DISSERTATION

WHEN IS THE UNPREDICTABLE (SLIGHTLY MORE) PREDICTABLE? AN ASSESSMENT OF OPPORTUNITIES FOR SKILLFUL DECADAL CLIMATE PREDICTION USING EXPLAINABLE NEURAL NETWORKS

Submitted by

Emily M. Gordon

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2023

Doctoral Committee:

Advisor: Elizabeth A. Barnes

James W. Hurrell
Maria Rugenstein
Charles Anderson

ABSTRACT

WHEN IS THE UNPREDICTABLE (SLIGHTLY MORE) PREDICTABLE? AN ASSESSMENT OF OPPORTUNITIES FOR SKILLFUL DECADAL CLIMATE PREDICTION USING EXPLAINABLE NEURAL NETWORKS

Predicting climate variability on decadal (2-10 year) timescales can have huge implications for society because it can provide better estimates of both global trends as well as regional climate variability for crucial, actionable lead times. The key to skillful decadal prediction is understanding and predicting oceanic variability. However, predictable signals in the ocean can be masked by the inherent noise in the system, and therefore, skillful prediction on decadal timescales is challenging. Machine learning, with its ability to extract non-linear signals from large sets of noisy data, has been shown a powerful tool for predicting and understanding processes across weather and climate applications. In this dissertation, I explore applications of machine learning to decadal prediction.

First, I present a machine learning approach to predicting the Pacific decadal oscillation (PDO) with artificial neural networks (ANNs) within the Community Earth System Model version 2 (CESM2) pre-industrial control simulation. Predicting PDO transitions and understanding the associated mechanisms has proven a critical but challenging task in climate science. As a form of decadal variability, the PDO is associated with both large-scale climate shifts and regional climate predictability. I show that ANNs predict PDO persistence and transitions at lead times of 12 months onward. Using layer-wise relevance propagation to investigate the ANN predictions, I demonstrate that the ANNs utilize oceanic patterns that have been previously linked to predictable PDO behavior. ANNs recognize a build-up of ocean heat content in the off-equatorial western Pacific 12–27

months before a transition occurs. The ANNs also distinguish transition mechanisms between positive-to-negative sign transitions, and negative-to-positive transitions.

Secondly, I demonstrate a technique for incorporating an uncertainty estimate into the prediction of a regression neural network, allowing the identification of predictable sea surface temperature (SST) anomalies on decadal timescales in the CESM2 pre-industrial control simulation. Predictability in SSTs can be masked by unpredictable variability, and one approach to extracting predictable signals is to investigate state-dependent predictability – how differences in prediction skill depend on the initial state of the system. I leverage the network's prediction of uncertainty to examine state-dependent predictability in SSTs by focusing on predictions with the lowest uncertainty. In particular, I study two regions of the global ocean–the North Atlantic and North Pacific–and find that skillful initial states identified by the neural network correspond to particular phases of low frequency variability in the North Pacific and North Atlantic Oceans.

Finally, I examine the potential role of predictable internal variability in a future, warmer climate by designing an interpretable neural network that can be decomposed to examine the relative contributions of external forcing and internal variability to future regional decadal SST trend predictions. I show that there is additional prediction skill to be garnered from internal variability in the CESM2 Large Ensemble in the near-term climate (2020-2050), even in a relatively high forcing future scenario. This predictability is especially apparent in the North Atlantic, North Pacific and Tropical Pacific Oceans as well as in the Southern Ocean. I further investigate how prediction skill covaries across the ocean and find three regions with distinct coherent prediction skill driven by internal variability. SST trend predictability is found to be associated with consistent patterns of interannual and decadal variability for the grid points within each region.

ACKNOWLEDGEMENTS

The work in this dissertation was partially funded by Fulbright New Zealand, and NSF CAREER AGS-1749261 under the Climate and Large-scale Dynamics program. I would also like to thank the Colorado State University Distinguished Professors for awarding me the Distinguished Professors award, allowing me to travel and present this work at the EGU general assembly in Vienna and spend two weeks at the National Center for Atmospheric Research in Boulder.

As the saying goes, it takes a village to raise a PhD Dissertation, so I must thank my village:

My wonderful advisor Libby Barnes deserves every good thing in the world. My committee Profs. Jim Hurrell, Maria Rugenstein and Chuck Anderson have provided excellent outside guidance and support. The Barnes research group throughout my time have been an unparalleled source of mentorship and, more importantly, friendship. And lastly all my friends and family in Fort Collins and back in New Zealand. And finally Rosa the dog for keeping me company during my 4am classes.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: Introduction

Trustworthy and actionable decadal (2-10 years) predictions of the Earth's climate have the potential to provide much needed information about future climate to a range of regions and sectors that are susceptible to climate risk and change [1, 2, 3]. Such predictions can provide information about near-term likelihood of, e.g., crop stress [4], regional flooding[5], drought [6], hurricane activity [7] and energy demand [3]. Forecasting on timescales beyond a season presents an interesting challenge, because these forecasts must account for both variations in the boundary conditions in the system (i.e., external forcing) and the initial conditions (state of climate at the beginning of the prediction window) (Figure 1.1a grey box). Specifically, initial condition predictability on decadal timescales largely derives from patterns of low frequency oceanic variability such as the Pacific decadal oscillation [PDO; 8, 9, 10], or Pacific Decadal Variability (PDV) as it is also termed, as well as Atlantic multi-decadal variability [AMV; 11, 12]. Furthermore, predictability on decadal timescales can also originate from the climate's response to anthropogenic forcing, such as greenhouse gas and aerosol emissions. For example, if we expect greenhouse gas emissions to continue to increase, then we would also expect global temperatures to rise [13]. This consequence of anthropogenic climate change therefore provides a measure of predictability in the Earth system for the next 10 years.

However, there is still debate around how much predictability is provided by internal variability on decadal timescales, because predictability varies both by location and initial state [14, 15]. For example, predictions of sea surface temperatures (SSTs) in mid-latitude ocean basins can be more predictable for 2-10 years depending on recent evolution of ocean heat content and transport, while the tropical oceans are generally only predictable for approximately a season to a year [16, 17]. Therefore, if an initial state for some variable is more predictable, then the forecasted range is constrained by the initial state of the system (Figure 1.1b). Conversely, for some initial conditions the range of forecasted outcomes

becomes indistinguishable from the full background range of variability soon after the initial prediction state (e.g., Figure 1.1c), so the initial conditions do not provide improved estimation of the future after the time when the initial condition information is lost.

It is therefore valuable to develop methods to identify and understand initial states that are more likely to evolve predictably on decadal timescales. This has led to an increased focus on identifying state-dependence in decadal predictions, or so-called "windows of opportunity" for increased prediction skill [18, 19]. This is an approach that has yielded results in the field of subseasonal-to-seasonal prediction, where it has been established that methods for *a priori* identification of more predictable initial states (i.e., windows of opportunity) can provide much needed forecast skill for these timescales [20, 21, 22]. On decadal timescales, there is evidence that predictability is state-dependent, and hence identifying these windows of opportunity can lead to enhanced prediction skill [23, 24, 25]. For example, positive heat content anomalies in the North Atlantic ocean can lead to enhanced prediction skill in the North Atlantic subpolar gyre up to 10 years later, while negative anomalies in the North Atlantic do not provide such skill [26]. While there has been some research into case studies of windows of opportunity for enhanced prediction skill on decadal timescales, there remains a need for a systematic assessment of state-dependent predictability to provide usable information about predictable variability.

In conjunction with an increased focus on identifying windows of opportunity across prediction horizons, there have been improvements in data-driven approaches to Earth system predictability, particularly machine learning approaches [27]. These methods harvest the enormous amount of climate model and observational data available to generate machine learning models that predict Earth system processes across timescales, e.g., the Madden-Julian oscillation [28], the El Nino Southern Oscillation [29], midlatitude circulation [21] and sea surface temperature [30, 31]. Simultaneously, advances in explainable AI (XAI) and machine learning interpretability methods have allowed examination into

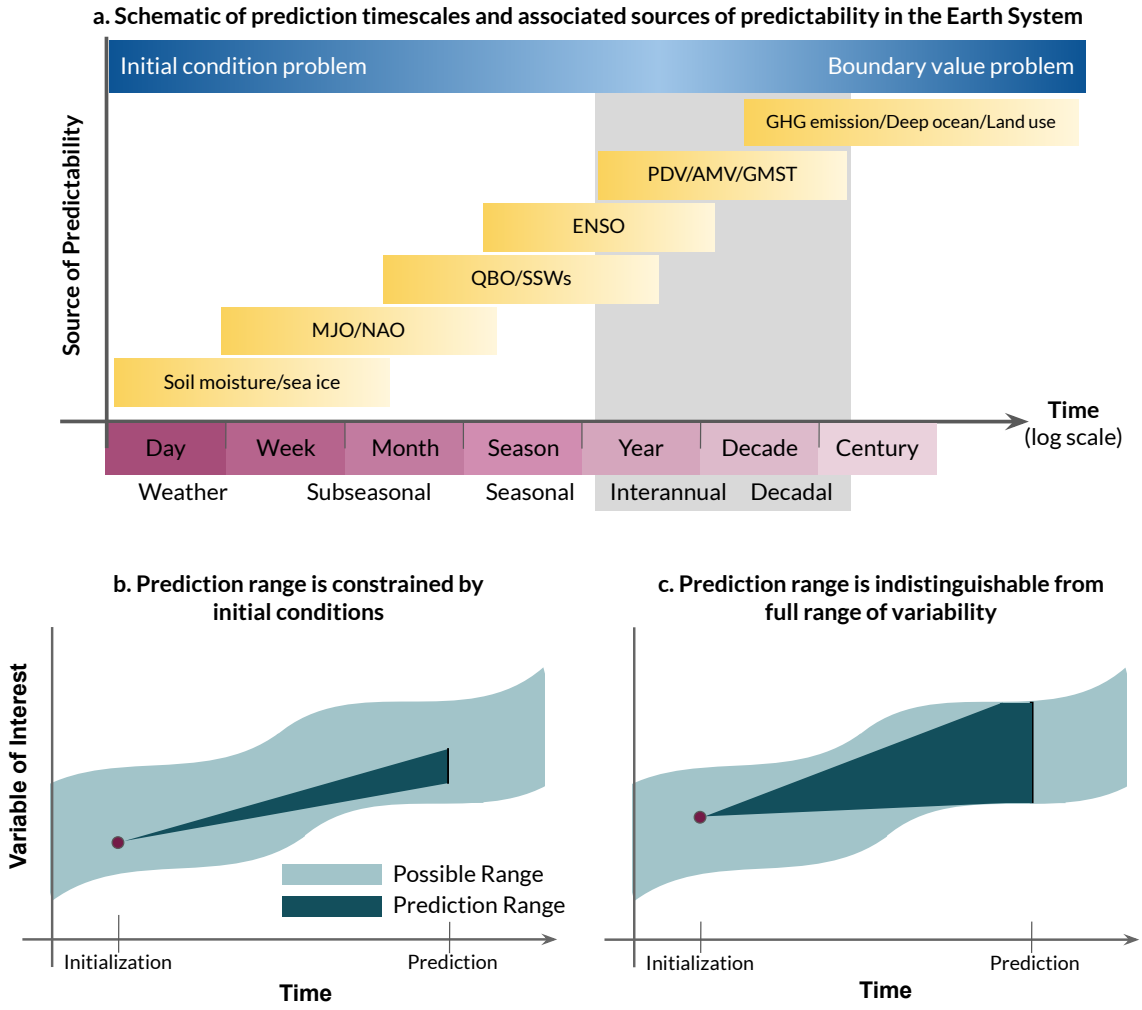

Figure 1.1: (a) Sources of predictability in the Earth system (yellow boxes) and their accompanying timescale (purple boxes, x axis). The grey box indicates the timescale of focus in this dissertation. This panel is adapted from Meehl et al. [1]. Sources include the Madden-Julian oscillation (MJO), North Atlantic oscillation (NAO), Quasi-biennial oscillation (QBO), El Nino Southern oscillation (ENSO), Pacific decadal variability (PDV), Atlantic multidecadal variability (AMV), global mean surface temperature (GMST) and greenhouse gas (GHG) emissions (b) Schematic of a climate variable whose predicted evolution (dark blue shading) from the initial state (purple dot) is smaller than the total possible range of variability (teal shading). (c) Same as panel b but for a variable whose predicted range is indistinguishable from the total range of variability at the prediction.

machine learning decision making processes to both increase trust in the model [32, 21, 33], and examine mechanisms leading to skillful predictions [34, 35, 36].

Therefore, there is an opportunity to provide much needed understanding of decadal prediction and predictability using the recent advances in machine learning. This dissertation focuses on three questions: Can machine learning predict low frequency modes of variability and increase our understanding of these processes? Can we then adapt existing machine learning methods for identifying state-dependent predictability, and its associated patterns of variability? And will these sources of predictability be usable in a future climate that is dominated by anthropogenic greenhouse warming? Chapter 2 addresses the first question by examining predictability of the PDO using a simple artificial neural network and explainability methods. In Chapter 3, I adapt a novel machine learning regression technique to provide a systematic method to identify decadal state-dependent predictability in SSTs across the ocean. In Chapter 4, I develop an interpretable neural network architecture to separate the contributions of internal variability and external forcing to decadal SST trend predictability in a future climate scenario. Finally, Chapter 5 summarizes the implications of the novel research in this dissertation, and discusses avenues for future investigation inspired by this work. The content in Chapters 2 and 3 has been peer-reviewed and published in Geophysical Research Letters, and Chapter 4 is currently in review in Environmental Research Letters. [37, 38, 39, full citations below]. As such, these papers have been reproduced in this dissertation without alteration.

- Gordon, E. M., Barnes, E. A., & Hurrell, J. W. (2021). Oceanic harbingers of Pacific Decadal Oscillation predictability in CESM2 detected by neural networks. Geophysical Research Letters, 48, e2021GL095392. doi: 10.1029/2021GL095392

- Gordon, E. M., & Barnes, E. A. (2022). Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in CESM2. Geophysical Research Letters, 49, e2022GL098635. doi: 10.1029/2022GL098635

- Gordon, E. M., Barnes E. A. & Davenport. F. V. (2023). Separating Internal and Forced Contributions to Near Term SST Predictability in the CESM2-LE. Submitted to Environmental Research Letters. Preprint Available: doi: 10.31223/X5BD5J

Chapter 2: Oceanic Harbingers of Pacific Decadal Oscillation Predictability in CESM2 Detected by Neural Networks

2.1 Introduction

The Pacific Decadal Oscillation [PDO; 8, 9] is recognised as one of the most important sources of predictability on decadal timescales [40]. As such it has been linked to increased predictability of surface variables, including precipitation and temperature, as well as being an important factor in marine ecosystems and resource management. The PDO is not itself considered a single mode of variability, but a manifestation of several different forcings operating on different timescales: the integration of stochastic atmospheric forcing associated with the Aleutian low; tropical-subtropical atmospheric teleconnections associated with the El Nino Southern Oscillation (ENSO) phenomenon; the re-emergence of winter-to-winter sea surface temperature (SST) anomalies; and ocean gyre dynamics [10, and the references therein]. In its positive phase, the PDO manifests as a pattern of negative SST anomalies in the central and western North Pacific Ocean, surrounded by positive anomalies around the eastern edge, extending southward to around 20°N (Figure 2.1a).

While the combination of mechanisms that contribute to the PDO are considered to be largely understood, challenges still exist in the realm of PDO predictability [40]. This is especially true in predicting PDO transitions, i.e., when the PDO shifts from one phase to the other. Stochastic models [41, 42, 43], linear inverse models [LIMs; 44, 45, 46], atmosphere-only models [47] and fully coupled climate models [48, 49] have been used to recreate the relevant processes that contribute to PDO variability and, by comparing to observations, attempt to estimate how these processes can lead to predictability. This has lead to a single robust theory for PDO transitions: studying periods of mega-droughts, Meehl and Hu [48] posited that tropical SST anomalies drive surface wind-stress anomalies in the off-equatorial Pacific ($\sim 25^{\circ}$) via atmospheric teleconnections, forcing oceanic Rossby waves

that propagate westward on decadal timescales. This results in a build-up of ocean heat content in the off-equatorial western Pacific. If an ENSO event subsequently switches the sign of the tropical Pacific SST anomaly, this off-equatorial heat is redistributed via Kelvin waves throughout the equatorial region, leading to a transition in the PDO. Meehl et al. [50] investigate this mechanism in the context of the Interdecadal Pacific Oscillation (IPO; similar to the PDO but the spatial domain spans the full meridional extent of the Pacific), finding that initialized hindcasts with the Community Climate System Model, Version 4, [CCSM4; 51] show skill in simulating past IPO transitions with this mechanism appearing to coincide with those particular transitions. Since the PDO is considered the North Pacific manifestation of the IPO, the mechanism outlined above is directly relevant to understanding and predicting PDO transitions [47, 52].

While stochastic climate models and LIMs model the climate system as linear, it has been suggested that predictive skill, especially of oceanic variability, could be gained using methods that better capture non-linearities in the system [44]. Artificial neural networks (ANNs), a form of machine learning, offer such a non-linear framework and have proven skillful at predicting processes in the climate system such as identifying the forced response to climate change, ENSO evolution and Madden-Julian Oscillation teleconnections [53, 54, 32, 21]. Specifically in the case of oceanic predictability, Ham et al. [54] used a convolutional neural network to predict ENSO evolution, showing significantly higher forecast skill than previous dynamical forecasts, while also identifying spatial SST patterns corresponding to increased predictability. Similarly, Nadiga [30] demonstrated how reservoir computing (a form of recurrent neural networks) increases predictability of oceanic variability in the North Atlantic Ocean on the interannual timescale, especially during period of infrequent or missing data. Together, these studies suggest that neural networks are effective for investigating and predicting climate processes related to oceanic variability. These, along with explainable AI (XAI, methods designed to aid the interpretation of the decision-

making process of a neural network) can identify signals associated with a neural network's prediction.

In this study we show that ANNs are effective tools for predicting persistence and transitions in the PDO. In our analysis we examine predictions with lead-times from 12 months onward. Recall the PDO is considered a combination of forcings that propagate on different timescales, from stochastic atmospheric forcing on the timescale of days to weeks, to oceanic Rossby wave propagation on multi-year scales [10]. We examine predictability on the shorter than "decadal" timescales to avoid averaging out the forcings on shorter timescales that may contribute to predictive skill. We choose to still use the PDO terminology, however, as we are investigating predictability of the PDO spatial pattern across various timescales.

Furthermore, we investigate mechanisms identified by the ANNs that lead to predictability, both long-term persistence and predicting transitions. Most notably, we leverage explainable AI methods to attribute patterns of ocean heat content anomalies to increased PDO predictability. We emphasize that not only are we concerned with optimizing an ANN to solve a prediction problem, but we also explore the decision making process of the ANN to uncover potential sources of predictability [32].

2.2 Data and Methods

2.2.1 Data

We use monthly mean sea surface temperature (SST) and ocean heat content (OHC) from the Community Earth System Model Version 2 [CESM2; 55] pre-industrial control run for the Coupled Model Intercomparison Project, Phase 6 [CMIP6; 56]. The presence of realistic ENSO and PDO variability in CESM2 was demonstrated by Capotondi et al. [57]. We use the full 2000 year run, with the large amount of data available (24000 months) desirable for training the ANNs. OHC is calculated as the vertical heat content integral from the surface to 100 m depth [58]. Both OHC and SST are interpolated to a $4^\circ \times 4^\circ$ grid

and we deseasonalize both the SST and OHC fields by subtracting their respective monthly mean annual cycles at each grid point. Furthermore for OHC (the input for the ANNs), we standardize each grid point by dividing it by its monthly standard deviation and apply a 6-month running mean. These preprocessing steps ensure that the input and output data is approximately symmetrical around 0, facilitating easier training of the neural networks.

The PDO is calculated from the deseasonalized SSTs, defined as the leading empirical orthogonal function (EOF) of the North Pacific (110E-260E, 20N-60N) monthly SSTs. This EOF, projected onto the global deseasonalized SST field, is presented in Figure 2.1a. In contrast to previous studies where the PDO index is defined using low pass filters with between 5–11 year cut-offs, here the PDO index is defined as the 6-month running mean of the principal component time series corresponding to the EOF patter. This is because PDO transitions are considered to be influenced by interannual variability associated with, e.g., ENSO [48, 1] and we want our ANNs to be able to account for these processes. The distribution of phase durations in CESM2 is shown in Figure 2.1b, demonstrating that there are a large number of phases of shorter duration, with decreasing samples as phase duration increases. The PDO representation in CESM2 is considerably improved over previous versions of the model, with periods of long term persistence similar to the observational record. However, the PDO within CESM2 contains extended periods of rapid fluctuation [57]. We choose to retain and investigate these periods because the observational record is relatively short, and furthermore it has been posited the PDO will become weaker and of shorter phase under climate change [59], hence high frequency PDO variability may become more relevant in future climate scenarios.

2.2.2 Artificial Neural Network

We use an artificial neural network (ANN) with a single hidden layer to predict whether a PDO phase transition will occur within 30 months, i.e. for some input, the output is a classification (yes or no) of whether a PDO transition will occur within the following

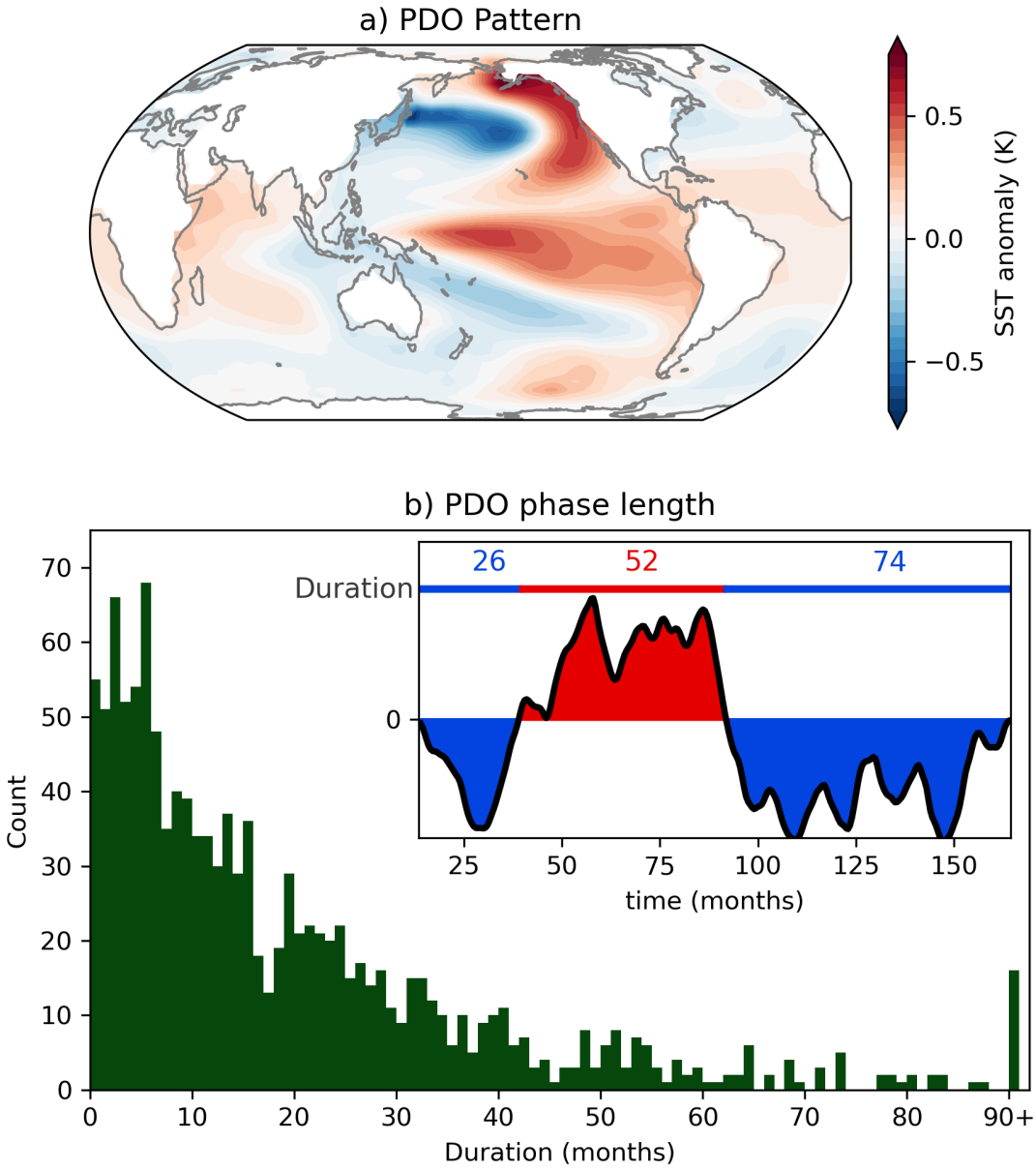

Figure 2.1: a) North Pacific PC 1 projected onto global de-seasoned SST. b) Histogram showing distribution of PDO phase lengths in CESM pre-industrial control run. Inset: slice of PDO index showing PDO phase length as number of months between phase changes.

