchi, Klaus M. Weickmann, and Michael J. Ventrice. A comparison of OLR and circulation-based indices for tracking the MJO. *Monthly Weather Review*, 142(5):1697–1715, 2014.

[98] Brant Liebmann and Catherine A Smith. Description of a complete (interpolated) outgoing longwave radiation dataset. *Bulletin of the American Meteorological Society*, 77(6):1275–1277, 1996.

[99] Paul E. Roundy and William M. Frank. Applications of a multiple linear regression model to the analysis of relationships between eastward-and westward-moving intraseasonal modes. *Journal of the atmospheric sciences*, 61(24):3041–3048, 2004.

[100] Chongbo Zhao, Tim Li, and Tianjun Zhou. Precursor signals and processes associated with MJO initiation over the tropical indian ocean. *Journal of Climate*, 26(1):291–307, 2013.

[101] Ronald Gelaro, Will McCarty, Max J. Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A. Randles, Anton Darmenov, Michael G. Bosilovich, Rolf Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, 2017.

[102] H. T. Lee. Climate algorithm theoretical basis document (C-ATBD): Outgoing longwave radiation (OLR)-daily. NOAA's Climate Data Record (CDR) Program. *CDR-ATBD-0526*, 2014.

[103] Seok-Woo Son, Yuna Lim, Changhyun Yoo, Harry H Hendon, and Joowan Kim. Stratospheric control of the Madden-Julian oscillation. *Journal of Climate*, 30(6):1909–1922, 2017.

[104] Casey R. Densmore, Elizabeth R. Sanabia, and Bradford S. Barrett. QBO influence on MJO amplitude over the Maritime Continent: Physical mechanisms and seasonality. *Monthly Weather Review*, 147(1):389–406, 2019.

[105] S. Abhik and Harry H. Hendon. Influence of the QBO on the MJO during coupled model multiweek forecasts. *Geophysical Research Letters*, 46(15):9213–9221, 2019.

[106] Zane Martin, Frederic Vitart, Shuguang Wang, and Adam Sobel. The impact of the stratosphere on the MJO in a forecast model. *Journal of Geophysical Research: Atmospheres*, 2020.

[107] Taroh Matsuno. Quasi-geostrophic motions in the equatorial area. *Journal of the Meteorological Society of Japan. Ser. II*, 44(1):25–43, 1966.

[108] Changhyun Yoo and Seok-Woo Son. Modulation of the boreal wintertime Madden-Julian oscillation by the stratospheric quasi-biennial oscillation. *Geophysical Research Letters*, 43(3):1392–1398, 2016.

[109] Chidong Zhang and Bosong Zhang. QBO-MJO Connection. *Journal of Geophysical Research: Atmospheres*, 123(6):2957–2967, 2018.

[110] Noah D. Brenowitz, Tom Beucler, Michael Pritchard, and Christopher S. Bretherton. Interpreting and stabilizing machine-learning parametrizations of convection. *arXiv preprint arXiv:2003.06549*, 2020.

[111] Ben P. Kirtman and Paul S. Schopf. Decadal variability in ENSO predictability and prediction. *Journal of Climate*, 11(11):2804–2822, 1998.

[112] Shang-Ping Xie and Youichi Tanimoto. A pan-Atlantic decadal climate oscillation. *Geophysical Research Letters*, 25(12):2185–2188, 1998.

[113] Tim P. Barnett, David W. Pierce, R. Saravanan, Niklas Schneider, Dietmar Dommenget, and Mojib Latif. Origins of the midlatitude Pacific decadal variability. *Geophysical Research Letters*, 26(10):1453–1456, 1999.

[114] Niklas Schneider, Arthur J. Miller, and David W. Pierce. Anatomy of North Pacific decadal variability. *Journal of climate*, 15(6):586–605, 2002.

[115] Matthew Newman, Michael A. Alexander, Toby R. Ault, Kim M. Cobb, Clara Deser, Emanuele Di Lorenzo, Nathan J. Mantua, Arthur J. Miller, Shoshiro Minobe, Hisashi Nakamura, et al. The Pacific decadal oscillation, revisited. *Journal of Climate*, 29(12):4399–4427, 2016.

[116] Na Wen, Claude Frankignoul, and Guillaume Gastineau. Active AMOC–NAO coupling in the IPSL-CM5A-MR climate model. *Climate Dynamics*, 47(7-8):2105–2119, 2016.

[117] Gerald A. Meehl, Lisa Goddard, James Murphy, Ronald J. Stouffer, George Boer, Gokhan Danabasoglu, Keith Dixon, Marco A. Giorgetta, Arthur M Greene, E. D. Hawkins, et al. Decadal prediction: Can it be skillful? *Bulletin of the American Meteorological Society*, 90(10):1467–1486, 2009.

[118] Geert Jan van Oldenborgh, Francisco J. Doblas-Reyes, Bert Wouters, and Wilco Hazeleger. Decadal prediction skill in a multi-model ensemble. *Climate dynamics*, 38(7-8):1263–1280, 2012.

[119] S. G. Yeager, G. Danabasoglu, N. A. Rosenbloom, W. Strand, S. C. Bates, G. A. Meehl, A. R. Karspeck, K. Lindsay, M. C. Long, H. Teng, et al. Predicting near-term changes

in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bulletin of the American Meteorological Society*, 99(9):1867–1886, 2018.

[120] D. M. Smith, R. Eade, Adam A. Scaife, L. P. Caron, G. Danabasoglu, T. M. DelSole, T. Delworth, F. J. Doblas-Reyes, N. J. Dunstone, L. Hermanson, et al. Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science*, 2(1):1–10, 2019.

[121] Riley X. Brady, Nicole S. Lovenduski, Stephen G. Yeager, Matthew C. Long, and Keith Lindsay. Skillful multiyear predictions of ocean acidification in the California Current System. *Nature Communications*, 11(1):1–9, 2020.

[122] Isla R Simpson, Stephen G Yeager, Karen A McKinnon, and Clara Deser. Decadal predictability of late winter precipitation in western Europe through an ocean–jet stream connection. *Nature Geoscience*, 12(8):613–619, 2019.

[123] Yujun He, Bin Wang, Mimi Liu, Li Liu, Yongqiang Yu, Juanjuan Liu, Ruizhe Li, Cheng Zhang, Shiming Xu, Wenyu Huang, et al. Reduction of initial shock in decadal predictions using a new initialization strategy. *Geophysical Research Letters*, 44(16):8538–8547, 2017.

[124] Jürgen Kröger, Holger Pohlmann, Frank Sienz, Jochem Marotzke, Johanna Baehr, Armin Köhl, Kameswarrao Modali, Iuliia Polkova, Detlef Stammer, Freja SE Vamborg, et al. Full-field initialized decadal predictions with the MPI earth system model: An initial shock in the North Atlantic. *Climate Dynamics*, 51(7-8):2593–2608, 2018.

[125] David E Black, Larry C Peterson, Jonathan T Overpeck, Alexey Kaplan, Michael N Evans, and Michaele Kashgarian. Eight centuries of North Atlantic Ocean atmosphere variability. *Science*, 286(5445):1709–1713, 1999.

[126] Ping Chang, Link Ji, and Hong Li. A decadal climate variation in the tropical Atlantic Ocean from thermodynamic air-sea interactions. *Nature*, 385(6616):516–518, 1997.

[127] Emilia K Jin, James L Kinter, Bin Wang, C-K Park, I-S Kang, BP Kirtman, J-S Kug, A Kumar, J-J Luo, J Schemm, et al. Current status of ENSO prediction skill in coupled ocean-atmosphere models. *Climate Dynamics*, 31(6):647–664, 2008.

[128] Hyemi Kim, Frédéric Vitart, and Duane E Waliser. Prediction of the Madden-Julian oscillation: A review. *Journal of Climate*, 31(23):9425–9443, 2018.

[129] Hyemi Kim, Matthew A Janiga, and Kathy Pegion. MJO propagation processes and mean biases in the SubX and S2S reforecasts. *Journal of Geophysical Research: Atmospheres*, 124(16):9314–9331, 2019.