30 months. An overview of neural networks is provided in Appendix A as well as our rationale for using a 30 month lead time in this study. The input layer to the ANN is three maps of deseasoned and standardized $4^\circ \times 4^\circ$ OHC anomalies, four months apart i.e. if the ANN is predicting PDO transition occurrence within some month $\tau = 0$, the three input maps are $\tau = -38$, $\tau = -34$, and $\tau = -30$ months. The input fields are flattened and concatenated resulting in an input vector of 12150 pixels. The input vector is fed into a densely connected hidden layer with 8 nodes which utilize the Rectified Linear Unit (ReLU) activation function. Finally, this is fed into an output layer of two nodes with softmax activation, representing the prediction. We interpret the ANN's prediction as the node with the higher value, and this value is termed the "ANN's confidence". For example, if the output is 0.63 on the persistence node, and 0.37 on the transition node, this represents a prediction of persistence with 0.63 confidence. For training, we use the categorical cross entropy loss function. We have found that setting the problem up as a binary classification task – will it or will it not transition in the next 30 months – yields insights into the mechanisms for PDO transition predictability. With that said, we have explored other architectures as well, including setting the problem up as a regression task whereby the network must predict the number of months until the next transition. In this instance, the network struggles to differentiate weak PDO states that may flip sign in the coming months from those weak PDO states that are on their way to persist for years. Since the main goal of this work is to identify mechanisms that offer PDO transition predictability, we present results from the binary classification architecture here although the regression architecture warrants further exploration.

We split the data into training and validation, using the first 90% (1800 years, 21600 samples) for training and final 10% (200 years, 2400 samples) for validation. Since there are more samples where transitions occur than persistence (see Figure 2.1b, there are more short duration phases than long), we manually balance the classes in both the training and validation sets. To generate the training data we use all of the persistence samples in the

training set, and randomly grab an equal number of transition samples from the training set. We do the same from the validation set. This results in 9386 training samples (4693 of each class) and 1110 validation samples (555 of each class) for each neural network. We train 60 networks total with identical architecture and vary only the random seed which controls how the weights in each network are initialized, and the random selection of transition samples. Here we present results as averages from the best 3 networks. Full model specifications, descriptions and analysis of all 60 networks is included in Table S1 and Appendix A. After training, we use the ANN to make predictions of both training and validation data. As we are able to rank an ANN's output by confidence, when presenting results as composites we choose to discard the 50% least confident predictions. Since the network is less confident about these predictions, removing them from our analysis suggests our results will focus on those with the strongest signals.

To investigate the decisions made by the ANNs, we use the neural network attribution technique called layer-wise relevance propagation [LRP; 60]. LRP propagates the prediction from an ANN back through the network and provides in our case, a map of relevance values corresponding to the input grid, with positive values indicating points that were relevant to the specific prediction, and negative values indicating points that detracted from the prediction. The higher the value, the more "relevant" the grid point. The utility of LRP in climate predictability studies has been discussed by Toms et al. [32], Mamalakis et al. [61] and used in studies by e.g. Mayer and Barnes [21], Toms et al. [31], Sonnewald and Lguensat [35]. Here, we present composites of LRP maps for predictions when the network is correct and confident. Each relevance map is first normalized by the prediction confidence (i.e. LRP map is divided by the winning confidence) before compositing, then the composite map is scaled by its maximum absolute value so that the composite map has a maximum absolute relevance value of 1.

2.3 Results

2.3.1 Detecting Persistence

The average total accuracy of the best three ANNs is 65%, with average conditional accuracy for predicting persistence of 55% (given no transition occurs, the ANN correctly predicts no transition). While this accuracy is above that expected by random chance, the low conditional accuracy across all persistence samples is likely due to the set up of this problem. Consider a sample that transitions 31 months after input; this sample would be designated persistence. However, a sample that transitions 29 months after input would be classified as a transition, despite the similarity of the input samples. Because of this, the samples that persist just beyond 30 months have very low accuracy while those with much longer phase duration (potentially more indicative of long-term PDO persistence) are more rare but have higher prediction accuracy (62% for durations > 40 months). This is demonstrated in Figure 2.2. In panel a we show the average distribution of phase duration (green line) with the blue line demonstrating the number of samples correctly identified by the ANN in the validation data. The increase of samples at month 30 is due to our method of balancing the number of samples per class for our neural network inputs. Recall that the number of samples in the transition class (area under green curve for durations 0-30 of months) is equal to the number of samples in the persistence class (area under green curve for durations of 30+ months), and to achieve this we sub-sampled the transition samples while maintaining all persistence samples. The sub-sampling maintains the shape of the distribution of phase duration in the transition class but reduces its size, resulting in a jump in the number of samples at phase duration > 30 months. Panel b shows the accuracy as a function of phase duration (i.e., blue divided by green). For example, when a transition occurs 10 months after input, (i.e., duration of 10 months on the horizontal axis), the ANNs are correct and predict a transition around 75% of the time. Similarly, when a transition occurs 60 months after input (i.e., the correct prediction is that no transition occurs within 30 months), the ANNs are correct around 90% of the time. To compare the

results to random chance, the dashed line indicates accuracy of 0.5, with shading indicating the 5th-95th percentile range for each phase duration bin. For samples around the cut off of 30 months, there is a dramatic drop in accuracy. However, as duration increases so does prediction accuracy with high accuracy for samples between 45 and 65 months. Note for samples of duration above 70 months accuracy is again very low. We propose that this is because these samples will occur early in a PDO phase (i.e., very soon after a transition) and hence having a weak PDO pattern for the ANNs to discern. It is hence difficult for the ANN to differentiate between these samples and those where the sign flips very soon after input. We hence propose that the ANNs have learned patterns relating to persistence especially for samples where the phase is of longer duration. We also consider the accuracy of the predictions with the top 50% confidence values, shown in the dashed red line in Fig. 2.2. This shows that predictions with higher confidence are more likely to also be accurate, especially for the regime we consider here (transitions that occur in 12-27 months). As higher confidence corresponds to higher accuracy, this implies that our networks have learned when patterns are more likely to lead to predictability.

Figure 2.3 shows the composite maps for correct predictions for cases when the PDO persists in its positive phase. The LRP heatmap of relevance values calculated for month $\tau = -30$ (the last input month) are shown in Figure 2.3a, while Figures 2.3b and 2.3c display the standardized OHC anomaly at the input month ($\tau = -30$) and the final month ($\tau = 0$). OHC anomalies at both the input time and the prediction show a positive PDO pattern in the North Pacific, with the horse-shoe shaped positive anomalies surrounding negative anomalies, verifying that indeed the ANNs have predicted a persisting pattern. Furthermore, the large magnitude anomalies in the North Pacific at input (Fig. 2.3b) are suggestive of PDO persistence as they correspond to a high magnitude PDO index which takes time to decay. It is thus encouraging that the largest relevance values in the LRP heatmap in Fig. 2.3a align with the positive horse-shoe shape in 2.3b. This suggests that the ANNs recognize large positive OHC anomalies in the North Pacific ocean as being an

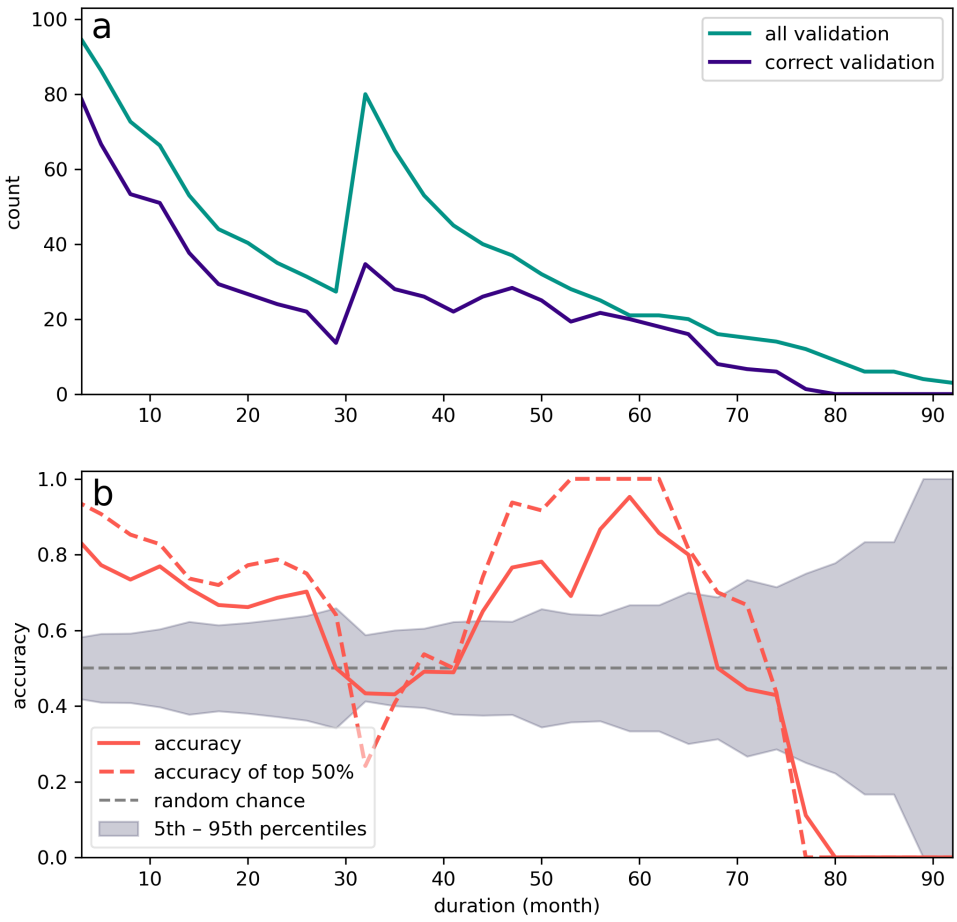

Figure 2.2: a: Average distribution of phase duration in the validation data for the three ANNs, green shows all the validation data and blue is number correctly predicted by the ANN with data binned into 3 month averages. b: Red line is accuracy of each phase duration bin (blue divided by green from above), red dashed line is accuracy of each phase duration when we only consider samples with highest 50% confidence. Grey dashed line indicates accuracy of 0.5, or random chance, with shading indicated 5th–95th percentile range for random chance.

indicator that the PDO will persist on the interannual timescale, and this is consistent with our physical understanding.

2.3.2 Detecting Transitions

We now consider the ANNs's ability to predict PDO transitions within CESM2. The average conditional accuracy for predicting a transition (i.e., given a transition occurs, the ANN predicts a transition) is 74%. The conditional accuracy of transitions 12-27 months after input (given a transition occurs 12-27 months after input, the ANN predicts the transition) is 69%. This is apparent in Figure 2.2b, with high accuracy for transitions that occur very soon after input (duration of 0-12 months on the horizontal axis) with reduced accuracy for transitions that occur in the 12-27 month window (duration of 12-27 months on the horizontal axis). These later transitions are hence more difficult for the ANNs to learn because they must learn to detect precursors of transitions more than 12 months before it occurs. Up to 27 months, accuracy values fall on or above the 95th percentile of random chance. This suggests that when correct, the ANNs have learned patterns that lead to PDO transitions and furthermore, that they can recognize them more than 12 months in advance.

Figure 2.4 shows the composite result for correct prediction of PDO transitions when the transition occurs 12-27 months after input. We choose this window because it means the ANNs must recognize patterns that signal transitions at least 12 months in advance while there is no loss in accuracy due to the 30 month cutoff. Positive to negative transitions are displayed in the left column and negative to positive transitions are displayed in the right column. Figures 2.4a and 2.4b are the LRP maps for the final input map (month $\tau = -30$) with Figure 2.4c and 2.4d the corresponding OHC. We highlight the strongest relevance regions from the LRP maps by superimposing LRP contours (Fig. 2.4a and 2.4b) onto the OHC (Fig. 2.4c and 2.4d), with solid lines contours outlining highest 5% relevance values. Similarly, dashed contours encircle regions with the lowest 5% relevance values.

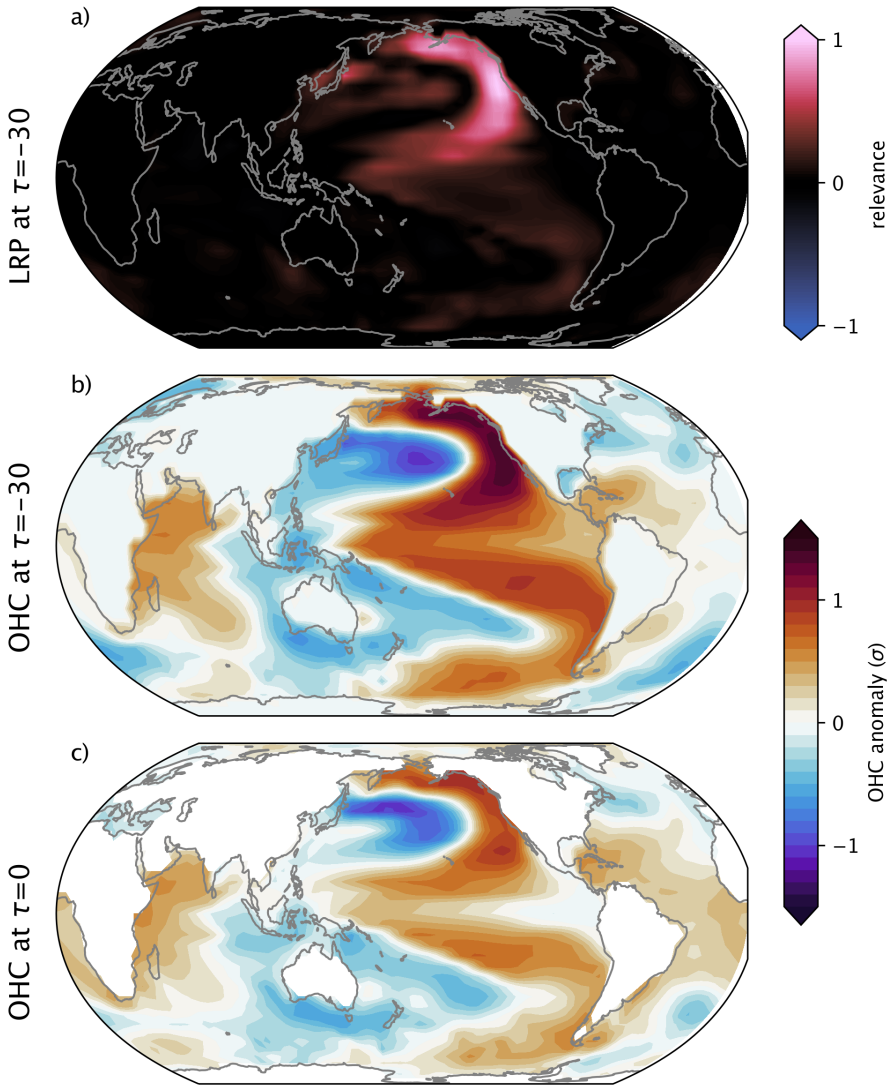

Figure 2.3: Composite maps when ANN correctly and confidently predicts persistence. a) Composite mean of LRP maps at final input month (τ=-30). Red areas correspond to positive relevance and blue to negative relevance. b) Composite mean of OHC input maps at τ=-30. Color scale is OHC anomaly in units of standard deviation σ at each grid-point. c) Composite mean of OHC at predicted month, color scale as in b).

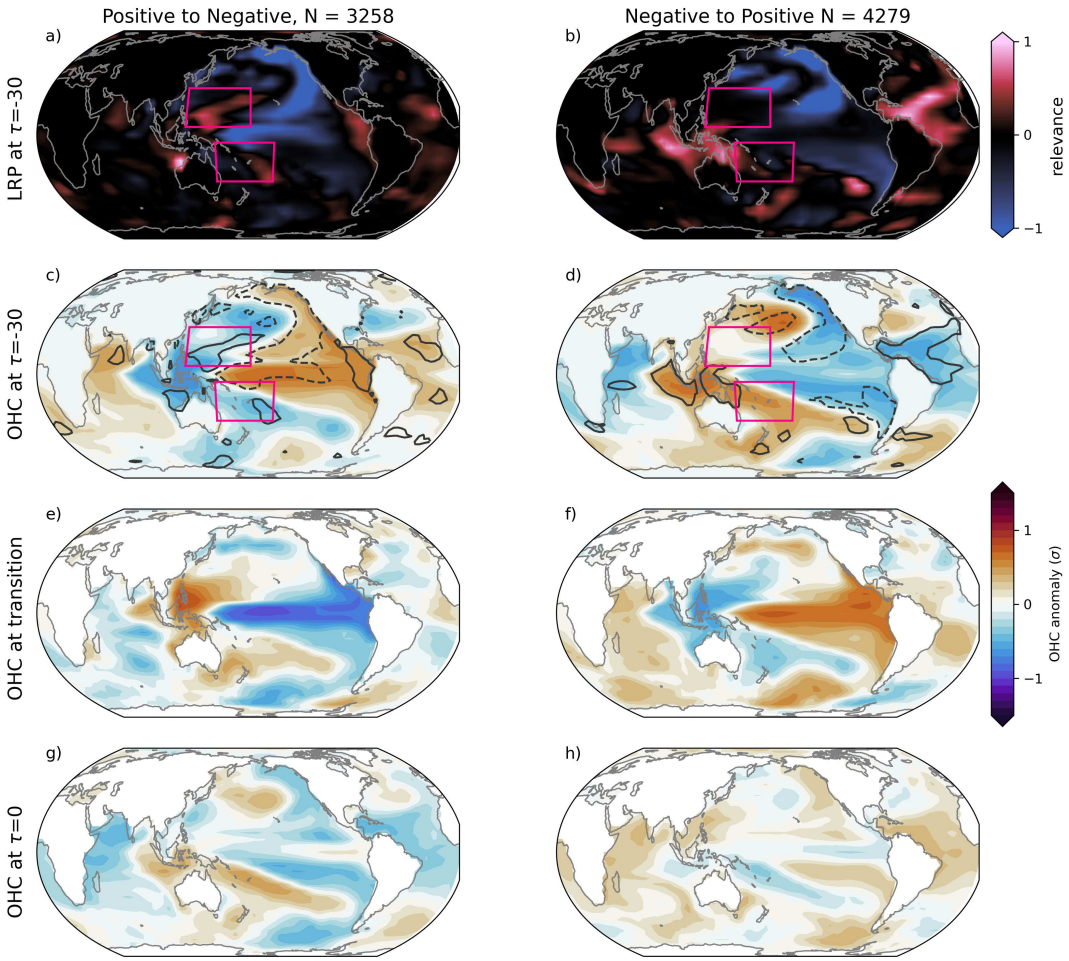

Figure 2.4: Composite maps of correct and confident predictions of PDO transition when transition occurs 12-27 months after input. Left column is positive to negative transitions, and right column is negative to positive transitions. Number of samples in each column is included in the title. Panels a) and b) are composite LRP 30 months before predictions. Red regions correspond to highest relevance and blue to lowest. Pink boxes highlight regions where OHC build-up is considered to precede PDO transitions (125E-180E, 5N-30N, and 150E-200E, 5S-30S). Panels c) and d) are the composite OHC maps 30 months before prediction, with color scale OHC anomaly in units of standard deviation. Dashed contours in c) and d) correspond to regions with highest 5% relevance in a) and b) respectively with dotted contour the lowest 5%. Panels e) and f) show composite OHC when transition occurs and panels g) and h) show OHC at the predicted month.

Furthermore, we include pink squares in Fig. 2.4a–d to emphasize the regions where a build-up of OHC has been suggested in the literature to precede a PDO transition [50]. Lastly, to track the OHC evolution throughout the transition process, panels 2.4e and 2.4f show the OHC when the transition occurs, and 2.4g and 2.4h the OHC at month $\tau = 0$. Note in Figure A.3-A.4 we show the LRP maps and associated OHC for each input grid ($\tau = -38$, $\tau = -34$ and $\tau = -30$) but we do not include them here as they are very similar but with lower relevance values.

Large negative anomalies in the northern and southern off-equatorial western Pacific precede the positive to negative PDO transitions (Fig. 2.4c), while large positive anomalies precede negative to positive transitions in the southern off-equatorial western Pacific (Fig. 2.4d). Together, these suggest the presence of a build up of OHC in either the northern or southern off-equatorial Pacific at least 12-27 months before a PDO transition occurs. In conjunction with the anomalies in Fig. 2.4c, the ANNs have recognized the northern region of heat content build up, with high relevance in the LRP composite in Fig. 2.4a. Conversely for negative to positive transitions, the ANNs mostly focus on the large positive anomalies over the maritime continent as well as the negative anomalies in the Atlantic, as shown by the high relevance values in Fig. 2.4b. The large relevance values in the Atlantic could signify the ANN detecting Atlantic teleconnections driving PDO transitions, which we discussion further in section 4. We also speculate that the lack of high relevance in the specific regions previously posited to contain anomalies leading to transitions [pink boxes in Fig. 2.4b 50] could be due to a westward shift of these anomalies in CESM2 leading to the high relevance values in the maritime continent. Conversely, the larger number of samples in Fig. 2.4b compared to positive to negative transitions (N = 4279 for negative to positive compared to N = 3258 for positive to negative), results in weaker relevance signals. In Figure A.5 we show by k-means clustering the LRP maps that there are indeed several distinct patterns within the LRP composite likely corresponding to different transition regimes detected by the ANNs, and cluster three of Fig. A.6 (middle column) shows

high relevance corresponding to the off-equatorial western Pacific for negative to positive transitions. So there appear to be different OHC patterns leading to PDO predictability. Furthermore the regions of high relevance in the composite in Fig. 2.4b suggest that the ANNs are using the OHC anomalies in these regions for its correct predictions, hence, we suggest future investigation into how these OHC anomaly patterns may preempt PDO transitions. Furthermore, the ANNs appear to be better at predicting negative to positive transitions than positive to negative transitions as there are more correct samples in the latter category (note there approximately the same number of transitions in each category). It is unclear whether this is due to PDO representation in CESM2, or whether there are fundamental differences in the transition process.

At the month the PDO transition occurs, note the large equatorial anomalies via La Nina and El Nino (Fig. 2.4e and 2.4f respectively). Furthermore, the anomalies in the western off-equatorial Pacific have switched sign in each panel at the transition as well. These factors are consistent with the mechanism posited by e.g. Meehl et al. [50], that an ENSO event following the OHC build-up causes the OHC to be redistributed by equatorial Kelvin waves. This redistribution of heat, and the associated atmospheric teleconnections, effect a PDO transition. Lastly, after the transition occurs (Fig. 2.4g and 2.4h), OHC anomalies have largely shifted into the opposite PDO phase pattern as we would expect.

The evolution of OHC throughout the PDO transition and corresponding LRP heatmaps suggest that not only are PDO transitions preceded by OHC build-up in the off-equatorial western Pacific 12-27 months before the transition, but for positive to negative transitions, our ANNs detect this heat build up as relevant to its predictions. Furthermore, we suggest that this is also the case for negative to positive transitions but it is likely that regimes where this is detected by the ANNs are averaged out in the composite (Fig. A.6). Conversely, there are other signals detected in the relevance maps (Figs. 2.4a and 2.4b), and in addition the OHC anomalies are not consistently strong in the off-equatorial regions (Fig. 2.4d) which suggests that there are likely mechanisms other than that proposed by Meehl et al.

[50] that contribute to PDO transitions, or at least precondition the OHC build-up and following ENSO event. The ability of the ANNs to apparently detect a known precursor to PDO transitions supports their use in climate variability problems to identify and possibly discover regions leading to predictability.

2.4 Discussion and Conclusion

We show that PDO transitions are preceded by large amplitude OHC anomalies in either the northern or southern off-equatorial western Pacific 12-27 months before the transition occurs. Furthermore, using LRP we show that these anomalies are detected by the ANNs and were relevant to their correct predictions of positive to negative transitions. This finding is similar to the work of Meehl et al. [50] however in their analysis they suggest that OHC must build up in the off-equatorial western Pacific over a period of 10-15 years before a transition occurs. The transition predictions analyzed here only have inputs 12-27 months before the transition occurs, yet the ANNs do make correct predictions above random chance, implying that perhaps the timescale of the OHC build-up is less important than the fact that the anomaly is present. This is similar to the finding of Lu et al. [52] whose network analysis did not necessarily require OHC to build-up over a long period of time as long as it reached a certain threshold. Moreover, as we have applied 6 month smoothing, it is perhaps surprising that mechanisms contributing to PDO transition predictability were able to be detected by the ANNs. This suggests that the decadal scale of OHC build-up, and the interannual scale of ENSO interact cooperatively and hence filtering out shorter duration signals may hinder the detection of mechanisms relating to PDO transitions. This was also suggested by Lu et al. [52], who found their method less likely to detect their "early warning signal" when an 11-year low pass filter is applied. Note that if we only focus on transition predictions for long PDO phases, i.e. the PDO must persist for a minimum 2.5 years before and following a transition, our results are essentially unchanged (see

Figure A.7). We use 2.5 years here as a balance between sample size and long duration phases.