[130] RD Koster, SPP Mahanama, TJ Yamada, Gianpaolo Balsamo, AA Berg, M Boisserie, PA Dirmeyer, FJ Doblas-Reyes, G Drewitt, CT Gordon, et al. The second phase of the global land–atmosphere coupling experiment: Soil moisture contributions to subseasonal forecast skill. *Journal of Hydrometeorology*, 12(5):805–822, 2011.

[131] Timothy DelSole and Arindam Banerjee. Statistical seasonal prediction based on regularized regression. *Journal of Climate*, 30(4):1345–1361, 2017.

[132] Daniel S Wilks. Improved statistical seasonal forecasts using extended training data. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 28(12):1589–1598, 2008.

[133] Gokhan Danabasoglu, J-F Lamarque, J Bacmeister, DA Bailey, AK DuVivier, Jim Edwards, LK Emmons, John Fasullo, R Garcia, Andrew Gettelman, et al. The Community Earth System Model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020.

[134] Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[135] Bo Qiu and Shuiming Chen. Variability of the Kuroshio Extension jet, recirculation gyre, and mesoscale eddies on decadal time scales. *Journal of Physical Oceanography*, 35(11):2090–2103, 2005.

[136] Richard Kleeman, Julian P McCreary Jr, and Barry A Klinger. A mechanism for generating ENSO decadal variability. *Geophysical Research Letters*, 26(12):1743–1746, 1999.

[137] Matthew Newman, Gilbert P Compo, and Michael A Alexander. ENSO-forced variability of the Pacific decadal oscillation. *Journal of Climate*, 16(23):3853–3857, 2003.

[138] Nathan J Mantua and Steven R Hare. The Pacific decadal oscillation. *Journal of oceanography*, 58(1):35–44, 2002.

[139] Jeff R Knight, Robert J Allan, Chris K Folland, Michael Vellinga, and Michael E Mann. A signature of persistent natural thermohaline circulation cycles in observed climate. *Geophysical Research Letters*, 32(20), 2005.

[140] Iselin Medhaug, Helene Reinertsen Langehaug, Tor Eldevik, Tore Furevik, and Mats Bentsen. Mechanisms for decadal scale variability in a simulated Atlantic meridional overturning circulation. *Climate dynamics*, 39(1-2):77–93, 2012.

[141] Evgenia Dimitriadou, Sara Dolničar, and Andreas Weingessel. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1):137–159, 2002.

[142] Jeff R Knight, Chris K Folland, and Adam A Scaife. Climate impacts of the Atlantic multidecadal oscillation. *Geophysical Research Letters*, 33(17), 2006.

[143] Arun Kumar and Martin P. Hoerling. Annual cycle of Pacific–North American seasonal predictability associated with different phases of ENSO. *Journal of Climate*, 11(12):3295–3308, 12 1998.

[144] S. J. Woolnough, F. Vitart, and M. A. Balmaseda. The role of the ocean in the Madden–Julian Oscillation: Implications for MJO prediction. *Quarterly Journal of the Royal Meteorological Society*, 133(622):117–128, 2007.

[145] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[146] Matt W. Gardner and S. R. Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[147] William M. Putman and Max Suarez. Cloud-system resolving simulations with the NASA Goddard Earth Observing System global atmospheric model (GEOS-5). *Geophysical Research Letters*, 38(16), 2011.

[148] Masaki Satoh, Bjorn Stevens, Falko Judt, Marat Khairoutdinov, Shian-Jiann Lin, William M. Putman, and Peter Düben. Global cloud-resolving models. *Current Climate Change Reports*, 5(3):172–184, 2019.

Appendix A

Appendix and Supplementary Material for Chapter 2

Within this supplemental material, we list additional rules for propagation, resources for implementing and using LRP in neural network packages other than $Keras$, and sample Python code.

A.1 Additional Relevance Propagation Rules

We list additional relevance propagation rules here that are most relevant for LRP as we present it within the main manuscript, although we do not intend for this list to be comprehensive. For additional information, readers should refer to Chapter 10 of Samek et al. (2019), which provides a detailed discussion of the theory behind the various propagation rules developed thus far for LRP.

We mention within the main manuscript that we use a set of relevance propagation rules that only propagates information that increases the value of the selected output, although there are rules that permit the inclusion of information that reduces the value of the selected output as well. There are caveats to these additional rules, however, as the relevance is not conserved from the output value back to the input layer, and so interpretation of the relevance heatmaps becomes more subjective. We share the rules below for completeness, but we encourage the reader to be careful in their application and to carefully read the theory behind their development as provided by Bach et al. (2015), Montavon et al. (2017), and Samek et al. (2019).

The relevance propagation rule that includes information that reduces the target output node is as follows:

$$R_i = \sum_j \left(\alpha \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j - \beta \frac{a_i w_{ij}^-}{\sum_i a_i w_{ij}^-} \right) R_j. \tag{A.1}$$

Within Equation A.1, α represents the relative amount of positive relevance to be propagated backwards, $beta$ represents the relative amount of negative relevance to be propagated backwards, the i

subscript represents the i-th node in the layer of the network to which the relevance is being translated backwards, the j subscript represents the j-th node in the layer of the network from which the relevance is being translated, R_i is the relevance translated backwards to the i-th node, R_j is the relevance of the j-th node, a_i is the output from the i-th node after the non-linearity has been applied when the sample is passed forward through the network, w_{ij} is the weight of the connection between the i-th and j-th nodes, and the $+$ and $-$ superscripts signifies that only positive or negative weights are considered, respectively.

There is also a separate propagation rule for relevance propagated backwards to the input layer from the first hidden layer for cases where the input is bounded between two values. In these cases, the relevance propagation rule is as follows:

$$R_i = \sum_j \left(\frac{x_i w_{ij} - l w_{ij}^{+} - h w_{ij}^{-}}{\sum_i x_i w_{ij} - l w_{ij}^{+} - h w_{ij}^{-}} \right) R_j. \tag{A.2}$$

Within Equation A.2, the i subscript represents the i-th node in the layer of the network to which the relevance is being translated backwards, the j subscript represents the j-th node in the layer of the network from which the relevance is being translated, x represents the input value associated with the i-th node, l represents the lower bound of the input dataset, h represents the upper bound of the input dataset, R_i is the relevance translated backwards to the i-th node, and R_j is the relevance of the j-th node. This rule also conserves the total relevance translated from the first hidden layer backwards to the input layer, and abides by the rules of deep Taylor decomposition. If this rule is used in tandem with the rules presented within the manuscript that only propagate information backwards that increases the value of the output, then the relevance is conserved as it is propagated from the output node to the input layer. Additional information about deep Taylor decomposition is6 available within Montavon et al. (2017).

A.2 Additional Resources for LRP

We offer a list of resources for programming LRP to be compatible with various $Python$ packages, although we do not intend for this list to be comprehensive. We use $innvestigate$, the first item in the below list, which is an implementation of LRP and various other neural network interpretation techniques for the $Keras$ neural network package within Python.

- Python, $Keras/Tensorflow$ package: https://github.com/albermax/innvestigate
- Python, $Keras/Tensorflow$ package: https://github.com/nielsrolf/tensorflow-lrp
- Python, $Keras/Tensorflow$ package (coming soon): https://github.com/sicara/tf-explain
- Python, $Caffe$ package: https://github.com/sebastian-lapuschkin/lrp_toolbox
- Python, $PyTorch$ package: https://github.com/moboehle/Pytorch-LRP
- MATLAB: http://www.heatmapping.org/lrptoolbox.html