The maps in Figures 2.3 and 2.4 are presented as composite means of correct predictions. As we have suggested, the signals detected by LRP and presented in these figures may not necessarily be cooperating on every prediction. We check for this by using cluster analysis on the LRP composites in Figure 2.4. Figures A.5-A.6 show how k-means clustering highlights different signals in the LRP maps. Notably, the off-equatorial western Pacific is highlighted in at least one cluster for both positive-to-negative transitions and negative-to-positive transitions. Interestingly, there are regimes when the Atlantic Ocean seems to be a highly relevant region for predictability. Since Atlantic teleconnections are hypothesized to influence both PDO variability and ENSO events, and an ENSO event is considered to be required to trigger a PDO transition [62, 63, 64, 65], it is not unrealistic that Atlantic OHC signals could assist in predicting PDO transitions. In particular, teleconnections from the Atlantic are considered a key influence for triggering El Nino events [29] whereas La Nina events are thought to be largely triggered by a preceding El Nino event. In Figure 2.4b, the neural networks concentrate relevance in the Atlantic basin preceding the El Nino event (and PDO transition) in Figure 2.4f. Given this, it appears that the neural network recognizes the precursors of the El Nino event required for the transition during negative to positive transitions. This highlights the ANNs's ability to detect distinct mechanisms contributing to predictability.

We show how ANNs and interpretability techniques can aid in the discovery and investigation of mechanisms behind climate predictability. In the future, we suggest investigating regions highlighted here as potentially connected to PDO transitions, such as the Atlantic Ocean. This is especially important in examining the possibility of different pathways that can lead to PDO transitions and hence we support the continued use of methods such as ANNs and k-means clustering in objectively identifying potential regimes. In a broader sense, we encourage the future use of ANNs and XAI in climate predictability studies. We

have shown that they are not just a tool for maximizing prediction accuracy, but also as a way of investigating potential mechanisms that lead to predictability, and to advance our understanding of our chaotic climate system.

Chapter 3: Incorporating Uncertainty into a Regression Neural Network Enables Identification of Decadal State-Dependent Predictability in CESM2

3.1 Introduction

Predicting the evolution of the climate on decadal timescales (2-10 year) has far reaching implications for both climate science and society. On these timescales, changes in climate patterns are associated with the forced response to anthropogenic emissions and internal variability in ocean [1]. For example, the forced response from climate change can manifest as the steady increase of global mean temperature which provides some predictability of future temperatures. Decadal predictability of oceanic temperature variability arises from the ocean's ability to store, release and transport heat on decadal timescales. Major modes of variability in the Pacific and Atlantic Oceans are therefore linked to decadal predictability as they indicate the spatial distribution of heat in these basins. Furthermore, this internal variability in the ocean can act to either mask or amplify the forced response from climate change [66]. The Pacific Ocean exhibits long-term variability via the interdecadal Pacific oscillation [IPO, 67, 68] and its related mode Pacific decadal variability [PDV, 8, 9]. Atlantic multi-decadal variability [AMV, 11, 69] is considered the dominant form of long-term variability in the Atlantic ocean, however whether variability arises due to internal Earth system processes or external forcing is still under debate [70, 71, 72]. Because these patterns of variability are associated with decadal predictability, decadal prediction is traditionally focused on either investigating and predicting the processes themselves, [e.g. 50, 37, 12], or exploring the predictability that arises from the atmospheric teleconnections driven by these patterns [e.g. 73, 74, 75].

As hinted at above, it is difficult to decipher the drivers of predictability in observations and historical simulations as it is influenced by the non-linear interactions between internal

variability and external forcing. Studies have diagnosed predictability in pre-industrial control runs [76], while others have deciphered predictability from internal variability in model hindcast ensembles with accompanying unforced ensembles [14, 24]. Another avenue of research has been to quantify (using various metrics) how much predictability is present in different regions of the ocean, and what the relative contributions of internal and external drivers may be [77, 78]. However, predictability in the climate system can vary drastically depending on region, timescale, and initial state [23, 1, 19] thus studies have encouraged a shift of focus towards the concept of state-dependent predictability [79, 80, 18, 19]. This paradigm intrinsically acknowledges that some initial states lead to more predictable behavior than others. The aim is therefore to identify these more predictable initial states, as they provide the opportunity to make more skillful forecasts. State-dependent predictability has been investigated on short (subseasonal to seasonal) timescales as the identification of "forecasts of opportunity" or "windows of opportunaity" [20, 21]. An example of an oceanic region with decadal state-dependent predictability is the North Atlantic Subpolar Gyre. It has been found that anomalously strong ocean heat transport in the North Atlantic ocean is associated with skillful predictions of sea surface temperature (SST) in the North Atlantic Subpolar Gyre for lead times up to 8 years [25, 26]. So enhanced heat transport in the North Atlantic could be considered a more predictable initial state for predicting North Atlantic SSTs.

With this increased focus on state-dependent predictability, it is necessary to explore methods that can identify state-dependent predictability. Machine learning is one such method that shows promise for identifying more predictable initial states. In fact, on subseasonal timescales, classification artificial neural networks (ANNs) have been shown to objectively identify states of the Madden-Julian oscillation that lead to enhanced predictability of circulation in the North Atlantic [21] by leveraging the network's confidence in a prediction to identify state-dependent predictability. Furthermore, on decadal timescales

it has been demonstrated that ANNs can skillfully predict decadal processes [37, 36] and identify states of enhanced predictability of surface temperature over land [31].

This study introduces the identification of state-dependent predictability on decadal timescales using a regression-based neural network to predict sea surface temperatures (SSTs) across the globe within the Community Earth System Model, version 2 [CESM2; 55] pre-industrial control simulation. We demonstrate a powerful technique for incorporating uncertainty into the prediction of regression neural networks which has previously only been used a handful of times in climate science [81, 82, 83]. We further leverage this uncertainty output to identify which initial states are associated with the lower uncertainty predictions. Lower uncertainty predictions imply more predictable inputs, hence this technique identifies state-dependent predictability. Furthermore, we link predictable initial states to major forms of variability so we are able to identify certain combinations of IPO and AMV phases that correspond to skillful decadal predictions of SSTs in CESM2.

3.2 Data and Methods

3.2.1 Data

We use sea surface temperature (SST) and ocean heat content (OHC) output from the CESM2 pre-industrial control run for the Coupled Model Intercomparison Project phase 6 [CMIP6; 56]. OHC is interpolated to a $4^\circ \times 4^\circ$ grid. We train ANNs at each SST grid point so SST is interpolated to a $5^\circ \times 5^\circ$ grid which captures the regional variation in predictability while not being too computationally demanding. We use monthly output of the 2000 year run with the first 100 years removed to allow the ocean circulation to spin-up. Both OHC and SST are then de-seasonalized by removing the mean annual cycle from each grid point. Furthermore, to account for model drift, after deseasonalizing we calculate the third degree polynomial trend via least squares and subtract this from each grid point. Model drift is unphysical changes in the model's background climatology, and its removal mean that each variable's statistics are approximately stationary for the remaining 1900 years of data.

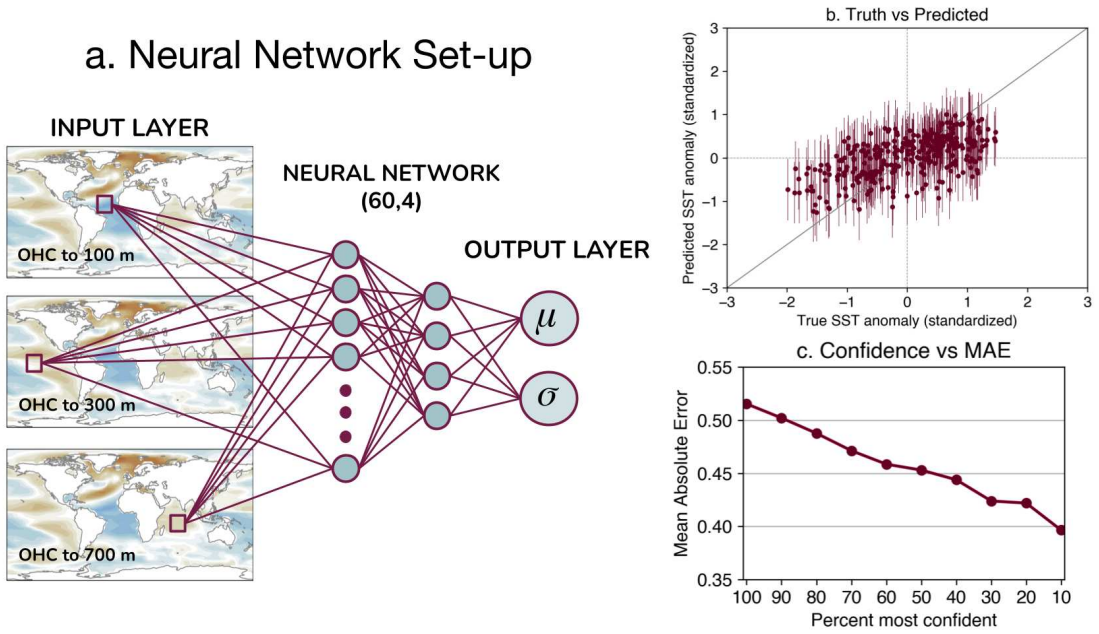

Figure 3.1: a. Schematic of the artificial neural network architecture. b. Scatter plot of predicted SST anomaly (y axis) vs true SST anomaly (x axis). Dots represent predicted μ values, while vertical lines represent the 1σ range. c. Prediction mean absolute error (MAE) as a function of prediction confidence (see text). Both b. and c. utilize the same network trained to predict SST in the North Atlantic Ocean (52.5°N, 325°E).

.

OHC is smoothed using a 60 month backward running mean to smooth high frequency variability. We divide the pre-processed data into training, validation and testing. The first 70% ($\sim$1300 years) is used for training, the next 15% ($\sim$300 years) for validation and the last 15% ($\sim$300 years) for testing. We calculate the mean and standard deviation for every point on both the OHC and SST grids in the training set. We then use these values to standardize all of the training, validation and testing data.

3.2.2 Artificial Neural Network

Artificial neural networks (ANNs) are used to predict the average SST anomaly at a lead time of 1-5 years and 3-7 years, i.e. the ANN predicts the average 60 month SST anomaly in the next 12-72 months, or 36-96 months respectively. In this experiment the ANN is trained to predict the SST evolution in the CESM2 pre-industrial control, so for

example, one input sample is OHC information from a specific time step in the control run, and the output prediction is the average SST anomaly over the next 12-72 months in the control run. A schematic of our neural network architecture is provided in Figure 3.1a and a brief overview of ANNs for geoscience applications can be found in e.g. Toms et al. [32]. The predictors are three OHC grids, where each grid is OHC integrated to a different depth (100 m, 300 m and 700 m). We chose varying depths of OHC because each contains information corresponding to different forms of climate variability. For example, the upper levels of the ocean integrate atmospheric forcing, and hence capture atmospheric variability as well as surface ocean dynamics [84]. The variability in lower levels of the ocean is guided by a combination of slow moving ocean circulation and the incorporation of mixed layer processes via the annual cycle in the thermocline [85]. By inputting three OHC depths into the neural network, it can theoretically combine different oceanic and atmospheric processes to make its predictions. The three ocean grids are vectorized with points over land removed resulting in a total 7947 input pixels. This input is connected to a hidden layer of 60 nodes which is then connected to another hidden layer of 4 nodes (see Fig.3.1). In this network, all layers are densely connected meaning all nodes in the previous layer are connected to all the nodes in the next layer. Furthermore, all nodes in the hidden layers use the rectified linear unit (ReLU) activation function. Finally this second layer is connected to the output layer of two nodes which serve as the parameters of the predicted conditional distribution (see details in the next paragraph). Here the distribution is a normal distribution as we found allowing skewness did not significantly improve the network's performance (not shown).

We use the $-\log(p)$ loss function described by e.g. Barnes et al. [86] which we will summarize briefly. For each input, the network outputs two values, μ and σ. To calculate loss, μ and σ are used to construct a conditional distribution, d and the negative log likelihood function is calculated at the true value (y_{true}), i.e. loss = -log($p(y_{true}|d)$). This means that the neural network can decrease loss (decrease -log($p(y_{true}|d)$)) in different

ways: either with a low σ value and μ that is close to y_{true}, or predict a larger σ value with μ that is further from y_{true}, or both. The neural network is therefore not penalized for high error predictions as long as it also guesses a correspondingly high σ value, that is, if it recognizes an input is less predictable by assigning a high σ value. The predictions of such an ANN are illustrated in Figure 3.1b, where we show an example scatter plot of prediction vs truth from an ANN trained to predict SST anomaly in the North Atlantic Subpolar Gyre. Note that we can plot both the predicted anomaly value (μ, colored dots) and an uncertainty range, with the error bars indicating the $\pm 1\sigma$ range predicted by the ANN. The ANN is trained using the training set, with the validation set evaluated at the end of each epoch. The results presented in this study are from the testing set. During training, we use a learning rate of 1×10^{-4} with stochastic gradient descent for up to 1000 epochs with early stopping when validation loss did not decrease for 100 epochs. To implement regularization, we include a dropout layer between the input layer and first hidden layer in training. We found that a high rate of dropout (80% dropout rate in this experiment) forced the ANN to learn information more slowly and greatly reduced over-fitting on the validation set.

3.2.3 AMV and IPO indices

We compute the AMV and IPO indices within CESM2 using the deseasoned and detrended SST data. For the AMV index, we calculate the monthly mean SST anomaly over the North Atlantic ocean ($0°$N to $80°$N, $280°$E to $360°$E) and then standardize by removing the mean and dividing by the standard deviation. Note we do not de-trend by the global mean SST as recommended by Trenberth and Shea [87] because the control run lacks a forced long term warming trend and model drift was removed during pre-processing. We calculate the IPO index following the tripole index proposed by Henley et al. [88]. We include plots of the spatial AMV and IPO patterns in CESM2 and the method for calculating IPO index in the Appendix B.

3.3 Results

3.3.1 Evaluating Performance

In this study, 10 networks (identical architecture, only varying the initial network random seed) are trained at each SST grid point in the ocean and we show the results of the best neural network at each grid point. To designate the "best" network, we select the ANN with the lowest mean absolute error (MAE, difference between predicted μ and true y) on the 10% of samples with the lowest σ predictions in the validation set. This designation leverages a fundamental characteristic of a network that has learned predictability in the data: prediction error should decrease as predicted σ decreases. We demonstrate this idea in Figure 3.1c where we show perfomance on the test set for a network trained to predict SST in 1-5 years in the North Atlantic (52.5°N, 325°E). Along the x-axis, we threshold by increasing confidence with the y-axis showing corresponding MAE for those predictions. For all samples, the MAE is $\sim$0.52 however for the 40% most confident predictions the MAE has dropped to 0.46. For the 10% most confident predictions, the MAE has dropped further to $\sim$0.39 implying the network has learned samples that lead to more predictable SST anomaly. We hence refer to lower σ predictions as more confident predictions, or more predictable inputs. For some grid points, all networks fail to learn anything, meaning they always predict an SST anomaly of zero (or very close to zero). The networks at these grid points are removed before analysis, resulting in 30% of networks (525/1709) removed for lead years 1-5, and 39% (675/1709) for lead years 3-7.

3.3.2 Predicting SST

We ensure that the ANNs are learning to skillfully predict SSTs on decadal timescales in CESM2 by examining prediction error in the testing data at each grid point. Fig. 3.2a is the MAE for ANN predictions for the testing set for lead years 1-5, with black indicating grid points where all 10 networks failed to learn anything. These regions are largely in the Southern Hemisphere subtropics. The lowest MAEs are found in the North Atlantic Ocean

and the Southern Ocean around South America. This spatial distribution of prediction skill (including regions where the networks failed) broadly agrees with that found to be attributable to internal variability in the decadal hindcast studies using the CESM1 decadal prediction large ensemble [14, 23]. These studies use a different model version (CESM1 vs CESM2), and the simulations include the effects of external forcing since 1850. However, the widespread agreement of spatially varying predictability suggests the results in Figure 3.2 are not a result of experiment design or network architecture but are rather due to differences in predictability between regions.

The prediction skill for lead years 3-7 is shown in Fig 3.2b and highlights similar regions as being more predictable as in lead years 1-5. Furthermore, there does not seem to be a substantial loss in skill between these two lead times. This, coupled with the spatial distribution of prediction skill, suggests that the ANNs are learning physical relationships to make their predictions.

To contextualize the predictions of the ANNs, we benchmark them against a simple persistence model. The persistence model predicts that the SST anomaly will be unchanged so that the SST anomaly at the time of input remains the same at the time of prediction. We calculate the MAE for the persistence model and subtract it from the MAE of the ANNs ($\Delta\text{MAE} = \text{MAE}_{ANN} - \text{MAE}_{persistence}$), and plot the results in Figure 3.2g and 3.2h. In regions where ΔMAE is negative, the ANN outperforms persistence (i.e. has lower error). These regions are illustrated in warm colors in Figure 3.2g and 3.2h and illustrates that the ANNs trained in this study out-perform persistence in all locations and at both lead times. These regions were all found to be significant to $\alpha = 0.05$ using a one-sided Wilcoxon signed-rank test. The greatest improvement in skill above persistence occurs in the cold tongue region of the Equatorial Pacific. This is unsurprising as this region exhibits large interannual variability due to the El Nino Southern Oscillation, and hence persistence performs poorly in this region. Also notable, the improvements over persistence do not necessarily align with grid points where the networks achieve lowest MAE. This

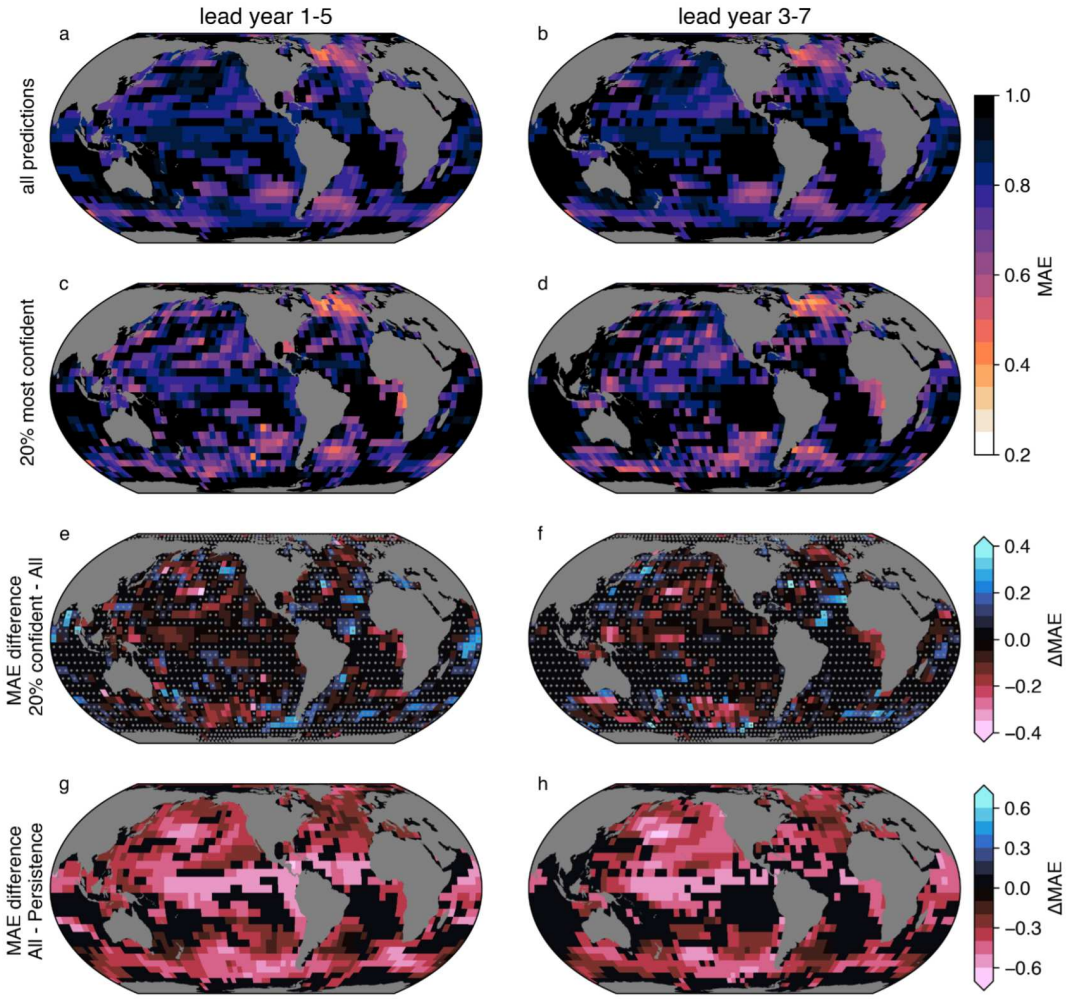

Figure 3.2: Evaluation of ANN prediction error. The left column is the prediction error for lead years 1-5, and the right column is for lead years 3-7. Panel a and panel b are mean absolute error (MAE) for all predictions in the testing set (i.e. all samples, N=3400). Panel c and panel d show MAE for only the 20% most confident predictions in the testing set as identified using the ANNs's uncertainty (N=680). Panel e and panel f are the differences between the 20% most confident predictions and all predictions (e.g. panel e = panel c – panel a). Stippling indicates areas where the skill improvement is not statistically significant to $\alpha = 0.05$. Panel f and panel g are the difference between MAE_{ANN} and $\text{MAE}_{persistence}$ ($\text{MAE}_{ANN} - \text{MAE}_{persistence}$) in the testing set.

is a fingerprint of regional decadal variability, that regions with longer memory (e.g. the mid-latitude North Atlantic) are better modeled by persistence, but in these cases our networks still out-perform persistence.

3.3.3 Identifying State-Dependent Predictability

The predictive power of ANNs for decadal prediction is now demonstrated by using them to identify state-dependent predictability. In Figure 3.2c and 3.2d we plot the MAE for only the 20% most confident predictions (20% lowest predicted σ) by the ANN for each SST grid point. That is, ANN objectively identifies more predictable initial states, and we do not directly use knowledge of the ground truth to identify these predictions. To aid in visualization, we also plot the difference in MAE between the 20% confident predictions and all predictions in Figure 3.2e. When comparing the most confident predictions with all predictions at lead years 1-5 (Figure 3.2e), MAE is largely reduced for more confident predictions in the mid-latitudes, implying that more confident predictions are associated with smaller prediction errors in these locations. Similarly for lead year 3-7 (Fig. 3.2f), we see that sorting for the most confident predictions leads to reduced error in most locations. For those regions where error increases, this is likely due to the network learning predictability in the testing and validation data that does not generalize to the testing data which either suggests over-fitting or unaccounted-for model drift. Interestingly, at both lead times, some regions that show very little skill across all predictions exhibit large increases in skill when considering only the most confident predictions (e.g. central Pacific and the Gulf of Guinea), demonstrating that a region may be considered not predictable when in fact it is just not *always* predictable.