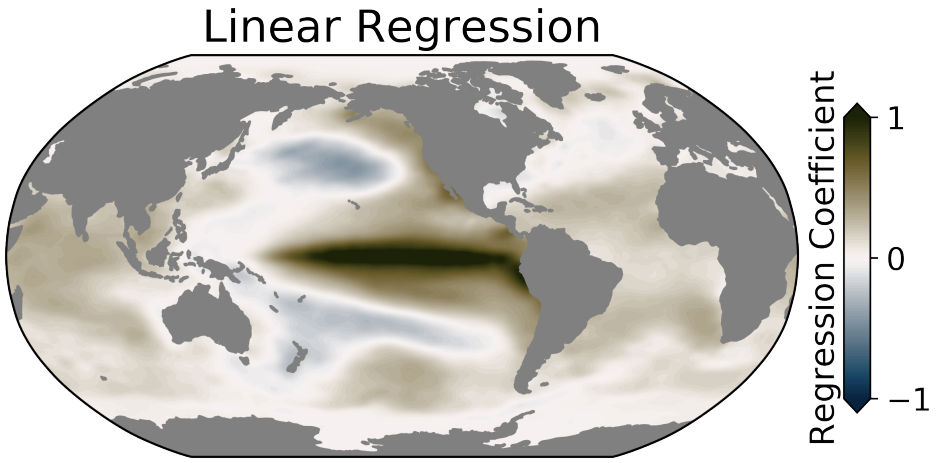

Figure A.1: Regression coefficients for the linear regression method to predict the phase of ENSO using global maps of sea-surface temperature anomalies. The regression coefficients are calculated by regressing the time series of global sea-surface temperature anomaly maps onto the standardized ENSO (Niño3.4) time series. We then project this map of regression coefficients onto the global sea-surface temperature anomalies to predict the sign of ENSO phase. The resultant accuracy predicting the ENSO phase using the regression coefficients is 82.5%.

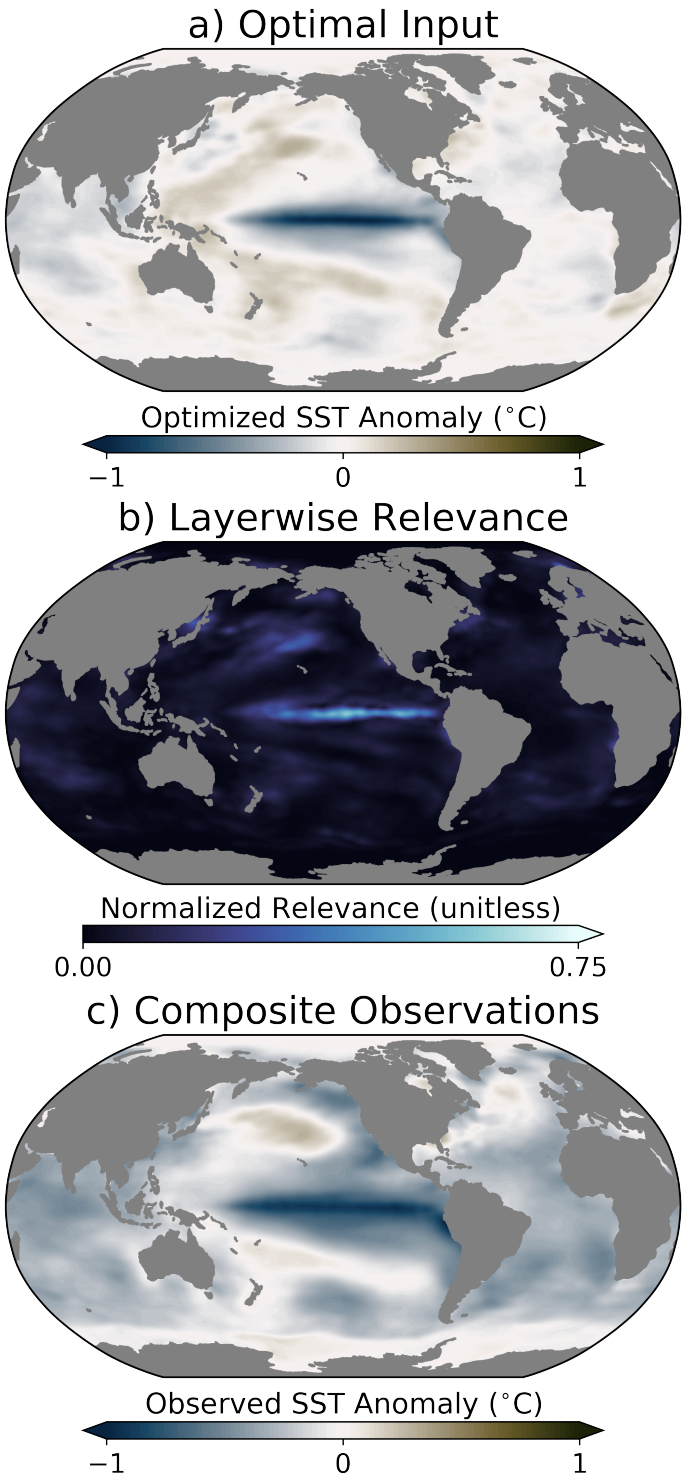

Figure A.2: As in Figure 2.6, but for the neural network's understanding of the spatial structure of La Niña based on a total of 485 samples (including both testing and training data).

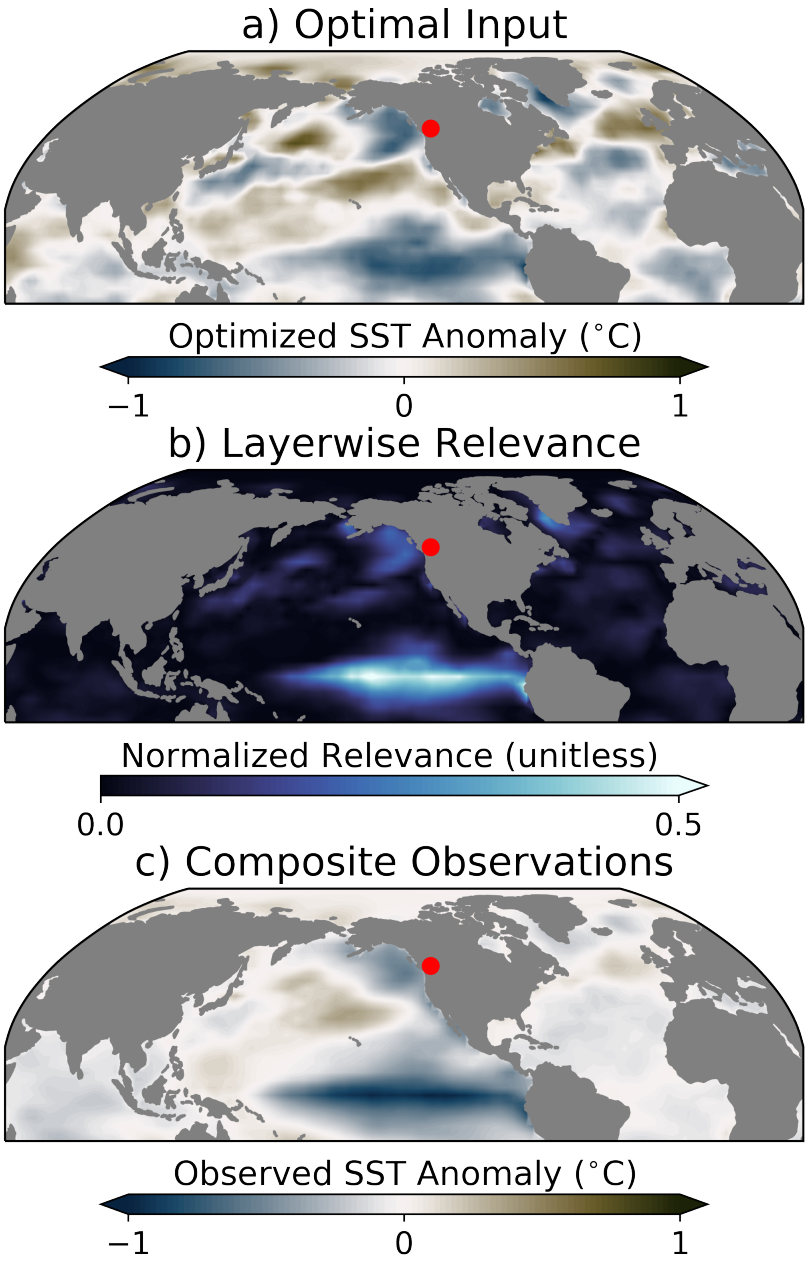

Figure A.3: As in Figure 2.9, but for the neural network's understanding of the sea-surface temperature anomalies that lend predictability for *negative* surface temperature anomalies at the red dot.