3.3.4 Investigating Skillful Decadal Predictions

By using ANN predictions to identify state dependent predictability, we can also investigate oceanic patterns that lead to predictability. Here we examine the predictions of two ANNs trained to predict SSTs in the North Atlantic and North Pacific oceans to investi-

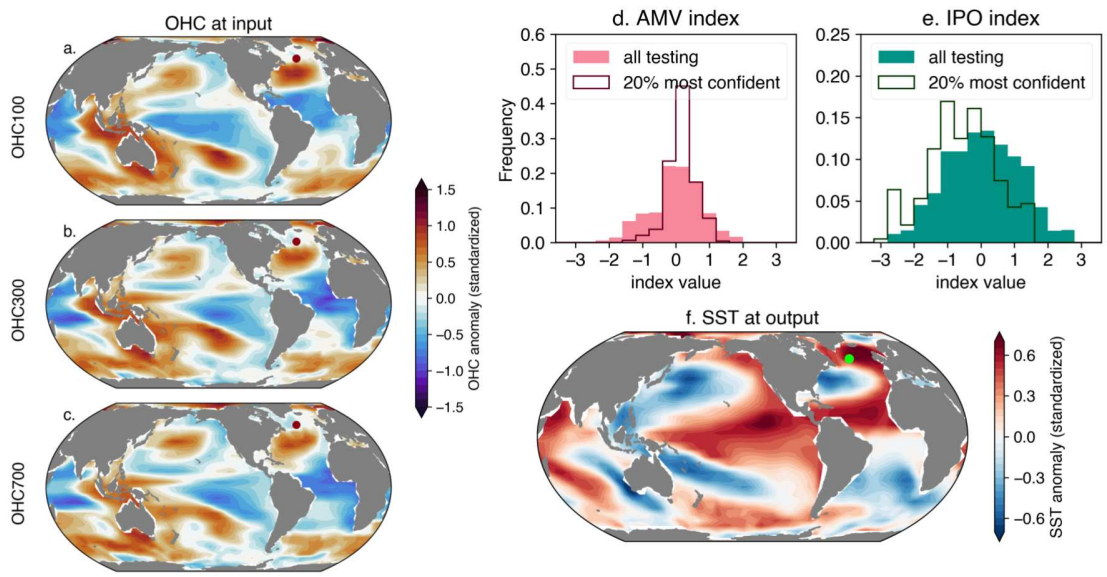

Figure 3.3: State-dependent predictability identified in the North Atlantic for predicting average SST anomaly at lead time 1-5 years. Panels a-c: Composite of OHC inputs for confident predictions of positive SST anomaly in a point in the North Atlantic (red dot). Panel d: histogram of AMV index for testing data (pink shading) and most confident predictions (pink outline). Panel e: as panel d but for IPO index. Panel f: Composite of SST map for confident predictions of SST in the North Atlantic (green dot).

gate processes that are contributing to enhanced prediction skill in these regions. In the following analysis we single out two particular grid points to investigate SST predictability but the results are largely unchanged for the directly adjacent grid cells. Here, we show results for the testing data but these results are consistent throughout the control run (see Appendix B).

Figure 3.3 shows the 20% most confident predictions of positive SST anomaly for a point in the North Atlantic Sub-Polar Gyre from the testing set (52.5°N, 325°E). We single out positive predictions because the ANN's confident predictions are preferentially positive (583 positive predictions out of 680 confident testing samples, where 680 is 20% of the testing set), implying that the ANN detects that particular positive predictions lead to lower uncertainty. As predictions are preferentially positive, this is evidence that the ANN is detecting state-dependent predictability in the North Atlantic.

We plot the correct and confident positive predictions to ensure we are analyzing the correct signals that contribute to predictability. This leaves 472 samples. Fig 3.3a – 3.3c show the composite of OHC input maps for correct and confident positive predictions to investigate the initial states that lead to predictability. At all three OHC levels there is a positive OHC anomaly in the subtropical to mid-latitude Atlantic Ocean. We verify that this signal was likely utilized by the ANN in its predictions by using an ANN explainability technique to investigate the input regions that are important to the network's prediction (see Section B.3 and Figure B.4). This shows the positive OHC anomaly in the North Atlantic at all three OHC levels was highlighted as contributing to the ANN's decisions. As the positive heat anomaly is slightly south of the predicted grid point, this could indicate northward heat transport to achieve a positive prediction. The composite SST anomaly in Fig 3.3f shows the positive anomaly is around the predicted grid point in the North Atlantic which implies that this anomaly has moved northward from the initial state (i.e. northward from the positive OHC anomaly in the subtropical North Atlantic in Fig 3.3a). From this evidence, we posit that the skillful SST prediction is preceded by a positive heat anomaly in North Atlantic ocean, which is transported into the gyre region. This is consistent with Borchert et al. [26] who identified periods of enhanced heat transport in the mid-latitude as a state of increased predictability of SSTs in the North Atlantic subpolar gyre for up to 8 years.

As an analogue for oceanic variability, we also consider the phases of the AMV and IPO during periods of increased network confidence. In Fig 3.3d we present the distribution of the AMV index during the entire testing period (pink shading, mean = 0.00) with the solid line showing the distribution for only 20% confident predictions which has a mean of 0.16. From this, it appears that confident predictions are most likely to occur during positive AMV. When randomly drawing 20% of the samples from the AMV distribution in testing, the likelihood of a mean of 0.16 occurring is less than 1%. This implies that more skillful SST predictions in the North Atlantic Sub-Polar Gyre coincide with northward

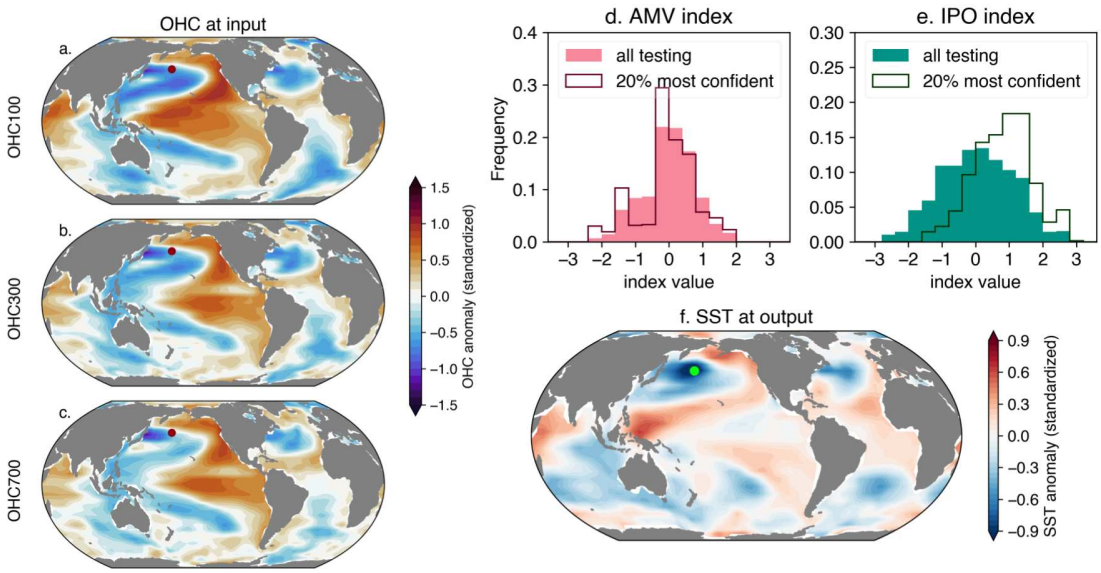

Figure 3.4: As Figure 3.3 but for the North Pacific

heat transport from the subtropics (from 3.3a-c and f) coupled with the positive phase of AMV (from 3.3d). This is consistent with previous results by e.g. Christensen et al. [23], Borchert et al. [26]. In 3.3e, we show the distribution of IPO phase for the testing data (green shading, mean = 0.05) and 20% most confident predictions outlined with the solid line, with a mean of -0.58. The likelihood drawing a mean of -0.58 from the IPO testing distribution is less than 1% which suggests that the negative phase of the IPO contributes to the predictability of North Atlantic SSTs. This is also apparent in Fig 3.3a-c which all show the negative IPO pattern in the Pacific Ocean. This may indicate some inter-basin teleconnection that contributes to the predictability of North Atlantic SSTs.

We now perform a similar analysis for an ANN trained to predict SST in 1-5 years at a point in the North Pacific (42.5°N, 175°E). In Figure 3.4 we show the results for the 20% most confident negative predictions. For this region, 632 out of the 680 most confident samples were predictions of negative anomaly, implying the ANN designated negative predictions as more confident. Again we plot only the correct predictions, resulting in 466 samples in these composites. Fig 3.4a-c shows the composite OHC inputs for confident negative

predictions, and the major signal appears to be a positive IPO/PDV pattern in all panels. It is likely the ANN utilized this pattern to make these confident negative predictions from the ANN explainability heat-maps (see section B.3 and Figure B.5). This is supported by the histogram of the IPO index in Fig 3.4e which shows the distribution of IPO phase in the confident samples is shifted such that confident samples significantly coincide with the positive phase of the IPO. There is no such strong signal in the AMV index (Fig 3.4d). Lastly, the confident predictions appear to relate to persistence in the positive IPO phase because the composite map of SST at output (Fig 3.4f) shows an IPO pattern in the Pacific Ocean. The largest SST anomalies are in the north Pacific mid-latitudes, in the traditional PDV region. From this, we posit that skillful predictions of SST in the North Pacific are associated with persistence in the positive phase of IPO (i.e. negative SST anomaly at the predicted grid point). Here, the ANN preferentially identifies negative SST predictions as skillful, perhaps implying that persistence in the positive phase of IPO is more predictable than persistence of the negative phase. We posit that this difference in predictability is due to the underlying non-linear mechanisms governing IPO dynamics and particularly the asymmetry in the dynamics governing ENSO events [89, 90]. Further investigation of this is an avenue for future work.

3.4 Discussion & Conclusion

We show that artificial neural networks (ANNs) skillfully predict SST evolution on decadal timescales and that they can objectively identify decadal state-dependent predictability due to internal variability in the North Pacific and North Atlantic Oceans. Specifically, we use a regression neural network where the predictions take the form of a conditional normal distribution which we leverage to isolate predictions that are more likely to have lower error. This approach allows us to investigate possible contributing mechanisms to decadal SST predictability, particularly Atlantic multi-decadal variability and the interdecadal Pacific oscillation (AMV and IPO, Figs 3.3 and 3.4). We chose to model

the conditional distributions as normal distributions as alternatives did not significantly improve skill. We suggest that future studies investigating state-dependent predictability for other timescales and variables may benefit from the addition of skewness to the predicted conditional distributions [86], as well as further exploring alternative network architectures to tease out additional skill.

We investigate state-dependent predictability in two regions, the North Atlantic Subpolar Gyre, and the North Pacific Ocean by identifying predictions in these regions that the ANNs assigned the lowest uncertainty and investigating the processes that correspond to these confident predictions. This study utilizes the CESM2 long control representation of the climate system and the results in the North Atlantic appear to agree with hindcast studies of Brune et al. [25], Borchert et al. [26], Yeager et al. [14] which use different models to that used here (MPI-ESM; Giorgetta et al. [91] and CESM1; Hurrell et al. [92]). These previous studies also incorporate observations or reanalysis to evaluate the prediction skill of the decadal hindcasts. Moreover, in a study of initialized decadal hindcasts in the CMIP6 archive, Borchert et al. [24] attribute predictable SSTs in the North Atlantic Subpolar gyre to the effects of external forcing in the historical era, particularly volcanic forcing. Since our findings are consistent with the state-dependent predictability investigated in these studies, this suggests that the ANN predictions and mechanisms investigated here are likely relevant to realistic climate variability and implies a role for internal variability in North Atlantic predictability. Further investigation is left for future work.

Here we present a data-driven approach to diagnosing state-dependent predictability in an unforced model simulation. In addition to the role of North Atlantic heat transport, we find evidence for a state-dependent inter-basin teleconnection, that is, the negative phase of the IPO influencing predictability of North Atlantic SSTs (Fig 3.3). The drivers of predictability and variability in the North Atlantic ocean are still debated, especially the relative roles of internal variability and external forcing [93, 70, 12, 71, 94, 95]. We hence

suggest that future work on decadal prediction should investigate the roles of internal variability and external forcing through the lens of state-dependent predictability.

This study emphasizes the importance of examining state-dependent predictability for decadal predictions. We stress that the *a priori* identification of more predictable initial states greatly increases prediction skill and can hence aid in estimating the evolution of the internal long-term variability of the climate system.

Chapter 4: Separating Internal and Forced Contributions to Near Term SST Predictability in the CESM2-LE

4.1 Introduction

Skillfully predicting both global and regional climate change on interannual to decadal timescales is an outstanding problem from both a scientific and societal standpoint [2]. Trustworthy forecasts on this timescale can provide actionable information about the future climate for various sectors affected by climate change [4, 96, 3]. However, making skillful forecasts on decadal timescales is a challenge because it requires both predicting the climate response to anthropogenic forcings like greenhouse gases and aerosols, as well as skillfully forecasting low frequency internal climate variability [1]. For example, it is has been argued that modulations in the rate of global mean surface temperature increase are associated with changes in the internal low frequency variability in the Pacific ocean so that even in the global mean, skillful temperature predictions require comprehensive understanding about internal variability [66, 36]. On regional scales, the amplitude of internal variability can be much larger relative to the forced response than on global scales, providing an even larger source of uncertainty in the range of variability [97, 98]. Identifying and understanding predictable internal variability on regional scales is therefore an opportunity to provide stakeholders with improved estimations of the future range of variability.

Internal variability on interannual to decadal timescales can be predictable both in model simulations of the pre-industrial climate [76, 37, 38], and in hindcast simulations of the historical era [50, 99, 17]. It has been suggested that future, global scale variability may too be predictable for certain initial states [36]. This predictability is generally associated with low frequency modes of variability, namely the Pacific decadal oscillation [PDO; 8, 10] and Atlantic multidecadal variability [AMV; 11]. However, prediction skill associated with internal variability is often sparse and limited to ocean heat predictability in mid-latitude

ocean basins [14]. Even with these limitations, it has been demonstrated that skillful sea surface temperature (SST) prediction can constrain predictions of surface climate evolution, for example temperature and precipitation over Europe [75, 100]. Therefore, studies have increasingly suggested studying "windows of opportunity" or state-dependent predictability to better leverage skill in decadal predictions [18, 19, 25, 26, 38]. This framework focuses on identifying initial states or "windows" where the climate is more predictable, inherently acknowledging that predictability depends on the initial state of the system, and these "windows" provide the best opportunity to make skillful decadal predictions.

In the next 30 years, it is expected that forced warming from anthropogenic greenhouse gasses will continue to increase. This raises the question of whether sources of predictability identified in the historical climate will still contribute to skillful decadal predictions, as the large signal of forced warming could overwhelm predictable signals from internal variability. Here, we aim to investigate whether, and to what extent, internal variability contributes to near term decadal prediction skill in a relatively high future forcing scenario. We develop a novel neural network architecture that separately ingests information about the forced response and internal variability to predict the SST trend over the next 10 years. This architecture is then used to diagnose the contribution of internal variability to prediction accuracy, in order to identify regions where internal variability is a significant source of predictability in a near-future climate. We further address the question of where and when we can attribute prediction skill to internal variability in the presence of high anthropogenic forcing.

4.2 Data & Methods

4.2.1 Community Earth System Model Version 2 Large Ensemble, CESM2-LE

This study uses output from the Community Earth System Model Version 2 Large Ensemble (CESM2-LE) [55, 101]. The CESM2-LE is a collection of 100 ensemble members

of CESM2 run under the specified historical forcing for the Coupled Model Intercomparison Project Phase 6 [CMIP6; 56] for years 1850-2014, and the SSP3-7.0 future radiative forcing for 2015-2100 [102]. Note that ensemble members 1-50 use the biomass burning specified for CMIP6 however members 51-100 use a smoothed version which affects end-of-century warming [103]. The ensemble members are also designed to sample the phase of the Atlantic Meridional Overturning Circulation (AMOC) by splitting the full ensemble into five groups based on their initialized AMOC. We account for both the biomass forcing and AMOC initialization in our experiment design (see next section). We use sea surface temperature (SST) output bilinearly regridded to $5^\circ \times 5^\circ$ resolution, and coarsened to annual means at each grid point.

The target predictions in this work are classifications of future decadal SST trends, specifically whether a future trend in a particular region will fall in the lower, middle or upper tercile of the 2020-2050 distribution in that region (Figure 4.1b). We therefore calculate sliding (i.e. starting each year) 10-year linear least-squares trends over the years 2020-2050 in $10^\circ \times 10^\circ$ boxes in the ocean for each ensemble member. Figure 4.1a demonstrates the annual mean time series for a single ensemble member (blue curve) and the forced response (black curve) in a $10^\circ \times 10^\circ$ boxes in the Southern Ocean (240°E - 250°E, 40°S - 50°S) over 1960-2100. SST trends for consecutive 10-year periods (green, orange and red lines) show that even though forced SST increases throughout the 21st century, regional internal variability can still contribute to reduced warming or even cooling in the later part of the 21st century. The distribution of decadal SST trends over 2020-2050 further demonstrates this point (Figure 4.1b), with trends in this particular grid box ranging from -1°C/decade to over 1.5°.

4.2.2 Artificial Neural Networks

Artificial neural networks (ANNs) are used here to both make predictions of future decadal SST trend and also to investigate sources of predictability. We use ANNs because

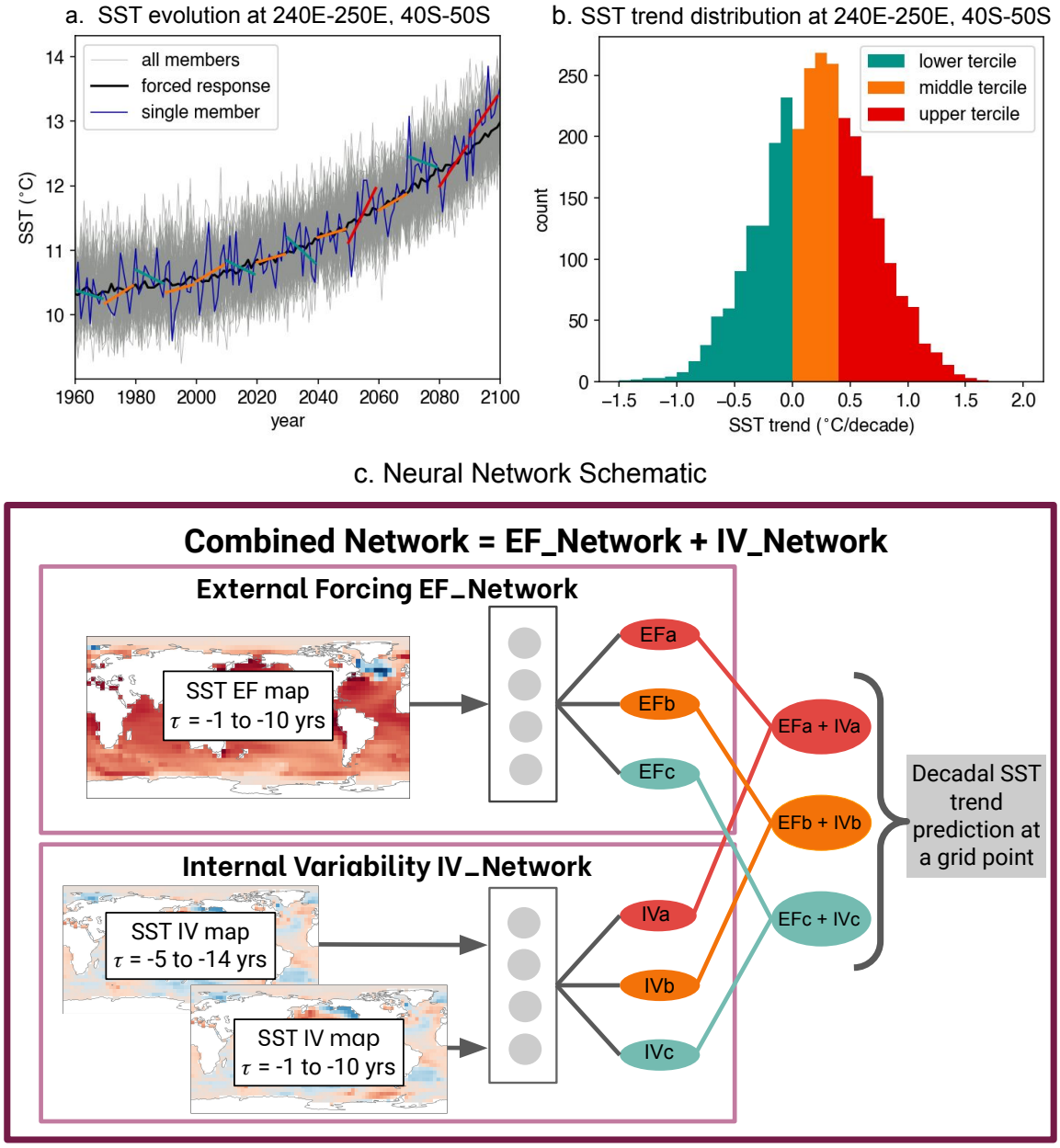

Figure 4.1: (a) CESM2-LE annual mean SST time series area averaged over 240°E - 250°E, 40°S - 50°S. The blue line is a single ensemble member (member 70) with gray lines indicating all other members. The black line is the forced response, defined as the ensemble mean. The green, orange and red lines are individual 10 year trends in SST in ensemble member 70, plotted every 10 years. (b) Distribution of SST trends (°C/decade) between 2020 and 2050 for all ensemble members at 240°E - 250°E, 40°S - 50°S. Color coding indicates the tercile cut-offs, green are in the lower third (or tercile) of the 2020-2050 distribution, orange in the middle third, and red in the upper third. (c) Schematic of the neural network architecture demonstrating the internal variability network (IV_Network) and external forcing network (EF_Network), and their summation to the Combined Network output.

they are an established method of identifying sources of predictability on decadal timescales [38] and for predicting future trends in SST [36]. In this application, neural networks can be considered a non-linear data driven model, taking information about the current and past state of the climate (the state of global SSTs) to make a prediction about a future quantity (regional decadal SST trends). Unlike some previous studies that use ANNs with post-hoc evaluation methods to examine their predictions and predictability [e.g. 36, 38], we separate the prediction problem into an internal variability component and an external forcing component by designing two separate ANNs and only coupling them at the final prediction step of the network. This design allows for a direct investigation of the relative contributions of internal variability and external forcing to the ANN's prediction. The external forcing component is defined as the ensemble mean across all ensemble members at each grid point, and the internal variability for a member is defined at each grid point as the full member minus the ensemble mean.

More specifically, we design two neural networks for the prediction task, named the EF_Network and the IV_Network (Figure 4.1c), which make predictions using only information about external forcing (EF) and internal variability (IV), respectively. The inputs to the neural networks are maps of global SST at $5^\circ \times 5^\circ$ resolution. For the IV_Network, we input two time-lagged maps of internal variability, the first, averaged over the 10 years prior to the prediction ($\tau = -1$ to -10) and the second, averaged over 10 years and lagged by five years ($\tau = -5$ to -14) (Figure 4.1c). These inputs provide the ANN with information about the current state of internal SST variability, and an earlier state, which gives information about the time evolution of internal variability before a prediction. The input to the EF_Network is simply a map of the forced response (i.e., ensemble mean at each grid point) averaged over the 10 years prior to a prediction ($\tau = -1$ to -10)(Figure 4.1c).

Each of the neural networks outputs a classification of the future SST trend (i.e. lower, middle or upper tercile) at a particular grid point, and these values are summed pair-wise (with no activation, weight or bias) to make the Combined Network prediction. This is

demonstrated schematically in Figure 4.1c. The upper box shows the EF_Network, with a single map of the externally forced SST input into an artificial neural network which outputs three values, EFa, EFb and EFc. The lower box is the IV_Network, with two time lagged internal variability maps input into an ANN which also outputs three values, IVa, IVb and IVc. The outputs of the individual models are summed pairwise (EFa+IVa, EFb+IVb, EFc+IVc) to make the final predictions of the future SST trend over the next 10 years. We name this full system the Combined Network. The Combined Network's prediction is taken to be the class with the highest value prediction in the final layer after the softmax activation function is applied. The softmax activation function converts the raw ANN outputs to probability-like values. The inclusion of the softmax activation means that higher value predictions correspond to higher ANN confidence in the prediction such that predictions can then be ranked by their value. For ease of comprehension, specific examples of predictions by the Combined Network with contributions from the IV_Network and EF_Network is provided schematically in Figure C.1.

We subset the 100 ensemble members into 60 members for training data for training the ANNs, 20 members for validation data for selecting the best performing networks, and 20 members for testing which is "unseen" by the ANNs which is exclusively used for performance evaluation and final analysis. The training, validation and testing data is created by grouping individual ensemble members such that each data set equally samples across the five AMOC initializations, and the two different biomass forcings in the individual ensemble members. We use model years 1960-2100 for training the ANNs, but only validate and test on 2020-2050. We use a wider span of model years for training so the neural networks can "see" more possible internal variability states. All results presented are from the testing set. We train 10 combined networks for each $10^\circ \times 10^\circ$ grid box in the ocean and present results from the best network at each location, but results do not qualitatively change if we instead use an average of networks. We define "best" as the

network that achieves the lowest loss on the validation data (not shown). Further, detailed information about the hyperparameters and training process is provided in Appendix C.

4.3 Results

4.3.1 Identifying Contributions to Prediction Skill

The Combined Networks skillfully predict SST trends over 2020-2050 (Figure 4.2a), with accuracy outperforming random chance (by design 33%) over most of the globe. Furthermore, the highest neural network accuracy corresponds to regions that are considered to be more predictable on decadal timescales: the North Pacific, North Atlantic and Southern Indian Ocean [1]. Recent literature has emphasized the importance of identifying so-called "windows of opportunity" for prediction skill on decadal timescales because this provides an indication of when variability may be at its most predictable [19, 38]. We therefore examine the existence the windows of opportunity in near-future decadal predictions, adopting a similar method used by Mayer and Barnes [21], Gordon and Barnes [38] by designating windows of opportunity as the 20% of samples that the Combined Networks assigned the highest confidence at each grid point (see methods). Note that other cutoffs could also be used because neural network accuracy increases with increasing confidence in prediction (see Appendix Figure C.2), however, we choose 20% as this provides the clearest signal of accuracy increase. The 20% most confident samples as designated by the neural networks generally have higher accuracy than all predictions (Figure 4.2b) demonstrating that the neural networks have learned inputs that are more likely to lead to a correct prediction, and hence initial states that are more predictable. Skill improvements are especially evident in the North Pacific PDO region, the North Atlantic Ocean and broadly across the Southern Ocean which aligns with previous work that points to windows of opportunity for enhanced prediction skill existing in these regions [38].