Appendix B

Appendix and Supplementary Material for Chapter 4

B.1 Neural Network Details

This section includes details of how the neural networks were trained for Chapter 4 of this dissertation. Each neural network was trained using the Adam optimizer, with an initial learning rate of 1E-4. We do not change the learning rate throughout training. The single hidden layer of neurons is regularized with a L2 (ridge) regularization coefficient of 10, which ensures the neural network uses information from broader spatial regions and can not overfit to individual locations. The networks were allowed to train for 100 epochs, which was sufficient for convergence in all cases. The model iteration that resulted in the highest accuracy on the validation data was selected and used for analysis. We train five neural networks for each location because it is possible that each network will find a different optimal solution, and so training numerous networks increases the likelihood that we capture the full range of optimal solutions. The accuracy values presented in Figure 2 represent the mean accuracy from the five networks. The interpretations presented in Figures 4.3, 4.4, and 4.5 are similar across each of the five network iterations, and so we randomly select one of the five neural networks and use this network for these analyses. We find that the networks converge on similar optimal solutions based on the LRP interpretations, and so training five models is sufficient for our purposes.

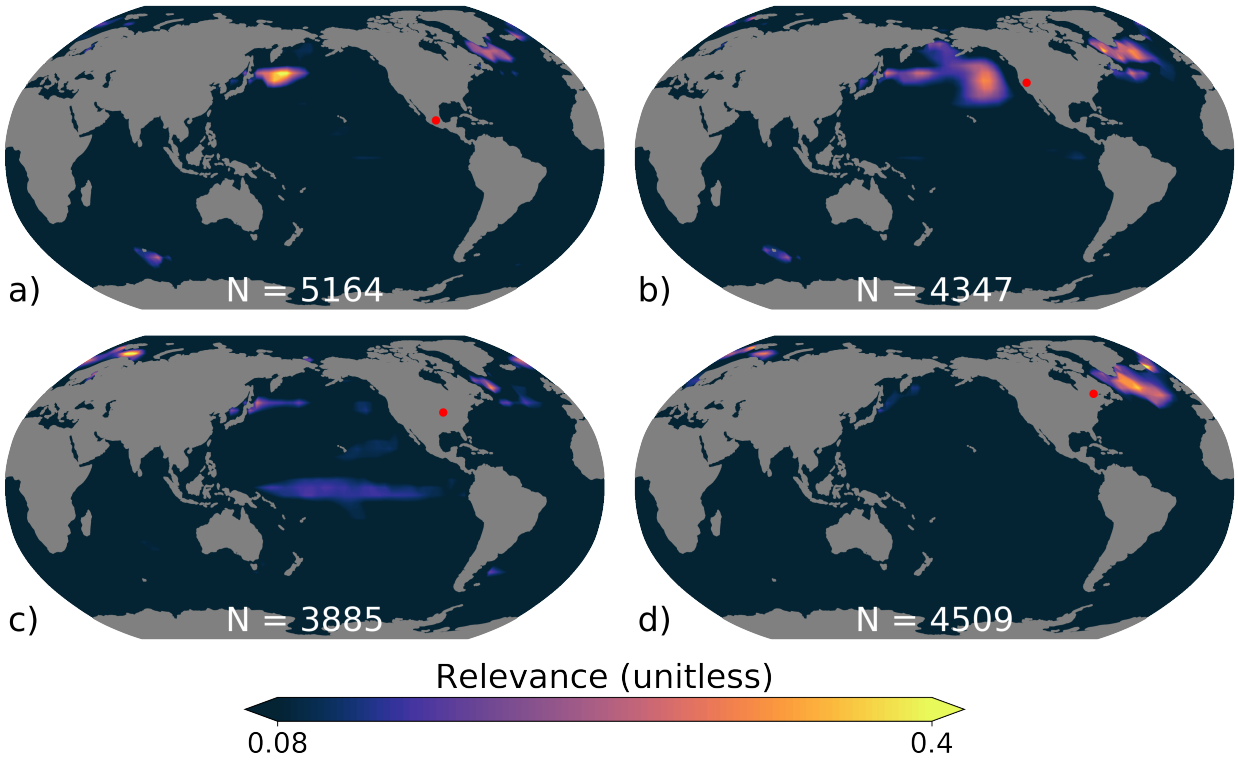

Figure B.1: As in Figure 4.3, but for negative surface temperature anomalies at the locations denoted by the red dots.

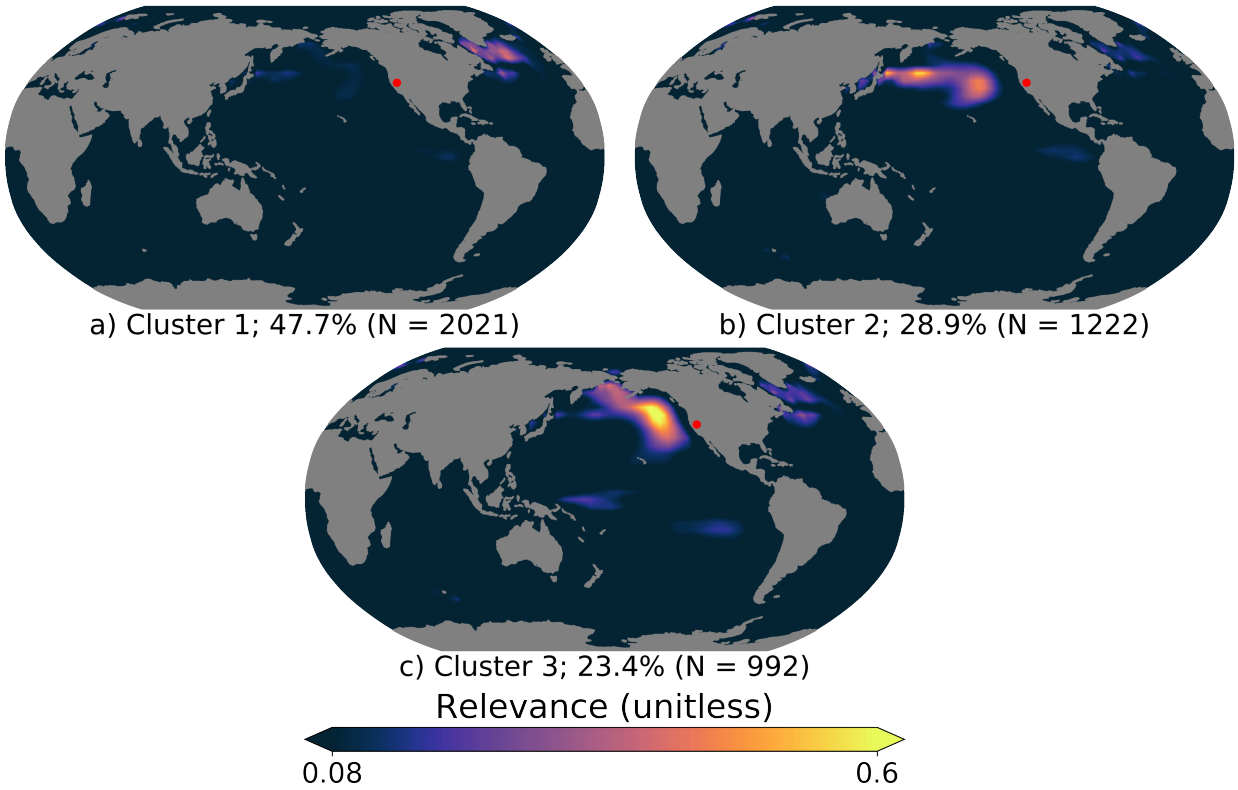

Figure B.2: As in Figure 4.4, but for negative surface temperature anomalies at the location denoted by the red dot.