We examine the contribution of internal variability to the Combined Network's skill by using permutation importance testing [104, 33] in Figure 4.2c. Permutation importance

measures a deep learning model's dependence on a certain predictor by scrambling that predictor while holding the others fixed and examining how the output is affected by the corrupted input data. Here, we test the sensitivity of the Combined Network's skill to the internal variability input and assume a null hypothesis that internal variability does not increase prediction skill. We scramble the internal variability input pixel-wise (randomly drawing each pixel individually from its distribution in the testing set) to create a corrupt testing set and calculate the accuracy of the neural network on this data. This process is repeated 500 times for each grid box and we plot the mean accuracy of each neural network on the scrambled data (Figure 4.2c). The neural networks still perform well in some regions (Indian Ocean, Subtropical North Atlantic) when the internal variability input is scrambled implying that network skill is largely derived from the forced response in these regions. The difference between the total accuracy (Figure 4.2a) and the accuracy from the permutation importance testing (Figure 4.2c) provides the contribution of internal variability to the neural network's skill (Figure 4.2e). We consider internal variability to significantly increase a network's skill (reject the null hypothesis) if the network's accuracy on the true testing set is greater than the 90th percentile of the accuracy on the scrambled data. Internal variability significantly contributes to the Combined Networks's skill in the northern and eastern edge of the North Pacific and extending into the tropical Pacific (i.e. the PDO "horseshoe"). There is also enhanced prediction skill from internal variability in the subpolar North Atlantic. Internal variability in these regions has previously been shown to be predictable in studies of pre-industrial and historical climate [38, 1], but these results further imply that internal variability in these regions can provide predictability in the presence of relatively high anthropogenic forcing.

In conjunction with identifying windows of opportunity for improved prediction skill, studies have underlined the difficulty in attributing prediction skill during windows of opportunity to either internal variability, or time varying changes in anthropogenic forcing in the historical period [24]. We therefore use permutation importance to decipher to what

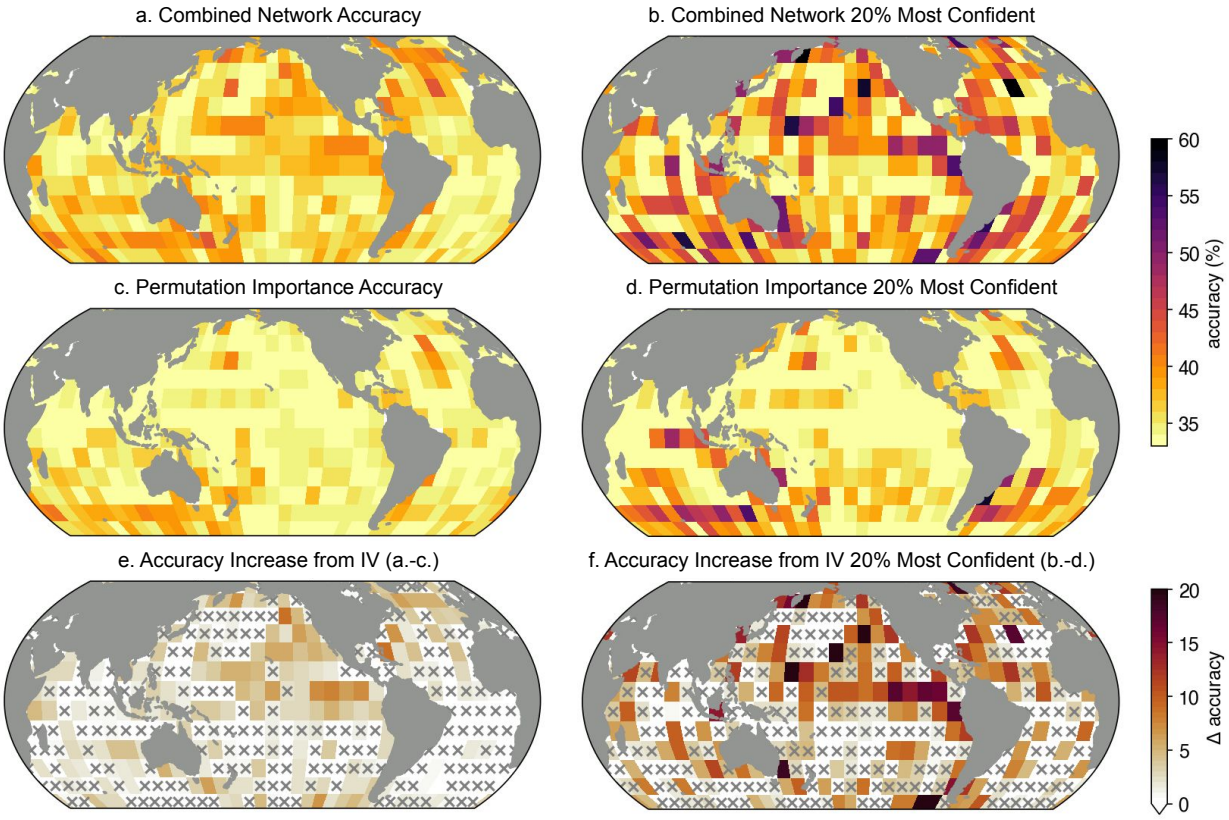

Figure 4.2: (a) Neural network accuracy at each grid point for testing samples in the years 2020-2050. (b) Neural network accuracy on 20% of testing samples with highest confidence. (c) Neural network accuracy when internal variability input is scrambled. (d) Accuracy for 20% most confident predictions when internal variability is scrambled. (e) Difference in accuracy between total accuracy and accuracy on scrambled internal variability (i.e. panel a minus panel c). Differences that are not significant at 90% are stippled. (f) Difference in accuracy between accuracy of confident predictions and accuracy of confident predictions when internal variability is scrambled (i.e. panel b minus panel d). Differences not significant at 90% are stippled.

extent internal variability contributes to prediction skill during windows of opportunity. First we calculate the skill of the 20% most confident predictions in the scrambled internal variability data (Figure 4.2d). We find skill increases in the Southern Indian and Atlantic Oceans even with scrambled internal variability input implying that internal variability did not contribute to the enhanced skill during windows of opportunity in these regions. This increased skill is hence likely derived solely from the external forcing input. Notably, skill enhancements in the North Pacific, Tropical Pacific and some of the North Atlantic Ocean during windows of opportunity can likely be attributed to the networks learning predictable internal variability (Figure 4.2f). Furthermore, in regions where internal variability contributes substantially to prediction skill, enhancements during windows of opportunity are even larger than across all predictions. For example, the general accuracy increase provided by internal variability in the Tropical Pacific is approximately 5-8 percentage points (Figure 4.2e) but accuracy increases to up to 15 percentage points during windows of opportunity (Figure 4.2f).

4.3.2 Internal Sources of Predictability over 2020-2050

Having identified regions where internal variability provides enhanced skill for SST trend prediction, we now examine how large-scale phenomena can lead to regional prediction skill in a future climate. We first isolate grid points where internal variability contributes significantly to prediction skill in 2020-2050 (i.e. regions un-stippled in Figure 4.2e). We then classify whether the IV_Network at each grid point was correct or incorrect for each test sample prediction. Four example IV_Network accuracy timeseries are demonstrated schematically in Figure 4.3a. We use K-means clustering to cluster these prediction outcome lists, resulting in clusters of grid points with correlated prediction skill. That is, where skill within a cluster is more likely to be associated with the same input pattern of variability. We choose six clusters as this number appeared a reasonable choice for this data (Figure C.3) and plot the assigned clusters in Figure 4.3b. Three distinct

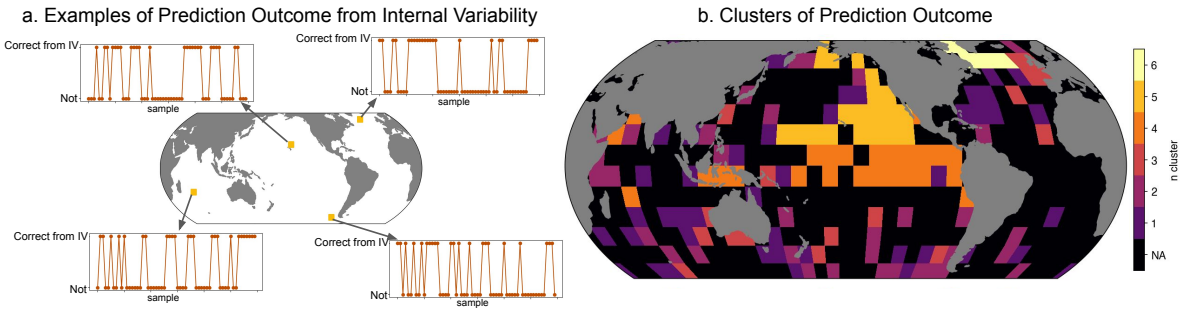

Figure 4.3: a. Example list of prediction outcome from internal variability at four grid points. b. K-means clusters of accuracy from internal variability. Black regions were not included in clustering.

spatial regions emerge; a Tropical Pacific cluster, a North Pacific cluster and a North Atlantic cluster (clusters 4, 5 and 6 respectively). These grid points are also well described by their respective centroids (Figure C.3b and C.3c). The remaining three clusters (clusters 1, 2 and 3) are not spatially distinct, and investigation into their sources of predictability are beyond the scope of this study. However, the underlying variability that leads to similar prediction skill within these clusters remains an intriguing avenue for future work. We choose to focus the remainder of this study on clusters 4, 5 and 6.

To investigate patterns of internal variability that correspond to common predictability within each cluster, we first compute the IV_Network accuracy within a cluster for each testing sample (i.e. percentage of grid points in the cluster that were correctly predicted) and isolate samples with cluster accuracy greater than 50% (i.e. more than 50% of grid cells within the cluster are correctly prediction for that sample). For example, Figure 4.4a and Figure 4.4d is the composite of input samples for which cluster accuracy in the North Pacific cluster is greater than 50%. We additionally separate the composites based on the predicted tercile within a cluster so that the opposing trend signals are not mixed in the composite. This distinction assumes that due to the spatial proximity of grid points within each cluster, that the correct target prediction will generally be the same.

In the North Pacific cluster (cluster #5), predictable upper tercile SST trends follow slightly positive SSTs in the PDO horseshoe 5-14 years before the prediction (Figure 4.4a) which appear to become more negative in the 10 years directly preceding the prediction

(Figure 4.4d). This SST evolution likely leads to predictable upper tercile trends because the leading mode of decadal variability in the North Pacific, the PDO, acts on approximately a 10-20 year timescale [10]. Increasingly negative SSTs in the horseshoe region over a 15 year period are thus likely followed by warming SSTs in the next decade. AppendixFigure C.4a shows the composite annual mean PDO index for the samples in this cluster, demonstrating the evolution of decreasing PDO index transitioning back to increasing PDO index over the input and prediction output window. We also suggest the opposite mechanism for predictions of lower tercile SST trends in the North Pacific cluster (Figure 4.4g and 4.4j), with strengthening positive SSTs in the horseshoe region over the input period leading to more predictable lower tercile SST trends. See Figure C.4d for composite PDO index over this period. These results provide evidence that some SST trends that are lower than that of the forced response may be predictable in the near future, and predictable trends are associated with the decadal evolution of the PDO.

In the North Atlantic cluster, precursors to predictable positive SST trends appear to be a strengthening SST dipole between the North Atlantic subpolar gyre and subtropical North Atlantic, with subpolar gyre SST anomalies becoming more negative, and the subtropical SST anomalies becoming more positive (Figure 4.4b and 4.4e). This SST pattern is likely driven by a similar mechanism to that identified by Borchert et al. [26], with strengthening positive SST anomalies in the subtropical Atlantic forming a predictable state for positive SST anomalies in the subpolar gyre. Appendix Figure C.4b shows the composite SST evolution for the North Atlantic subpolar gyre for these predictions, supporting this theory. Predictable lower tercile SST trends in the subpolar gyre are preceded by strengthening positive SST anomalies in the 15 years before a prediction (Figure 4.4h and 4.4k). This positive anomaly is likely followed by a negative anomaly within the 10-year prediction window (Figure C.4e), resulting in net negative SST trends following these initial states. Here we find that there are more samples in the upper tercile trend composite (132) than the lower tercile trend composite (38), implying that a warming trend in the subpolar gyre

region may be more predictable than a cooling trend which aligns with previous findings [26, 38]. Furthermore, though these mechanisms leading to predictability in the North Atlantic have been studied previously, here we provide evidence that it may continue to be a source of predictability in the presence of relatively high anthropogenic forcing in the near future.

Lastly, the initial state that corresponds to correct predictions of upper tercile trends in the Tropical Pacific cluster (Figure 4.4c and f) appears to coincide with a strengthening El Nino like pattern in the central Pacific Ocean for the 15 year period preceding the prediction of positive trend, as equatorial Pacific SST anomalies strengthen from prediction lead years 5-14 (Fig. 4.4c) to lead years 1-10 (Fig. 4.4f). We hypothesize that the ANN's are forecasting a shift to La Nina in the early part of the 10-year prediction window as large El Nino events are often followed by a rebound to a La Nina within both observations and CESM2 [105, 57]. This La Nina hence results in large negative SST anomalies in the tropical Pacific in the early part of the prediction window. The SSTs will then likely follow the approximate timescale of the El Nino Southern Oscillation (ENSO) cycle by growing into neutral conditions, and then likely another El Nino within the later part of the future 10-year period, resulting in a net positive trend over the forecast period. Appendix Figure C.4c demonstrates the composite annual mean Nino3.4 index for the 10 years preceding and following the upper tercile predictions for samples in Fig 4.4c and 4.4f, showing a robust Nino3.4 decrease early in the prediction window, followed by a positive trend for the next decade. Conversely, a common initial state for predictable lower tercile SST trends for the ENSO cluster (Figure 4.4i and 4.4l) shows a strengthening La Nina-like cooling in the central Pacific over the preceding 15 years. Similar to the positive trend prediction, we hypothesize that the ANNs forecast a substantial El Nino early in the prediction window which dramatically increases tropical SSTs. These positive SST anomalies then decay to neutral conditions and a further La Nina event within the 10 year window (see appendix Figure C.4f, like S4c, shows the composite annual mean Nino3.4 index for the preceding and following 10 years of predictions in the

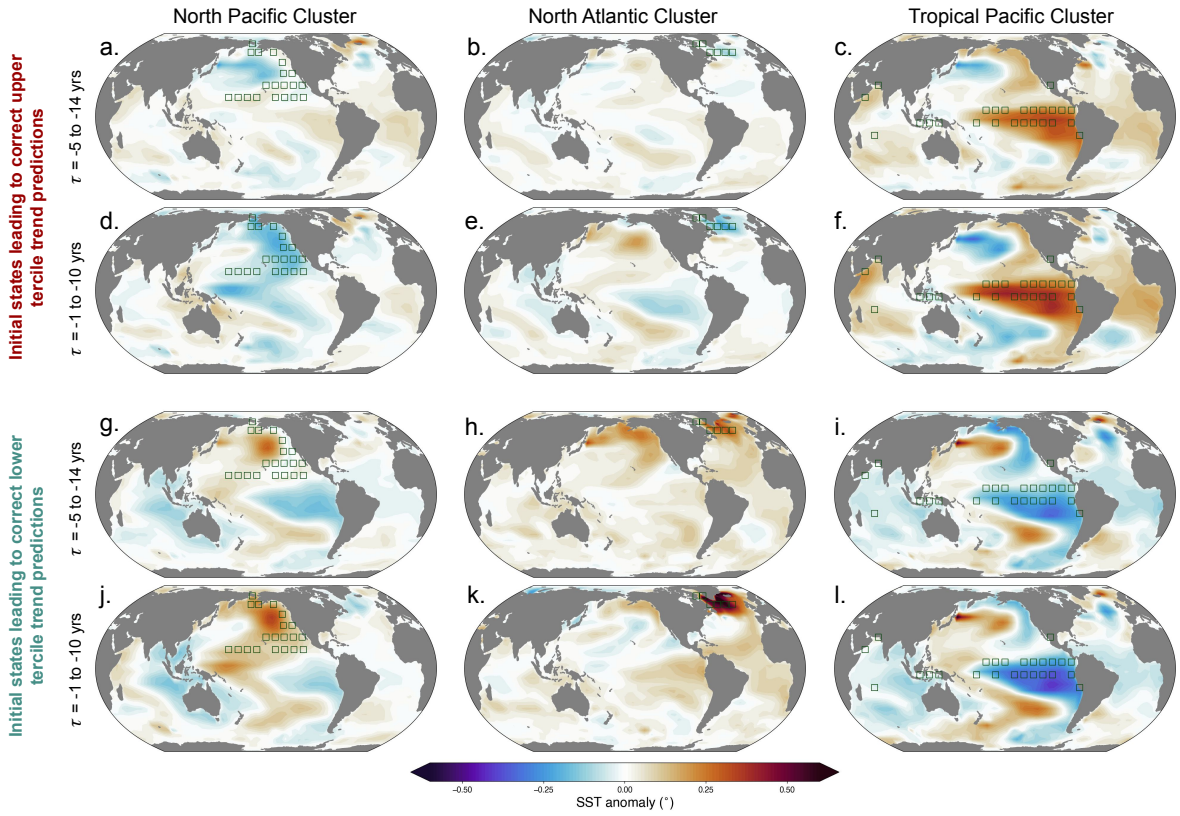

Figure 4.4: Initial states of internal variability that lead to predictable SST trends in the three clusters. Left column is the North Pacific cluster, middle column is the North Atlantic cluster, and the right column is the Tropical Atlantic cluster. The top two rows are initial state composites for upper tercile trend predictability, and the bottom two rows are for the lower tercile trend predictability. All plots are SST anomaly from the ensemble in mean in °C. Grid points included in each cluster are illustrated with boxes in each plot

composite). This tropical SST evolution therefore results in a net negative trend prediction for grid points in the Tropical Pacific cluster (El Nino to neutral to La Nina).

4.4 Conclusion

This study demonstrates that internal variability is a source of predictability in the years 2020-2050 in the CESM2-LE, even with the relatively high anthropogenic climate forcing in the SSP3-7.0 scenario [102]. This result has interesting implications for future projections of regional climate change, as it implies that there may be some periods where SST trend predictions can be more skillful than just predicting the forced response. SST patterns like ENSO and PDO are associated with atmospheric teleconnection which affect temperature

and precipitation over land, so any improved predictions of these patterns can potentially couple to improved future estimates of land surface processes [106]. Furthermore, it is becoming clear that identifying windows of opportunity for improved prediction skill will continue to be a crucial method for making skillful near-term climate forecasts [19]. Our results further suggest that windows of opportunity for increased prediction skill will likely exist in a future with increased anthropogenic greenhouse gas forcing. Interestingly, in places where internal variability is a significant source of skill, correctly forecasting internal variability becomes even more important during windows of opportunity. The importance of internal variability during windows of opportunity reinforces the need to investigate predictable states of internal variability in the climate system as it likely provides the best opportunity for skillful decadal predictions.

Our findings may potentially aid in the communication of climate change and its impacts since internal variability modulates the forced climate change signal, particularly on regional scales. For example, there is still potential for continued warming and extreme events due to internal variability even after aggressive climate change mitigation efforts [107]. Much of the public perceives climate change based on short-term, regional trends [108] so continued warming could harm continued mitigation efforts if previous efforts are perceived to have failed. With improved understanding of predictable internal variability, we can better attribute whether future variability stems from either successful mobilization against climate change, or irreducible internal variability.

Chapter 5: Conclusion

5.1 Summary

This dissertation provides an assessment of predictability on decadal timescales using machine learning. In Chapter 2, I demonstrate that ANNs can be a powerful tool for predicting low frequency variability. Specifically, I use ANNs to predict persistence and transitions in the PDO, showing that ANNs can predict PDO transitions up to 2 years in advance. Furthermore, XAI techniques are used to dig into sources of predictability used by the ANNs to make correct PDO predictions. The ANNs appeared to use off-equatorial heat content anomalies in the western Pacific Ocean to make skillful predictions of PDO transitions. Interestingly, the XAI methods revealed that the ANNs appear to detect different regions of relevance between positive-to-negative transitions and negative-to-positive transitions. It is posited that this difference is likely linked to the asymmetry in predictability of the ENSO cycle, as ENSO events are crucial for triggering a PDO transitions [109]. This work therefore demonstrates the utility of neural networks and explainable AI for understanding decadal climate variability, and supports growing evidence for the mechanisms behind PDO predictability.

The second study in this dissertation (Chapter 3), demonstrates an application of regression neural networks to decadal SST prediction. This chapter leverages a relatively novel technique in climate science – incorporating an uncertainty estimation into an ANN's output – to identify state-dependent predictability in SSTs in a pre-industrial climate simulation. It is found that, when viewed through the lens of state-dependence, predictability may be more widespread in the ocean than previously suspected. By sorting predictions by those assigned the lowest uncertainty, areas exhibiting increased skill are revealed across the North Pacific and Atlantic oceans, as well as the Southern Ocean. Regional case studies reveal sources of predictability in the North Atlantic and North Pacific oceans, which are likely associated with ocean heat transport in the North Atlantic subpolar gyre, and

persistence in the PDO/IPO, respectively. These case studies also revealed asymmetry in predictability, with positive North Atlantic SST anomalies more predictable than negative anomalies, and negative SST anomalies appearing more predictable in the North Pacific. This further underlines the importance of nonlinear methods for examining predictability in the Earth system as it allows the investigation of nonlinearity in predictability.

Lastly, in Chapter 4, I investigate the contribution of internal variability to the predictability of SST trends in a future climate simulation. This study demonstrates that regions that are considered important for initial condition predictability, such as the North Pacific and North Atlantic Oceans, may continue to contribute to predictability in the near future (2020-2050), even under an increasing anthropogenic warming signal. I design an interpretable neural network to dissect the individual contribution of external forcing and internal variability to a future prediction. This approach demonstrates that internal variability is even more important to predictability during windows of opportunity than across all predictions, building on the results of Chapter 3 and further underlining the importance of these initial states for making skillful decadal predictions. Cluster analysis is then used to examine how predictability covaries within different ocean basins. This demonstrates that the North Pacific, North Atlantic and Tropical Pacific oceans could be considered regions within which predictability is governed by distinct patterns of variability. Together, these studies provide a thorough grounding for applying machine learning to decadal prediction, and underline the importance of understanding predictable internal variability to make skillful future forecasts of decadal climate evolution.

5.2 Future Work

A running theme throughout this work is the interconnectedness of predictability between the North Pacific and North Atlantic oceans. For example, Chapter 2 suggests that tropical Atlantic variability may be linked to negative-to-positive PDO transition predictability. Moreover, in Chapter 3 it is found that predictability in the North Atlantic subpolar

gyre may be linked to negative PDO-like variability in the North Pacific Ocean. Future work should therefore explore how these two basins interact and hence how predictable basin interactions influence decadal predictability. Importantly, future investigations should first identify windows of opportunity for increased predictability, and only then further decompose these times into their contributing mechanisms because it is established that the mechanisms leading to low frequency variability do not have consistent contributions in time. For example, even though much of the observed variability in the North Atlantic is due to anthropogenic forcing [71, 72, 15], there are periods where the internal variability signal plays a larger role than the external signal in the observed range of variability [24]. Additionally, this dissertation has demonstrated that windows of opportunity for increased predictability exist across the global ocean. Any study that assumes the governing mechanisms of low frequency variability are constant in time could therefore overlook times where the basins are more likely to interact, particularly if the signal is slight on average. Importantly, this could lead to extended predictability of certain processes, for example if it is found that negative PDO does indeed precondition predictable SSTs in the North Atlantic subpolar gyre, then this may increase the lead time by which these SSTs are predictable.

The work in this dissertation provides evidence that internal variability has a key role in SST anomaly and SST trend prediction, even in a future climate. Therefore, future work should explore the implication of more skillful of predictions of ocean evolution for the skill of predicting atmospheric and land surface variables. For example, it has been shown that skillful decadal timescale predictions of SSTs in the North Atlantic ocean can lead to enhanced skill in predictions of temperature and precipitation over Europe [75, 100]. This finding is predicated on the existence of predictability in the North Atlantic ocean, which, until this dissertation, was considered one of the only places in the ocean that exhibited predictability above climatology on decadal timescales. Chapters 3 and 4, provide a framework for systematically identifying state-dependent predictability (or windows of opportunity for enhanced prediction skill). Future should work leverage sources of

predictability across the ocean and connect these predictable processes with constrained estimates of relevant surface processes. For example, the PDO is associated with sustained teleconnection patterns over the western United States that are thought to be related to low frequency variability in temperature and precipitation [48, 10]. Initial states that lead to enhanced skill of PDO persistence (or indeed transitions) should be investigated for how they can be connected to better predictions of temperature and precipitation over land.

The studies in this dissertation solely utilize output from CESM2, which means that some of the results presented here may be specific to this realization of the Earth system. Therefore, future work should compare predictability across CMIP6 models and future climate scenarios. This approach would further verify the results presented in this dissertation and may also lead to further discoveries of sources of predictability.

In a similar vein, the results of this work should be grounded in the observational record of the climate system. The relative lack of observations (at least compared with the enormous amount of model data) means that neural networks are perhaps not the optimal method for this task because limited data means neural networks cannot be trained on observations alone. We hence suggest that methods that leverage both climate model data and the observational record should be at the forefront of understanding sources of decadal predictability. Recent work has demonstrated that neural networks trained on climate model data can be transferred to predicting observations, either with or without retraining the neural networks [54, 110]. This suggests that the ANN approaches in this dissertation could be applied to observational data after training on large ensembles to examine historical sources of predictability. Furthermore, climate models are becoming better calibrated for Earth system prediction, as well as moving to higher resolution. This has lead to much better skill, as well as likely better resolved processes leading to predictability [111]. Future work should hence continue to leverage these valuable runs to examine predictability. Furthermore, as these model runs are verified against the observational record, they can hence lend understanding to when our observed climate may be more predictable.

This dissertation demonstrates that machine learning is a powerful tool for examining Earth system predictions and predictability. It shows how explainability techniques can support scientific discovery, that the addition of uncertainty to regression tasks can dramatically increase trust in the application of a neural network's prediction, and that data-driven models can be flexibly combined to attribute prediction skill. It is therefore clear that machine learning should continue to be a cornerstone method in the field of climate prediction, and the methods demonstrated here have the potential to be flexibly applied to future prediction tasks. For example, the coupled neural network technique demonstrated in Chapter 4 could be used to attribute the relative roles of any two entangled processes to some future prediction at any time scale, or the uncertainty technique in Chapter 3 could be used to identify predictability in any climate variable. Moreover, these methods do not need to stand in isolation and future work should consider pipelines between the methods used here and other machine learning and explainability techniques. For example, the complementary use of clustering with machine learning to diagnose different sources of skill for machine learning models is well established both in the literature [31, 21], and within this dissertation. Finally, the field of machine learning in climate is evolving quickly, and novel methods are constantly being invented and investigated for their utility in climate science. Our field has been improved for the explosion of people willing to try out these methods, and hence the most important aspect of future work in machine learning applications to climate science is for the practitioners to continue to keep an open mind.

5.3 Concluding Remarks

As the climate continues to warm, and humans continue to feel the effect of anthropogenic climate change, there will only be increased demand for trustworthy predictions of decadal-scale climate evolution. This dissertation provides an assessment of locations and times where we can possibly expect enhanced predictability to exist on these crucial,

actionable timescales. This dissertation also investigates examples of how machine learning can be used to exploit the large amounts of climate data available, allowing for deeper understanding of nonlinear processes in the climate system, and identification of the sources of predictability that currently provide the best opportunity for making skillful predictions of the near future.

Bibliography

[1] Gerald A Meehl, Jadwiga H Richter, Haiyan Teng, Antonietta Capotondi, Kim Cobb, Francisco Doblas-Reyes, Markus G Donat, Matthew H England, John C Fyfe, Weiqing Han, Hyemi Kim, Ben P Kirtman, Yochanan Kushnir, Nicole S Lovenduski, Michael E Mann, William J Merryfield, Veronica Nieves, Kathy Pegion, Nan Rosenbloom, Sara C Sanchez, Adam A Scaife, Doug Smith, Aneesh C Subramanian, Lantao Sun, Diane Thompson, Caroline C Ummenhofer, and Shang-Ping Xie. Initialized Earth System prediction from subseasonal to decadal timescales. *Nat. Rev. Earth Environ.*, 2(5): 340–357, April 2021. ISSN 2662-138X, 2662-138X. doi: 10.1038/s43017-021-00155-x. URL https://www.nature.com/articles/s43017-021-00155-x.

[2] Kirsten L Findell, Rowan Sutton, Nico Caltabiano, Anca Brookshaw, Patrick Heimbach, Masahide Kimoto, Scott Osprey, Doug Smith, James S Risbey, Zhuo Wang, Lijing Cheng, Leandro B Diaz, Markus G Donat, Michael Ek, June-Yi Lee, Shoshiro Minobe, Matilde Rusticucci, Frederic Vitart, and Lin Wang. Explaining and Predicting Earth System Change: A World Climate Research Programme Call to Action. *Bull. Am. Meteorol. Soc.*, 104(1):E325–E339, January 2023. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-21-0280.1. URL https://journals.ametsoc.org/view/journals/bams/104/1/BAMS-D-21-0280.1.xml.

[3] Nick Dunstone, Julia Lockwood, Balakrishnan Solaraju-Murali, Katja Reinhardt, Eirini E Tsartsali, Panos J Athanasiadis, Alessio Bellucci, Anca Brookshaw, Louis-Philippe Caron, Francisco J Doblas-Reyes, Barbara Früh, Nube González-Reviriego, Silvio Gualdi, Leon Hermanson, Stefano Materia, Andria Nicodemou, Dario Nicolì, Klaus Pankatz, Andreas Paxian, Adam Scaife, Doug Smith, and Hazel E Thornton. Towards Useful Decadal Climate Services. *Bull. Am. Meteorol. Soc.*, 103(7):E1705–E1719, July 2022. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-21-0190.1.

URL https://journals.ametsoc.org/view/journals/bams/103/7/BAMS-D-21-0190.1.xml.

[4] Balakrishnan Solaraju-Murali, Nube Gonzalez-Reviriego, Louis-Philippe Caron, Andrej Ceglar, Andrea Toreti, Matteo Zampieri, Pierre-Antoine Bretonnière, Margarita Samsó Cabré, and Francisco J Doblas-Reyes. Multi-annual prediction of drought and heat stress to support decision making in the wheat sector. *npj Climate and Atmospheric Science*, 4(1):1–9, June 2021. ISSN 2397-3722, 2397-3722. doi: 10.1038/s41612-021-00189-4. URL https://www.nature.com/articles/s41612-021-00189-4.

[5] S Moulds, L J Slater, N J Dunstone, and D M Smith. Skillful decadal flood prediction. *Geophys. Res. Lett.*, 50(3), February 2023. ISSN 0094-8276, 1944-8007. doi: 10.1029/2022gl100650. URL https://onlinelibrary.wiley.com/doi/10.1029/2022GL100650.

[6] Andreas Paxian, Katja Reinhardt, Klaus Pankatz, Alexander Pasternack, Maria Paula Lorza-Villegas, Marc Scheibel, Amelie Hoff, Birgit Mannig, Philip Lorenz, and Barbara Früh. High-Resolution Decadal Drought Predictions for German Water Boards: A Case Study for the Wupper Catchment. *Frontiers in Climate*, 4, 2022. ISSN 2624-9553. doi: 10.3389/fclim.2022.867814. URL https://www.frontiersin.org/articles/10.3389/fclim.2022.867814.

[7] Julia F Lockwood, Nick Dunstone, Leon Hermanson, Geoffrey R Saville, Adam A Scaife, Doug Smith, and Hazel E Thornton. A decadal climate service for insurance: Skilful multi-year predictions of North Atlantic hurricane activity and US hurricane damage. *J. Appl. Meteorol. Climatol.*, -1(aop), June 2023. ISSN 1558-8424, 1558-8432. doi: 10.1175/JAMC-D-22-0147.1. URL https://journals.ametsoc.org/view/journals/apme/aop/JAMC-D-22-0147.1/JAMC-D-22-0147.1.xml.

[8] Nathan J Mantua, Steven R Hare, Yuan Zhang, John M Wallace, and Robert C Francis. A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production.

Bull. Am. Meteorol. Soc., 78(6):1069–1080, June 1997. ISSN 0003-0007. doi: 10.1175/1520-0477(1997)078<1069:APICOW>2.0.CO;2. URL http://journals.ametsoc.org/bams/article/78/6/1069/55942/A-Pacific-Interdecadal-Climate-Oscillation-with.

[9] Yuan Zhang, John M Wallace, and David S Battisti. ENSO-like Interdecadal Variability: 1900–93. *J. Clim.*, 10(5):1004–1020, May 1997. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(1997)010<1004:ELIV>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/clim/10/5/1520-0442_1997_010_1004_eliv_2.0.co_2.xml.

[10] 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, Niklas Schneider, Daniel J Vimont, Adam S Phillips, James D Scott, and Catherine A Smith. The Pacific Decadal Oscillation, Revisited. *J. Clim.*, 29(12):4399–4427, June 2016. ISSN 0894-8755. doi: 10.1175/JCLI-D-15-0508.1. URL https://journals.ametsoc.org/jcli/article/29/12/4399/34340/The-Pacific-Decadal-Oscillation-Revisited.

[11] David B Enfield, Alberto M Mestas-Nuñez, and Paul J Trimble. The Atlantic Multidecadal Oscillation and its relation to rainfall and river flows in the continental U.S. *Geophys. Res. Lett.*, 28(10):2077–2080, 2001. ISSN 0094-8276, 1944-8007. doi: 10.1029/2000GL012745. URL http://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000GL012745.

[12] Rong Zhang, Rowan Sutton, Gokhan Danabasoglu, Young-Oh Kwon, Robert Marsh, Stephen G Yeager, Daniel E Amrhein, and Christopher M Little. A review of the role of the Atlantic meridional overturning circulation in Atlantic multidecadal variability and associated climate impacts. *Rev. Geophys.*, 57(2):316–375, June 2019. ISSN 8755-1209, 1944-9208. doi: 10.1029/2019rg000644. URL https://onlinelibrary.wiley.com/doi/10.1029/2019RG000644.

[13] IPCC. Climate change 2023: Synthesis report. a report of the intergovernmental panel on climate change. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change, 2023. (in press).

[14] 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, and N S Lovenduski. Predicting Near-Term Changes in the Earth System: A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model. *Bull. Am. Meteorol. Soc.*, 99(9):1867–1886, September 2018. ISSN 0003-0007. doi: 10.1175/BAMS-D-17-0098.1. URL https://journals.ametsoc.org/bams/article/99/9/1867/70398/Predicting-Near-Term-Changes-in-the-Earth-System-A.

[15] Jeremy M Klavans, Mark A Cane, Amy C Clement, and Lisa N Murphy. NAO predictability from external forcing in the late 20th century. *npj Climate and Atmospheric Science*, 4(1):1–8, March 2021. ISSN 2397-3722, 2397-3722. doi: 10.1038/s41612-021-00177-8. URL https://www.nature.com/articles/s41612-021-00177-8.

[16] Stephen G Yeager, Nan Rosenbloom, Anne A Glanville, Xian Wu, Isla Simpson, Hui Li, Maria J Molina, Kristen Krumhardt, Samuel Mogen, Keith Lindsay, Danica Lombardozzi, Will Wieder, Who M Kim, Jadwiga H Richter, Matthew Long, Gokhan Danabasoglu, David Bailey, Marika Holland, Nicole Lovenduski, Warren G Strand, and Teagan King. The Seasonal-to-Multiyear Large Ensemble (SMYLE) prediction system using the Community Earth System Model version 2. *Geosci. Model Dev.*, 15(16):6451–6493, August 2022. ISSN 1991-959X, 1991-9603. doi: 10.5194/gmd-15-6451-2022. URL https://doi.org/10.5194/gmd-15-6451-2022.

[17] Carlos Delgado-Torres, Markus G Donat, Nube Gonzalez-Reviriego, Louis-Philippe Caron, Panos J Athanasiadis, Pierre-Antoine Bretonnière, Nick J Dunstone, An-Chi Ho, Dario Nicoli, Klaus Pankatz, Andreas Paxian, Núria Pérez-Zanón, Mar-

garida Samsó Cabré, Balakrishnan Solaraju-Murali, Albert Soret, and Francisco J Doblas-Reyes. Multi-Model Forecast Quality Assessment of CMIP6 Decadal Predictions. *J. Clim.*, 35(13):4363–4382, July 2022. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-21-0811.1. URL https://journals.ametsoc.org/view/journals/clim/35/13/JCLI-D-21-0811.1.xml.

[18] William J Merryfield, Johanna Baehr, Lauriane Batté, Emily J Becker, Amy H Butler, Caio A S Coelho, Gokhan Danabasoglu, Paul A Dirmeyer, Francisco J Doblas-Reyes, Daniela I V Domeisen, Laura Ferranti, Tatiana Ilynia, Arun Kumar, Wolfgang A Müller, Michel Rixen, Andrew W Robertson, Doug M Smith, Yuhei Takaya, Matthias Tuma, Frederic Vitart, Christopher J White, Mariano S Alvarez, Constantin Ardilouze, Hannah Attard, Cory Baggett, Magdalena A Balmaseda, Asmerom F Beraki, Partha S Bhattacharjee, Roberto Bilbao, Felipe M de Andrade, Michael J DeFlorio, Leandro B Díaz, Muhammad Azhar Ehsan, Georgios Fragkoulidis, Sam Grainger, Benjamin W Green, Momme C Hell, Johnna M Infanti, Katharina Isensee, Takahito Kataoka, Ben P Kirtman, Nicholas P Klingaman, June-Yi Lee, Kirsten Mayer, Roseanna McKay, Jennifer V Mecking, Douglas E Miller, Nele Neddermann, Ching Ho Justin Ng, Albert Ossó, Klaus Pankatz, Simon Peatman, Kathy Pegion, Judith Perlwitz, G Cristina Recalde-Coronel, Annika Reintges, Christoph Renkl, Balakrishnan Solaraju-Murali, Aaron Spring, Cristiana Stan, Y Qiang Sun, Carly R Tozer, Nicolas Vigaud, Steven Woolnough, and Stephen Yeager. Current and Emerging Developments in Subseasonal to Decadal Prediction. *Bull. Am. Meteorol. Soc.*, 101(6): E869–E896, June 2020. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-19-0037.1. URL https://journals.ametsoc.org/view/journals/bams/101/6/bamsD190037.xml.

[19] Annarita Mariotti, Cory Baggett, Elizabeth A Barnes, Emily Becker, Amy Butler, Dan C Collins, Paul A Dirmeyer, Laura Ferranti, Nathaniel C Johnson, Jeanine Jones, Ben P Kirtman, Andrea L Lang, Andrea Molod, Matthew Newman, Andrew W Robertson, Siegfried Schubert, Duane E Waliser, and John Albers.

Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond. *Bull. Am. Meteorol. Soc.*, 101(5):E608–E625, May 2020. ISSN 0003-0007. doi: 10.1175/BAMS-D-18-0326.1. URL http://journals.ametsoc.org/bams/article/101/5/E608/345558/Windows-of-Opportunity-for-Skillful-Forecasts.

[20] John R Albers and Matthew Newman. A Priori identification of skillful extratropical subseasonal forecasts. *Geophys. Res. Lett.*, 46(21):12527–12536, November 2019. ISSN 0094-8276, 1944-8007. doi: 10.1029/2019gl085270. URL https://onlinelibrary.wiley.com/doi/10.1029/2019GL085270.

[21] Kirsten J Mayer and Elizabeth A Barnes. Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network. *Geophys. Res. Lett.*, 48(10): e2020GL092092, 2021. ISSN 0094-8276, 1944-8007. doi: 10.1029/2020GL092092. URL http://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL092092.

[22] K-C Tseng, E A Barnes, and E D Maloney. Prediction of the midlatitude response to strong Madden-Julian oscillation events on S2S time scales. *Geophys. Res. Lett.*, 45(1): 463–470, January 2018. ISSN 0094-8276, 1944-8007. doi: 10.1002/2017gl075734. URL https://onlinelibrary.wiley.com/doi/10.1002/2017GL075734.

[23] H M Christensen, J Berner, and S Yeager. The Value of Initialization on Decadal Timescales: State-Dependent Predictability in the CESM Decadal Prediction Large Ensemble. *J. Clim.*, 33(17):7353–7370, September 2020. ISSN 0894-8755. doi: 10.1175/JCLI-D-19-0571.1. URL https://journals.ametsoc.org/jcli/article/33/17/7353/348619/The-Value-of-Initialization-on-Decadal-Timescales.

[24] Leonard F Borchert, Matthew B Menary, Didier Swingedouw, Giovanni Sgubin, Leon Hermanson, and Juliette Mignot. Improved decadal predictions of north Atlantic subpolar gyre SST in CMIP6. *Geophys. Res. Lett.*, 48(3), February 2021. ISSN 0094-8276, 1944-8007. doi: 10.1029/2020gl091307. URL https://onlinelibrary.wiley.com/doi/10.1029/2020GL091307.

[25] Sebastian Brune, André Düsterhus, Holger Pohlmann, Wolfgang A Müller, and Johanna Baehr. Time dependency of the prediction skill for the North Atlantic subpolar gyre in initialized decadal hindcasts. *Clim. Dyn.*, 51(5):1947–1970, September 2018. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-017-3991-4. URL https://doi.org/10.1007/s00382-017-3991-4.

[26] Leonard F Borchert, Wolfgang A Müller, and Johanna Baehr. Atlantic Ocean Heat Transport Influences Interannual-to-Decadal Surface Temperature Predictability in the North Atlantic Region. *J. Clim.*, 31(17):6763–6782, September 2018. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-17-0734.1. URL https://journals.ametsoc.org/view/journals/clim/31/17/jcli-d-17-0734.1.xml.

[27] Maria J Molina, Travis A O'Brien, Gemma Anderson, Moetasim Ashfaq, Katrina E Bennett, William D Collins, Katherine Dagon, Juan M Restrepo, and Paul A Ullrich. A Review of Recent and Emerging Machine Learning Applications for Climate Variability and Weather Phenomena. *Artificial Intelligence for the Earth Systems*, -1(aop):1–46, June 2023. ISSN 2769-7525. doi: 10.1175/AIES-D-22-0086.1. URL https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-22-0086.1/AIES-D-22-0086.1.xml?tab_body=pdf.

[28] Zane K Martin, Elizabeth A Barnes, and Eric Maloney. Using simple, explainable neural networks to predict the Madden-Julian oscillation. *J. Adv. Model. Earth Syst.*, 14(5), May 2022. ISSN 1942-2466. doi: 10.1029/2021ms002774. URL https://onlinelibrary.wiley.com/doi/10.1029/2021MS002774.

[29] Yoo-Geun Ham, Jong-Seong Kug, and Jong-Yeon Park. Two distinct roles of Atlantic SSTs in ENSO variability: North Tropical Atlantic SST and Atlantic Niño. *Geophys. Res. Lett.*, 40(15):4012–4017, August 2013. ISSN 0094-8276, 1944-8007. doi: 10.1002/grl.50729. URL http://dx.doi.org/10.1002/grl.50729.

[30] Balasubramanya T Nadiga. Reservoir Computing as a Tool for Climate Predictability Studies. *J. Adv. Model. Earth Syst.*, 13(4):e2020MS002290, 2021. ISSN 1942-2466. doi: 10.1029/2020MS002290. URL http://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002290.

[31] Benjamin A Toms, Elizabeth A Barnes, and James W Hurrell. Assessing decadal predictability in an earth-system model using explainable neural networks. *Geophys. Res. Lett.*, 48(12), June 2021. ISSN 0094-8276, 1944-8007. doi: 10.1029/2021gl093842. URL https://onlinelibrary.wiley.com/doi/10.1029/2021GL093842.

[32] Benjamin A Toms, Elizabeth A Barnes, and Imme Ebert-Uphoff. Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *J. Adv. Model. Earth Syst.*, 12(9):e2019MS002002, 2020. ISSN 1942-2466. doi: 10.1029/2019MS002002. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002002.

[33] Amy McGovern, Ryan Lagerquist, David John Gagne, G Eli Jergensen, Kimberly L Elmore, Cameron R Homeyer, and Travis Smith. Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bull. Am. Meteorol. Soc.*, 100(11):2175–2199, November 2019. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-18-0195.1. URL https://journals.ametsoc.org/view/journals/bams/100/11/bams-d-18-0195.1.xml.

[34] Frances V Davenport and Noah S Diffenbaugh. Using machine learning to analyze physical causes of climate change: A case study of U.s. midwest extreme precipitation. *Geophys. Res. Lett.*, 48(15), August 2021. ISSN 0094-8276, 1944-8007. doi: 10.1029/2021gl093787. URL https://onlinelibrary.wiley.com/doi/10.1029/2021GL093787.

[35] Maike Sonnewald and Redouane Lguensat. Revealing the impact of global heating on North Atlantic circulation using transparent machine learning. *J. Adv. Model.*

Earth Syst., July 2021. ISSN 1942-2466. doi: 10.1029/2021ms002496. URL https://onlinelibrary.wiley.com/doi/10.1029/2021MS002496.

[36] Zachary M Labe and Elizabeth A Barnes. Predicting slowdowns in decadal climate warming trends with explainable neural networks. *Geophys. Res. Lett.*, 49 (9), May 2022. ISSN 0094-8276, 1944-8007. doi: 10.1029/2022gl098173. URL https://onlinelibrary.wiley.com/doi/10.1029/2022GL098173.

[37] Emily M Gordon, Elizabeth A Barnes, and James W Hurrell. Oceanic harbingers of Pacific decadal oscillation predictability in CESM2 detected by neural networks. *Geophys. Res. Lett.*, 48(21), November 2021. ISSN 0094-8276, 1944-8007. doi: 10.1029/2021gl095392. URL https://onlinelibrary.wiley.com/doi/10.1029/2021GL095392.

[38] Emily M Gordon and Elizabeth A Barnes. Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in CESM2. *Geophys. Res. Lett.*, 49(15), August 2022. ISSN 0094-8276, 1944-8007. doi: 10.1029/2022gl098635. URL https://onlinelibrary.wiley.com/doi/10.1029/2022GL098635.

[39] Emily M Gordon, Elizabeth A Barnes, and Frances V Davenport. Separating internal and forced contributions to near term SST predictability in the CESM2-LE. *EarthArXiv*, July 2023. doi: 10.31223/x5bd5j. URL https://eartharxiv.org/repository/view/5699/.

[40] Christophe Cassou, Yochanan Kushnir, Ed Hawkins, Anna Pirani, Fred Kucharski, In-Sik Kang, and Nico Caltabiano. Decadal Climate Variability and Predictability: Challenges and Opportunities. *Bull. Am. Meteorol. Soc.*, 99(3):479–490, March 2018. ISSN 0003-0007. doi: 10.1175/BAMS-D-16-0286.1. URL http://journals.ametsoc.org/bams/article/99/3/479/70287/Decadal-Climate-Variability-and-Predictability.

[41] Clara Deser, Michael A Alexander, and Michael S Timlin. Understanding the Persistence of Sea Surface Temperature Anomalies in Midlatitudes. *J. Clim.*, 16(1): 57–72, January 2003. ISSN 0894-8755. doi: 10.1175/1520-0442(2003)016<0057:UTPOSS>2.0.CO;2. URL http://journals.ametsoc.org/jcli/article/16/1/57/29818/Understanding-the-Persistence-of-Sea-Surface.

[42] Matthew Newman, Gilbert P Compo, and Michael A Alexander. ENSO-Forced Variability of the Pacific Decadal Oscillation. *J. Clim.*, 16(23):3853–3857, December 2003. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(2003)016<3853:EVOTPD>2.0.CO;2. URL http://journals.ametsoc.org/view/journals/clim/16/23/1520-0442_2003_016_3853_evotpd_2.0.co_2.xml.

[43] Niklas Schneider and Bruce D Cornuelle. The Forcing of the Pacific Decadal Oscillation. *J. Clim.*, 18(21):4355–4373, November 2005. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI3527.1. URL https://journals.ametsoc.org/view/journals/clim/18/21/jcli3527.1.xml.

[44] Matthew Newman. Interannual to Decadal Predictability of Tropical and North Pacific Sea Surface Temperatures. *J. Clim.*, 20(11):2333–2356, June 2007. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI4165.1. URL https://journals.ametsoc.org/view/journals/clim/20/11/jcli4165.1.xml.

[45] Michael A Alexander, Ludmila Matrosova, Cécile Penland, James D Scott, and Ping Chang. Forecasting Pacific SSTs: Linear Inverse Model Predictions of the PDO. *J. Clim.*, 21(2):385–402, January 2008. ISSN 0894-8755, 1520-0442. doi: 10.1175/2007JCLI1849.1. URL http://journals.ametsoc.org/view/journals/clim/21/2/2007jcli1849.1.xml.

[46] Daniela Faggiani Dias, Aneesh Subramanian, Laure Zanna, and Arthur J Miller. Remote and local influences in forecasting Pacific SST: a linear inverse model and

a multimodel ensemble study. *Clim. Dyn.*, 52(5):3183–3201, March 2019. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-018-4323-z. URL https://doi.org/10.1007/s00382-018-4323-z.

[47] Riccardo Farneti, Franco Molteni, and Fred Kucharski. Pacific interdecadal variability driven by tropical–extratropical interactions. *Clim. Dyn.*, 42(11):3337–3355, June 2014. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-013-1906-6. URL https://doi.org/10.1007/s00382-013-1906-6.

[48] Gerald A Meehl and Aixue Hu. Megadroughts in the Indian Monsoon Region and Southwest North America and a Mechanism for Associated Multidecadal Pacific Sea Surface Temperature Anomalies. *J. Clim.*, 19(9):1605–1623, May 2006. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI3675.1. URL http://journals.ametsoc.org/view/journals/clim/19/9/jcli3675.1.xml.

[49] Gerald A Meehl, Haiyan Teng, and Julie M Arblaster. Climate model simulations of the observed early-2000s hiatus of global warming. *Nat. Clim. Chang.*, 4(10):898–902, October 2014. ISSN 1758-678X, 1758-6798. doi: 10.1038/nclimate2357. URL http://www.nature.com/articles/nclimate2357.

[50] Gerald A Meehl, Aixue Hu, and Haiyan Teng. Initialized decadal prediction for transition to positive phase of the Interdecadal Pacific Oscillation. *Nat. Commun.*, 7(1):11718, June 2016. ISSN 2041-1723. doi: 10.1038/ncomms11718. URL http://www.nature.com/articles/ncomms11718.

[51] Peter R Gent, Gokhan Danabasoglu, Leo J Donner, Marika M Holland, Elizabeth C Hunke, Steve R Jayne, David M Lawrence, Richard B Neale, Philip J Rasch, Mariana Vertenstein, Patrick H Worley, Zong-Liang Yang, and Minghua Zhang. The Community Climate System Model Version 4. *J. Clim.*, 24(19):4973–4991, October 2011. ISSN 0894-8755, 1520-0442. doi: 10.1175/2011JCLI4083.1. URL https://journals.ametsoc.org/view/journals/clim/24/19/2011jcli4083.1.xml?tab_body=fulltext-display.

[52] Zhenghui Lu, Naiming Yuan, Qing Yang, Zhuguo Ma, and Jürgen Kurths. Early warning of the Pacific decadal oscillation phase transition using complex network analysis. *Geophys. Res. Lett.*, 48(7), April 2021. ISSN 0094-8276, 1944-8007. doi: 10.1029/2020gl091674. URL https://onlinelibrary.wiley.com/doi/10.1029/2020GL091674.

[53] 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. *J. Adv. Model. Earth Syst.*, 12(9), September 2020. ISSN 1942-2466. doi: 10.1029/2020ms002195. URL https://onlinelibrary.wiley.com/doi/10.1029/2020MS002195.

[54] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775):568–572, September 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1559-7. URL http://www.nature.com/articles/s41586-019-1559-7.

[55] G Danabasoglu, J-F Lamarque, J Bacmeister, D A Bailey, A K DuVivier, J Edwards, L K Emmons, J Fasullo, R Garcia, A Gettelman, C Hannay, M M Holland, W G Large, P H Lauritzen, D M Lawrence, J T M Lenaerts, K Lindsay, W H Lipscomb, M J Mills, R Neale, K W Oleson, B Otto-Bliesner, A S Phillips, W Sacks, S Tilmes, L van Kampenhout, M Vertenstein, A Bertini, J Dennis, C Deser, C Fischer, B Fox-Kemper, J E Kay, D Kinnison, P J Kushner, V E Larson, M C Long, S Mickelson, J K Moore, E Nienhouse, L Polvani, P J Rasch, and W G Strand. The Community Earth System Model Version 2 (CESM2). *J. of Adv. in Model. Earth Syst.*, 12(2): e2019MS001916, 2020. ISSN 1942-2466. doi: 10.1029/2019MS001916. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001916.

[56] V Eyring, S Bony, G A Meehl, C A Senior, B Stevens, R J Stouffer, and K E Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) exper-

imental design and organization. *Geosci. Model Dev.*, 9(5):1937–1958, 2016. doi: 10.5194/gmd-9-1937-2016. URL https://gmd.copernicus.org/articles/9/1937/2016/.

[57] A Capotondi, C Deser, A S Phillips, Y Okumura, and S M Larson. ENSO and Pacific Decadal Variability in the Community Earth System Model Version 2. *J. Adv. Model. Earth Syst.*, 12(12):e2019MS002022, 2020. ISSN 1942-2466. doi: 10.1029/2019MS002022. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002022.

[58] John T Fasullo and Robert S Nerem. Interannual Variability in Global Mean Sea Level Estimated from the CESM Large and Last Millennium Ensembles. *Water*, 8(11):491, October 2016. ISSN 0305-3105. doi: 10.3390/w8110491. URL https://www.mdpi.com/2073-4441/8/11/491.

[59] Shujun Li, Lixin Wu, Yun Yang, Tao Geng, Wenju Cai, Bolan Gan, Zhaohui Chen, Zhao Jing, Guojian Wang, and Xiaohui Ma. The Pacific Decadal Oscillation less predictable under greenhouse warming. *Nat. Clim. Chang.*, 10(1):30–34, December 2019. ISSN 1758-678X. doi: 10.1038/s41558-019-0663-x. URL https://www.nature.com/articles/s41558-019-0663-x.

[60] 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):1–46, 2015. doi: 10.1371/journal.pone.0130140. URL https://doi.org/10.1371/journal.pone.0130140.

[61] Antonios Mamalakis, Imme Ebert-Uphoff, and Elizabeth A Barnes. Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environ. Data Sci.*, 1:e8, January 2022. ISSN 2634-4602. doi: 10.1017/eds.2022.7.

[62] Fred Kucharski, Farah Ikram, Franco Molteni, Riccardo Farneti, In-Sik Kang, Hyun-Ho No, Martin P King, Graziano Giuliani, and Kristian Mogensen. Atlantic forcing of Pacific decadal variability. *Clim. Dyn.*, 46(7):2337–2351, April 2016. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-015-2705-z. URL https://doi.org/10.1007/s00382-015-2705-z.

[63] Y Chikamoto, Z F Johnson, S-Y Simon Wang, M J McPhaden, and T Mochizuki. El Niño–Southern Oscillation Evolution Modulated by Atlantic Forcing. *J. Geophys. Res. C: Oceans*, 125(8):e2020JC016318, 2020. ISSN 2169-9291. doi: 10.1029/2020JC016318. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JC016318.

[64] Zachary F Johnson, Yoshimitsu Chikamoto, S-Y Simon Wang, Michael J McPhaden, and Takashi Mochizuki. Pacific decadal oscillation remotely forced by the equatorial Pacific and the Atlantic Oceans. *Clim. Dyn.*, 55(3):789–811, August 2020. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-020-05295-2. URL https://doi.org/10.1007/s00382-020-05295-2.

[65] Gerald A Meehl, Aixue Hu, Frederic Castruccio, Matthew H England, Susan C Bates, Gokhan Danabasoglu, Shayne McGregor, Julie M Arblaster, Shang-Ping Xie, and Nan Rosenbloom. Atlantic and Pacific tropics connected by mutually interactive decadal-timescale processes. *Nat. Geosci.*, pages 1–7, December 2020. ISSN 1752-0894, 1752-0908. doi: 10.1038/s41561-020-00669-x. URL http://www.nature.com/articles/s41561-020-00669-x.

[66] Kevin E Trenberth and John T Fasullo. An apparent hiatus in global warming? *Earths Future*, 1(1):19–32, December 2013. ISSN 2328-4277. doi: 10.1002/2013ef000165. URL https://onlinelibrary.wiley.com/doi/10.1002/2013EF000165.

[67] S Power, T Casey, C Folland, A Colman, and V Mehta. Inter-decadal modulation of the impact of ENSO on australia. *Clim. Dyn.*, 15(5):319–324, May 1999. ISSN

0930-7575, 1432-0894. doi: 10.1007/s003820050284. URL https://doi.org/10.1007/s003820050284.

[68] Gerald A Meehl, Aixue Hu, Julie M Arblaster, John Fasullo, and Kevin E Trenberth. Externally Forced and Internally Generated Decadal Climate Variability Associated with the Interdecadal Pacific Oscillation. *J. Clim.*, 26(18):7298–7310, September 2013. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-12-00548.1. URL https://journals.ametsoc.org/view/journals/clim/26/18/jcli-d-12-00548.1.xml.

[69] Shang-Ping Xie and Youichi Tanimoto. A pan-Atlantic decadal climate oscillation. *Geophys. Res. Lett.*, 25(12):2185–2188, 1998. ISSN 0094-8276, 1944-8007. doi: 10.1029/98GL01525. URL http://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/98GL01525.

[70] Amy Clement, Katinka Bellomo, Lisa N Murphy, Mark A Cane, Thorsten Mauritsen, Gaby Rädel, and Bjorn Stevens. The Atlantic Multidecadal Oscillation without a role for ocean circulation. *Science*, 350(6258):320–324, October 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aab3980. URL https://science.sciencemag.org/content/350/6258/320.

[71] Michael E Mann, Byron A Steinman, Daniel J Brouillette, and Sonya K Miller. Multidecadal climate oscillations during the past millennium driven by volcanic forcing. *Science*, 371(6533):1014–1019, March 2021. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abc5810. URL http://dx.doi.org/10.1126/science.abc5810.

[72] Ben B B Booth, Nick J Dunstone, Paul R Halloran, Timothy Andrews, and Nicolas Bellouin. Aerosols implicated as a prime driver of twentieth-century north atlantic climate variability. *Nature*, 484(7393):228–232, April 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature10946. URL http://dx.doi.org/10.1038/nature10946.

[73] Rong Zhang and Thomas L Delworth. Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. *Geophys. Res. Lett.*, 33(17), 2006. ISSN 0094-8276, 1944-8007. doi: 10.1029/2006gl026267. URL http://doi.wiley.com/10.1029/2006GL026267.

[74] Isla R Simpson, Clara Deser, Karen A McKinnon, and Elizabeth A Barnes. Modeled and Observed Multidecadal Variability in the North Atlantic Jet Stream and Its Connection to Sea Surface Temperatures. *J. Clim.*, 31(20):8313–8338, October 2018. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-18-0168.1. URL https://journals.ametsoc.org/view/journals/clim/31/20/jcli-d-18-0168.1.xml.

[75] 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. *Nat. Geosci.*, 12(8):613–619, August 2019. ISSN 1752-0894, 1752-0908. doi: 10.1038/s41561-019-0391-x. URL https://www.nature.com/articles/s41561-019-0391-x.

[76] Grant Branstator, Haiyan Teng, Gerald A Meehl, Masahide Kimoto, Jeff R Knight, Mojib Latif, and A Rosati. Systematic estimates of Initial-Value decadal predictability for six AOGCMs. *J. Clim.*, 25(6):1827–1846, March 2012. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-11-00227.1. URL https://journals.ametsoc.org/view/journals/clim/25/6/jcli-d-11-00227.1.xml.

[77] George J Boer. Decadal potential predictability of twenty-first century climate. *Clim. Dyn.*, 36(5):1119–1133, March 2011. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-010-0747-9. URL https://doi.org/10.1007/s00382-010-0747-9.

[78] Grant Branstator and Haiyan Teng. Two limits of Initial-Value decadal predictability in a CGCM. *J. Clim.*, 23(23):6292–6311, December 2010. ISSN 0894-8755, 1520-0442. doi: 10.1175/2010JCLI3678.1. URL https://journals.ametsoc.org/view/journals/clim/23/23/2010jcli3678.1.xml.

[79] Holger Pohlmann, Michael Botzet, Mojib Latif, Andreas Roesch, Martin Wild, and Peter Tschuck. Estimating the Decadal Predictability of a Coupled AOGCM. *J. Clim.*, 17(22):4463–4472, November 2004. ISSN 0894-8755, 1520-0442. doi: 10.1175/3209.1. URL https://journals.ametsoc.org/view/journals/clim/17/22/3209.1.xml.

[80] R Msadek, K W Dixon, T L Delworth, and W Hurlin. Assessing the predictability of the atlantic meridional overturning circulation and associated fingerprints. *Geophys. Res. Lett.*, 37(19), October 2010. ISSN 0094-8276, 1944-8007. doi: 10.1029/2010gl044517. URL http://doi.wiley.com/10.1029/2010GL044517.

[81] Dallas Foster, David John Gagne, II, and Daniel B Whitt. Probabilistic machine learning estimation of ocean mixed layer depth from dense satellite and sparse in situ observations. *J. Adv. Model. Earth Syst.*, 13(12), December 2021. ISSN 1942-2466. doi: 10.1029/2021ms002474. URL https://onlinelibrary.wiley.com/doi/10.1029/2021MS002474.

[82] Arthur P Guillaumin and Laure Zanna. Stochastic-deep learning parameterization of ocean momentum forcing. *J. Adv. Model. Earth Syst.*, 13(9), September 2021. ISSN 1942-2466. doi: 10.1029/2021ms002534. URL https://onlinelibrary.wiley.com/doi/10.1029/2021MS002534.

[83] Elizabeth A Barnes and Randal J Barnes. Controlled abstention neural networks for identifying skillful predictions for regression problems. *J. Adv. Model. Earth Syst.*, 13(12), December 2021. ISSN 1942-2466. doi: 10.1029/2021ms002575. URL https://onlinelibrary.wiley.com/doi/10.1029/2021MS002575.

[84] Claude Frankignoul and Klaus Hasselmann. Stochastic climate models, part II application to sea-surface temperature anomalies and thermocline variability. *Tell'Us*, 29(4):289–305, 1977. ISSN 0804-6042, 2153-3490. doi: 10.1111/j.2153-3490.1977.tb00740.x. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1977.tb00740.x.

[85] Michael A Alexander and Clara Deser. A mechanism for the recurrence of wintertime midlatitude SST anomalies. *J. Phys. Oceanogr.*, 25(1):122–137, January 1995. ISSN 0022-3670, 1520-0485. doi: 10.1175/1520-0485(1995)025<0122:AMFTRO>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/phoc/25/1/1520-0485_1995_025_0122_amftro_2_0_co_2.xml?tab_body=fulltext-display.

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

[87] Kevin E Trenberth and Dennis J Shea. Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.*, 33(12), 2006. ISSN 0094-8276, 1944-8007. doi: 10.1029/2006gl026894. URL http://doi.wiley.com/10.1029/2006GL026894.

[88] Benjamin J Henley, Joelle Gergis, David J Karoly, Scott Power, John Kennedy, and Chris K Folland. A Tripole Index for the Interdecadal Pacific Oscillation. *Clim. Dyn.*, 45(11):3077–3090, December 2015. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-015-2525-1. URL https://doi.org/10.1007/s00382-015-2525-1.

[89] Kit-Yan Choi, Gabriel A Vecchi, and Andrew T Wittenberg. ENSO transition, duration, and amplitude asymmetries: Role of the nonlinear wind stress coupling in a conceptual model. *J. Clim.*, 26(23):9462–9476, December 2013. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-13-00045.1. URL https://journals.ametsoc.org/view/journals/clim/26/23/jcli-d-13-00045.1.xml.

[90] Yuko M Okumura and Clara Deser. Asymmetry in the duration of el niño and la niña. *J. Clim.*, 23(21):5826–5843, November 2010. ISSN 0894-8755, 1520-0442. doi: 10.1175/2010JCLI3592.1. URL https://journals.ametsoc.org/view/journals/clim/23/21/2010jcli3592.1.xml.

[91] Marco A Giorgetta, Johann Jungclaus, Christian H Reick, Stephanie Legutke, Jürgen Bader, Michael Böttinger, Victor Brovkin, Traute Crueger, Monika Esch, Kerstin

Fieg, Ksenia Glushak, Veronika Gayler, Helmuth Haak, Heinz-Dieter Hollweg, Tatiana Ilyina, Stefan Kinne, Luis Kornblueh, Daniela Matei, Thorsten Mauritsen, Uwe Mikolajewicz, Wolfgang Mueller, Dirk Notz, Felix Pithan, Thomas Raddatz, Sebastian Rast, Rene Redler, Erich Roeckner, Hauke Schmidt, Reiner Schnur, Joachim Segschneider, Katharina D Six, Martina Stockhause, Claudia Timmreck, Jörg Wegner, Heinrich Widmann, Karl-H Wieners, Martin Claussen, Jochem Marotzke, and Bjorn Stevens. Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the coupled model intercomparison project phase 5. *J. Adv. Model. Earth Syst.*, 5(3):572–597, July 2013. ISSN 1942-2466. doi: 10.1002/jame.20038. URL http://doi.wiley.com/10.1002/jame.20038.

[92] James W Hurrell, M M Holland, P R Gent, S Ghan, Jennifer E Kay, P J Kushner, J-F Lamarque, W G Large, D Lawrence, K Lindsay, W H Lipscomb, M C Long, N Mahowald, D R Marsh, R B Neale, P Rasch, S Vavrus, M Vertenstein, D Bader, W D Collins, J J Hack, J Kiehl, and S Marshall. The community earth system model: A framework for collaborative research. *Bull. Am. Meteorol. Soc.*, 94(9):1339–1360, September 2013. ISSN 0003-0007. doi: 10.1175/BAMS-D-12-00121.1. URL https://journals.ametsoc.org/view/journals/bams/94/9/bams-d-12-00121.1.xml.

[93] Shu Wu, Zhengyu Liu, Rong Zhang, and Thomas L Delworth. On the observed relationship between the pacific decadal oscillation and the atlantic multi-decadal oscillation. *J. Oceanogr.*, 67(1):27–35, February 2011. ISSN 0916-8370, 1573-868X. doi: 10.1007/s10872-011-0003-x. URL https://doi.org/10.1007/s10872-011-0003-x.

[94] Shih-Wei Fang, Myriam Khodri, Claudia Timmreck, Davide Zanchettin, and Johann Jungclaus. Disentangling internal and external contributions to atlantic multidecadal variability over the past millennium. *Geophys. Res. Lett.*, 48(23), December 2021. ISSN 0094-8276, 1944-8007. doi: 10.1029/2021gl095990. URL https://onlinelibrary.wiley.com/doi/10.1029/2021GL095990.

[95] T Fenske and A Clement. No internal connections detected between low frequency climate modes in north atlantic and north pacific basins. *Geophys. Res. Lett.*, 49 (5), March 2022. ISSN 0094-8276, 1944-8007. doi: 10.1029/2022gl097957. URL https://onlinelibrary.wiley.com/doi/10.1029/2022GL097957.

[96] Yochanan Kushnir, Adam A Scaife, Raymond Arritt, Gianpaolo Balsamo, George Boer, Francisco Doblas-Reyes, Ed Hawkins, Masahide Kimoto, Rupa Kumar Kolli, Arun Kumar, Daniela Matei, Katja Matthes, Wolfgang A Müller, Terence O'Kane, Judith Perlwitz, Scott Power, Marilyn Raphael, Akihiko Shimpo, Doug Smith, Matthias Tuma, and Bo Wu. Towards operational predictions of the near-term climate. *Nat. Clim. Chang.*, 9(2):94–101, January 2019. ISSN 1758-678X. doi: 10.1038/s41558-018-0359-7. URL https://www.nature.com/articles/s41558-018-0359-7.

[97] Flavio Lehner, Clara Deser, Nicola Maher, Jochem Marotzke, Erich M Fischer, Lukas Brunner, Reto Knutti, and Ed Hawkins. Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth Syst. Dyn.*, 11(2):491–508, May 2020. ISSN 2190-4979, 2190-4987. doi: 10.5194/esd-11-491-2020. URL https://esd.copernicus.org/articles/11/491/2020/.

[98] Flavio Lehner and Clara Deser. Origin, importance, and predictive limits of internal climate variability. *Environ. Res.: Climate*, 2(2):023001, May 2023. ISSN 2752-5295. doi: 10.1088/2752-5295/accf30. URL https://iopscience.iop.org/article/10.1088/2752-5295/accf30/meta.

[99] D M Smith, R Eade, A A Scaife, L-P Caron, G Danabasoglu, T M DelSole, T Delworth, F J Doblas-Reyes, N J Dunstone, L Hermanson, V Kharin, M Kimoto, W J Merryfield, T Mochizuki, W A Müller, H Pohlmann, S Yeager, and X Yang. Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science*, 2(1):1–10, May 2019. ISSN 2397-3722. doi: 10.1038/s41612-019-0071-y. URL https://www.nature.com/articles/s41612-019-0071-y.

[100] L F Borchert, V Koul, M B Menary, D J Befort, D Swingedouw, G Sgubin, and J Mignot. Skillful decadal prediction of unforced southern European summer temperature variations. *Environ. Res. Lett.*, 16(10):104017, September 2021. ISSN 1748-9326. doi: 10.1088/1748-9326/ac20f5. URL https://iopscience.iop.org/article/10.1088/1748-9326/ac20f5/meta.

[101] Keith B Rodgers, Sun-Seon Lee, Nan Rosenbloom, Axel Timmermann, Gokhan Danabasoglu, Clara Deser, Jim Edwards, Ji-Eun Kim, Isla R Simpson, Karl Stein, Malte F Stuecker, Ryohei Yamaguchi, Tamás Bódai, Eui-Seok Chung, Lei Huang, Who M Kim, Jean-François Lamarque, Danica L Lombardozzi, William R Wieder, and Stephen G Yeager. Ubiquity of human-induced changes in climate variability. *Earth Syst. Dyn.*, 12(4):1393–1411, December 2021. ISSN 2190-4979, 2190-4987. doi: 10.5194/esd-12-1393-2021. URL https://doi.org/10.5194/esd-12-1393-2021.

[102] Brian C O'Neill, Claudia Tebaldi, Detlef P van Vuuren, Veronika Eyring, Pierre Friedlingstein, George Hurtt, Reto Knutti, Elmar Kriegler, Jean-Francois Lamarque, Jason Lowe, Gerald A Meehl, Richard Moss, Keywan Riahi, and Benjamin M Sanderson. The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geosci. Model Dev.*, 9(9):3461–3482, September 2016. ISSN 1991-959X, 1991-9603. doi: 10.5194/gmd-9-3461-2016. URL https://gmd.copernicus.org/articles/9/3461/2016/.

[103] J T Fasullo, Jean-Francois Lamarque, Cecile Hannay, Nan Rosenbloom, Simone Tilmes, Patricia DeRepentigny, Alexandra Jahn, and Clara Deser. Spurious late historical-era warming in CESM2 driven by prescribed biomass burning emissions. *Geophys. Res. Lett.*, 49(2), January 2022. ISSN 0094-8276, 1944-8007. doi: 10.1029/2021gl097420. URL https://onlinelibrary.wiley.com/doi/10.1029/2021GL097420.

[104] Leo Breiman. Random Forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125, 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

[105] Yann Planton, Jérôme Vialard, Eric Guilyardi, Matthieu Lengaigne, and Takeshi Izumo. Western pacific oceanic heat content: A better predictor of la Niña than of El Niño. *Geophys. Res. Lett.*, 45(18):9824–9833, September 2018. ISSN 0094-8276, 1944-8007. doi: 10.1029/2018gl079341. URL https://onlinelibrary.wiley.com/doi/10.1029/2018GL079341.

[106] Justin S Mankin, Flavio Lehner, Sloan Coats, and Karen A McKinnon. The value of initial condition large ensembles to robust adaptation decision-making. *Earth's Future*, 8(10), October 2020. ISSN 2328-4277. doi: 10.1029/2020ef001610. URL https://onlinelibrary.wiley.com/doi/10.1029/2020EF001610.

[107] Noah S Diffenbaugh, Elizabeth A Barnes, and Patrick W Keys. Probability of continued local-scale warming and extreme events during and after decarbonization. *Environ. Res.: Climate*, 2(2):021003, May 2023. ISSN 2752-5295. doi: 10.1088/2752-5295/accf2f. URL https://iopscience.iop.org/article/10.1088/2752-5295/accf2f/meta.

[108] Wanyun Shao, James C Garand, Barry D Keim, and Lawrence C Hamilton. Science, scientists, and local weather: Understanding mass perceptions of global warming. *Soc. Sci. Q.*, 97(5):1023–1057, November 2016. ISSN 0038-4941, 1540-6237. doi: 10.1111/ssqu.12317. URL https://onlinelibrary.wiley.com/doi/10.1111/ssqu.12317.

[109] Gerald A Meehl, Haiyan Teng, Doug Smith, Stephen Yeager, William Merryfield, Francisco Doblas-Reyes, and Anne A Glanville. The effects of bias, drift, and trends in calculating anomalies for evaluating skill of seasonal-to-decadal initialized climate predictions. *Clim. Dyn.*, April 2022. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-022-06272-7. URL https://doi.org/10.1007/s00382-022-06272-7.

[110] Marybeth Arcodia, Elizabeth A Barnes, Kirsten Mayer, Jiwoo Lee, Ana Ordonez, and Min-Seop Ahn. Assessing decadal variability of subseasonal forecasts of opportunity using explainable AI, June 2023. URL https://eartharxiv.org/repository/view/5479/.

[111] Stephen G Yeager, Ping Chang, Gokhan Danabasoglu, Nan Rosenbloom, Qiuying Zhang, Fred S Castruccio, Abishek Gopal, M Cameron Rencurrel, and Isla R Simpson. Reduced Southern Ocean warming enhances global skill and signal-to-noise in an eddy-resolving decadal prediction system. *npj Climate and Atmospheric Science*, 6(1): 1–13, July 2023. ISSN 2397-3722, 2397-3722. doi: 10.1038/s41612-023-00434-y. URL https://www.nature.com/articles/s41612-023-00434-y.

[112] Antonios Mamalakis, Elizabeth A Barnes, and Imme Ebert-Uphoff. Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience. *Artificial Intelligence for the Earth Systems*, 1(4), October 2022. ISSN 2769-7525. doi: 10.1175/AIES-D-22-0012.1. URL https://journals.ametsoc.org/view/journals/aies/1/4/AIES-D-22-0012.1.xml.

Appendix A: Supporting Information for Chapter 2

A.1 Introduction

Here we provide a short overview of neural networks, along with the specifications of the artificial neural network (ANN) used in this study. We also describe the rationale behind the choice of a 30 month lead time followed by various statistics of the three ANNs used. Lastly we include supplementary figures to support our discussion and conclusions.

A.2 Neural Network Overview

A general description of an artificial neural network (ANN) is thus: the neural network learns from some training data to map an input to some output, with hidden weights and connections optimized in the training process, and an activation function which allows for non-linearities. The network is trained for a set number of passes though the training data (called epochs), updating hidden weights based on minimizing the so-called loss function. The ANN architecture and training procedure in this study has been optimized for the specific problem that we consider. The use of regularization, dropout layers, training epoch and sample weights were carefully chosen to balance accuracy, but prevent over-fitting. Values used are included in Table S1. A more in-depth description of ANNs, as well as a broad background on their application to climate studies can be found in Toms et al. [32].

A.3 Rationale behind 30 month lead time

Our ANN learns to predict whether a PDO phase transition will occur within some cut-off time. Consider an input such that by the time of the output, a transition has occurred (i.e. the true output is 1). If, for example, the lead time is 30 months and the transition occurred 29 months after the input, then this would be classified transition however it would be difficult for the ANN to guess as it is similar to inputs where transitions occur at 31 months (which are classified persistence). The accuracy of the ANN dramatically

decreases for samples where the transition occurs within around 3 months of the lead time. On the other hand, we want to focus on transitions occurring at least 12 months after input in order to benchmark our networks against previous work. Hence, in order to optimize for the accuracy of samples with transitions at least 12 months after input, retain good general accuracy, and a reasonable cut-off for recognizing persistence, we choose a lead time of 30 months (2.5 years).

A.4 Summary of the 'best' neural networks

In order to find the best models for our problem setup we have trained 60 neural networks of the identical architecture, each with a different random seed. Note this seed is the same for both initializing the neural network and for choosing the transition samples to grab from the training/validation data. We train many models because we do not use all of the available data in the training process. This, along with the inherent randomness in the ANN training process can result in variation in the ANNs's accuracy. The random seed is set and recorded before the training/validation data is selected and the model is trained.

In Figure A.1 we show various statistics of each individual neural network. The left panel compares the total accuracy of each ANN (x axis) with its persistence recall (percentage of the time that when persistence occurs, the ANN guesses persistence, y axis). This plot shows the difficulty in guessing persistence for this particular problem, with no ANNs above 56% recall. We comment on the reason for this in Chapter 2. As persistence appears to be more difficult for the ANNs to learn, we designate the 'best' ANNs as those that combine high accuracy and high persistence recall. These are indicated in each plot by the pink dots.

The right panel demonstrates the ANNs's ability to predict transitions that occur 12-27 months after input, with total accuracy on the x axis and 12-27 month transition recall (percentage of the time that when a transition occurs 12-27 months after input, the ANN predicts the transition) on the y axis. This shows that the NNs we have designated as the

'best' (again in pink dots) have recall of 12-27 month transitions of around 65%-72%. While these are not the best ANNs for this task in particular, we choose them for this study as they are the best at *both* persistence and transitions, with their recall implying they have learned both, and are least likely to be over-fit.

In Figure A.2 we show the confusion matrices for the best three ANNs described above. These demonstrate how the ANNs perform at the classification task on the validation data (1110 samples; 555 persistence, 555 transitions). Each row is the actual class the samples belong to, while the columns show how the ANN designated them, i.e. the top row are samples that are *true* persistence while the left column is the samples that were *predicted* as persistence. This means the main diagonal is where the ANN was correct and the off-diagonal is where the ANN was wrong. The number in each box is the number of samples placed in that category e.g. the top left box is number of samples with actual persistence *and* the ANN predicted persistence, whereas the bottom left is where an actual transition occurred but the ANN predicted persistence. In all cases, the ANNs were better at correctly predicting transitions than persistence while the largest source of inaccuracy is due to the ANNs predicting transitions when the true class is persistence.

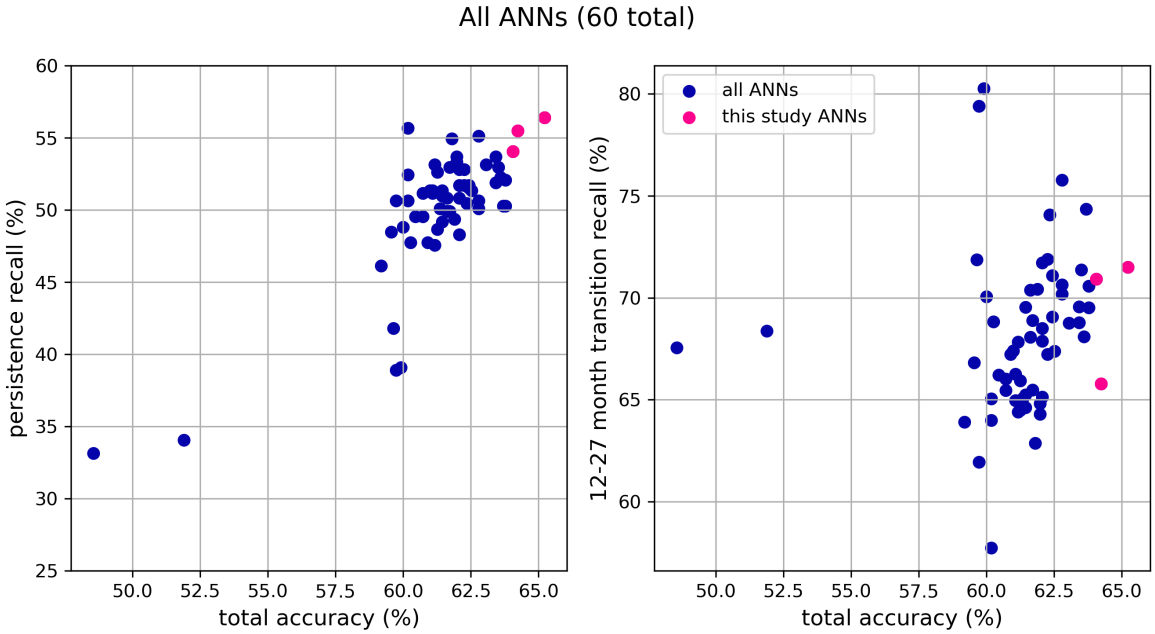

Figure A.1: (left) Comparison of total accuracy (horizontal) and persistence recall (vertical) for all ANNs trained. Blue dots are all ANNs with pink dots representing the ANNs used in the study. (right) Comparison of total accuracy (horizontal) and 12-27 month transition recall.

Table A.1: Table of neural network specifications and accuracy for the ANNs used in this study.

Input	3 deseasoned and standardized $4° \times 4°$ OHC grids, 4 months apart
Architecture	3 vectorized OHC grids (12150 pixels total) connected to a single hidden layer with 8 nodes and rectified linear unit (ReLU) activation function, then connected to 2 output nodes representing positive and negative phase prediction with softmax activation to normalize outputs to probabilities.
Training	L2 regularization coefficient of 12 and dropout of one node per epoch on hidden layer. Adam optimization algorithm, with initial learning rate of 10^{-3}, dropping by a factor of 2 every 25 epochs. Trained for 300 epochs total. Categorical cross entropy loss function. First 1800 years (21600 samples) used for training, latter 200 years (2400 samples) used for validation (see Chapter 2).
Output	Prediction of whether PDO transition occurs within 30 months of last input map.

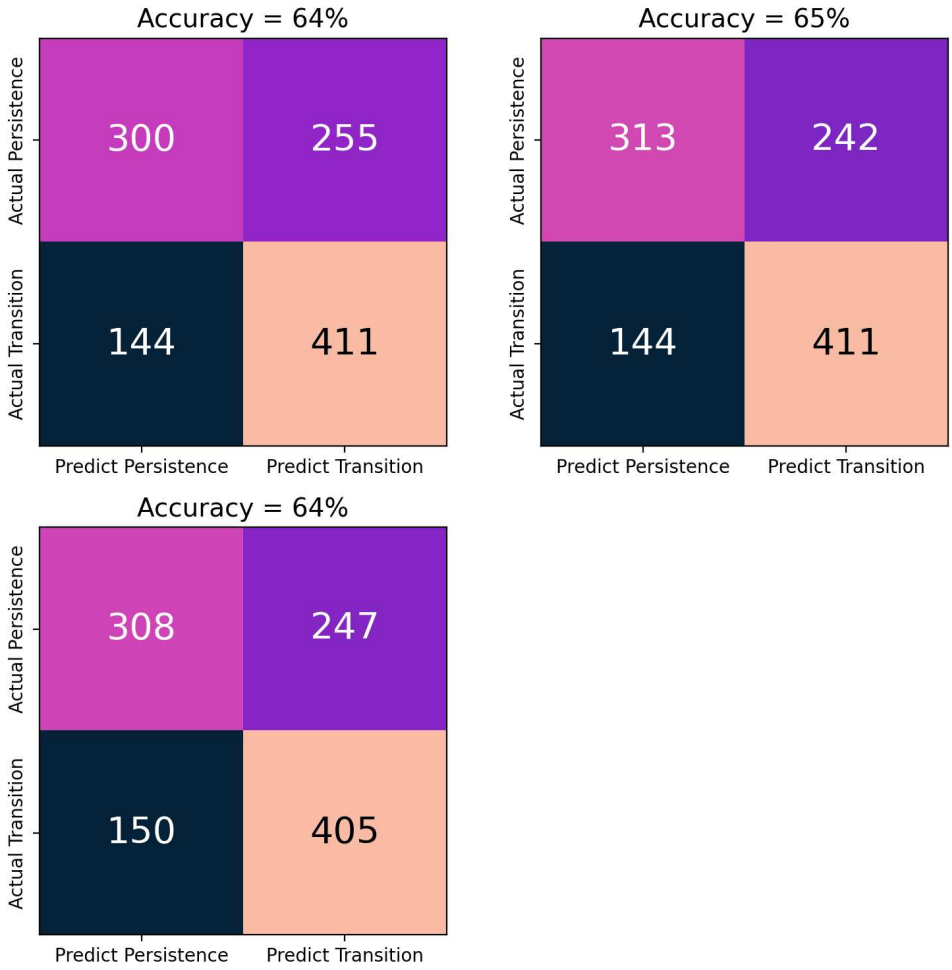

Figure A.2: Confusion matrices for the 3 models used in this study. Vertical axis is the actual class and horizontal axis is the predicted class. Number of samples in each bin is printed in each square and total accuracy of each ANN in the title.

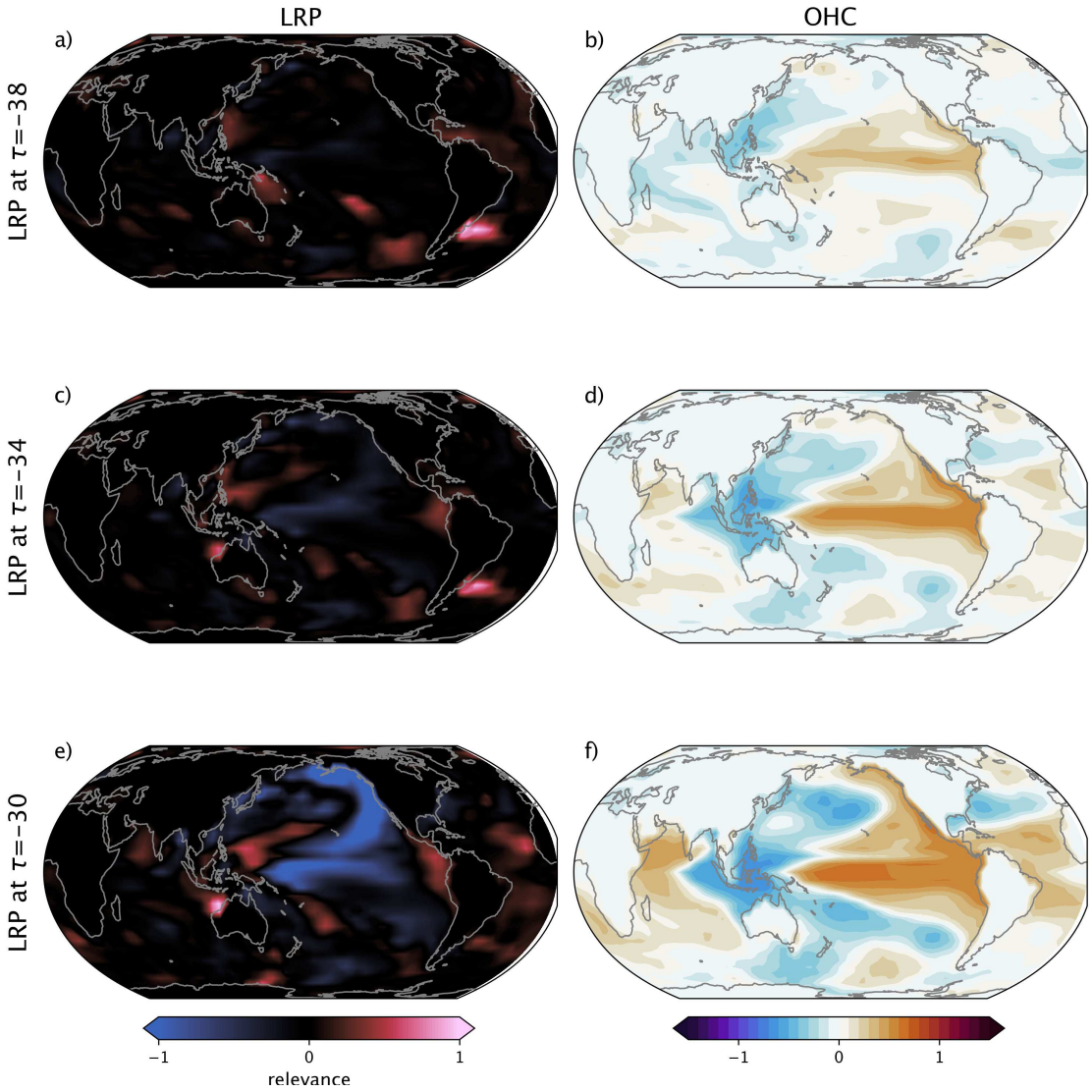

Figure A.3: Left column: composite LRP maps for input maps where model correctly guesses transition from positive to negative occurs 12-27 months after final input. a) 38 months before output, c) 34 months before output, e) 30 months before output (and panel a in Figure 2.4). Right column: As left column but for composite OHC anomaly, with units of standard deviation at each grid point and color scale as in Figure 2.4.

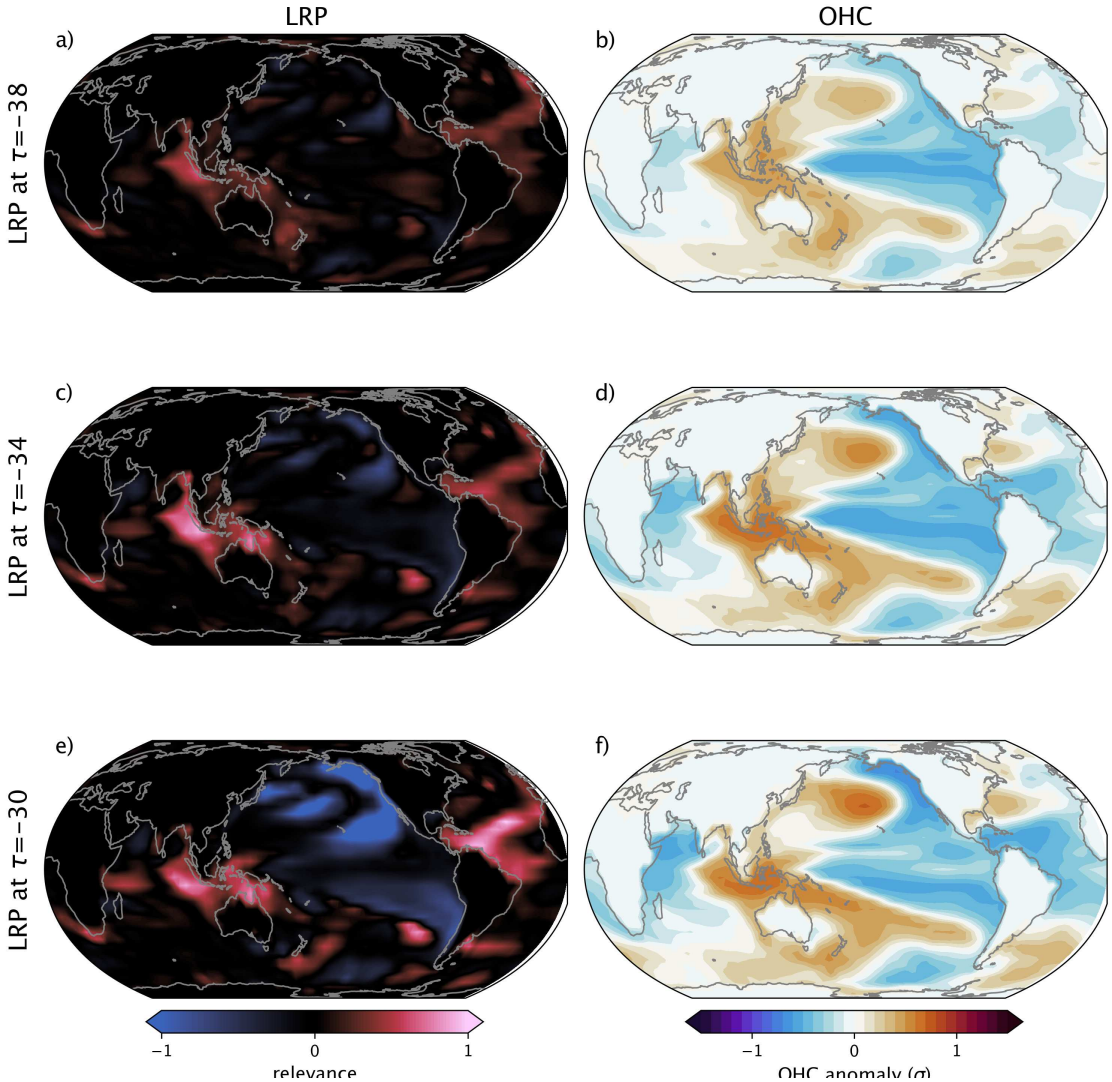

Figure A.4: As Figure A.3 but for negative to positive transitions.

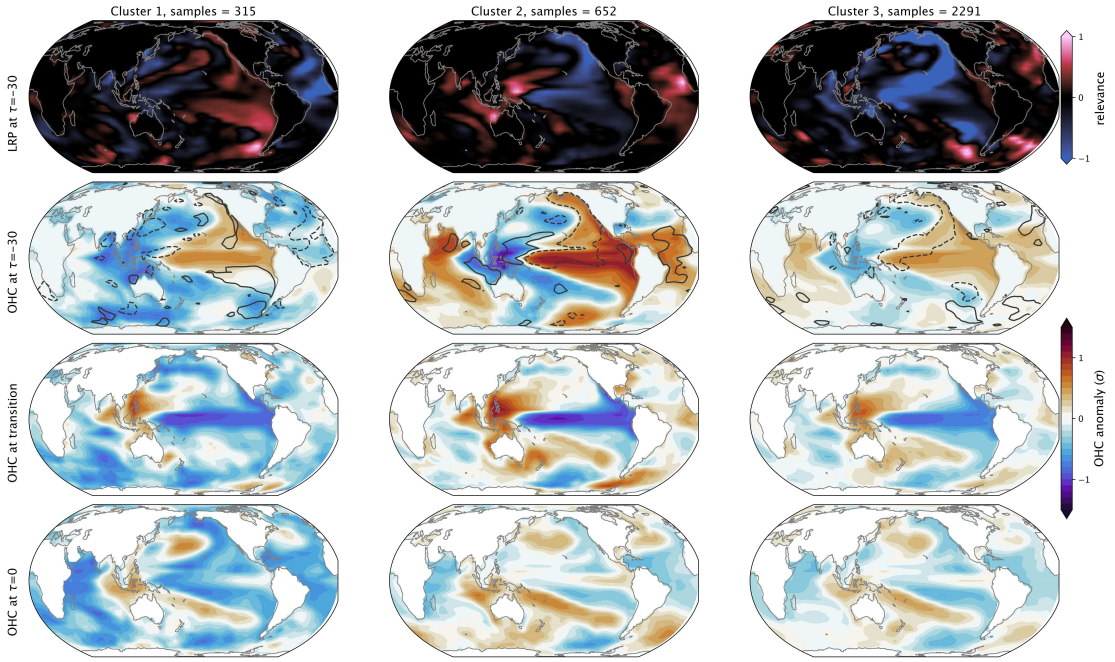

Figure A.5: K-means of LRP maps when model correctly predicts positive to negative transition 12-27 months after input. Each column represents a cluster. Top row is LRP maps at month $\tau = -30$, second row is corresponding OHC with top and bottom 5% from the LRP contoured (dashed and dotted respectively as in Figure 2.4). The third row is OHC at the transition while the bottom row is OHC at month $\tau = 0$.

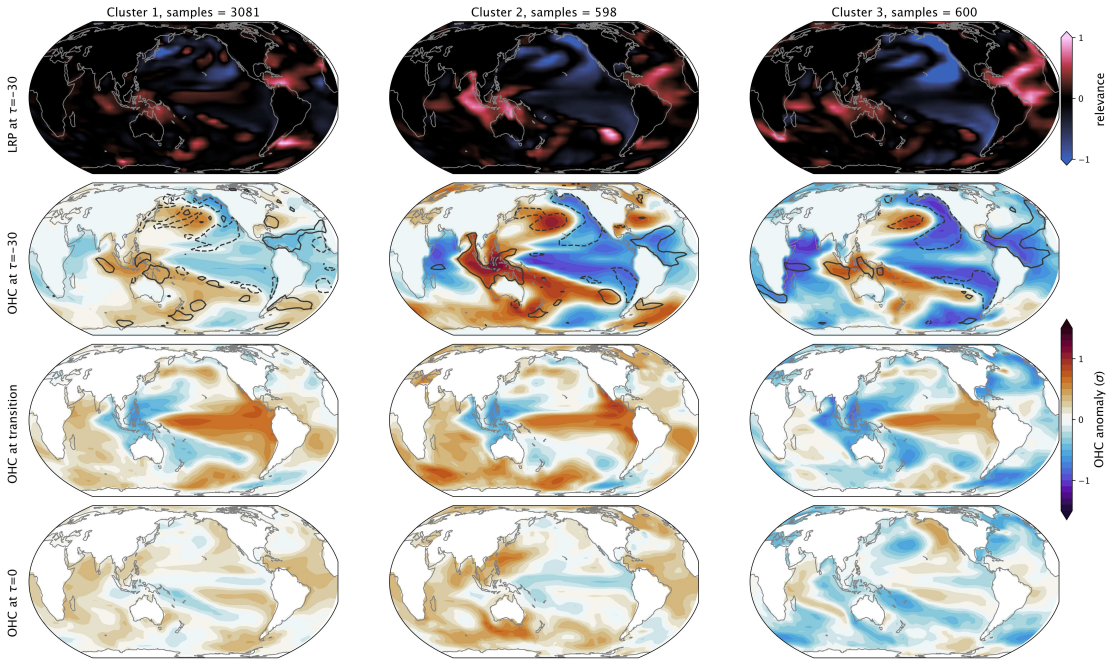

Figure A.6: As Figure A.5 but for negative to positive transitions

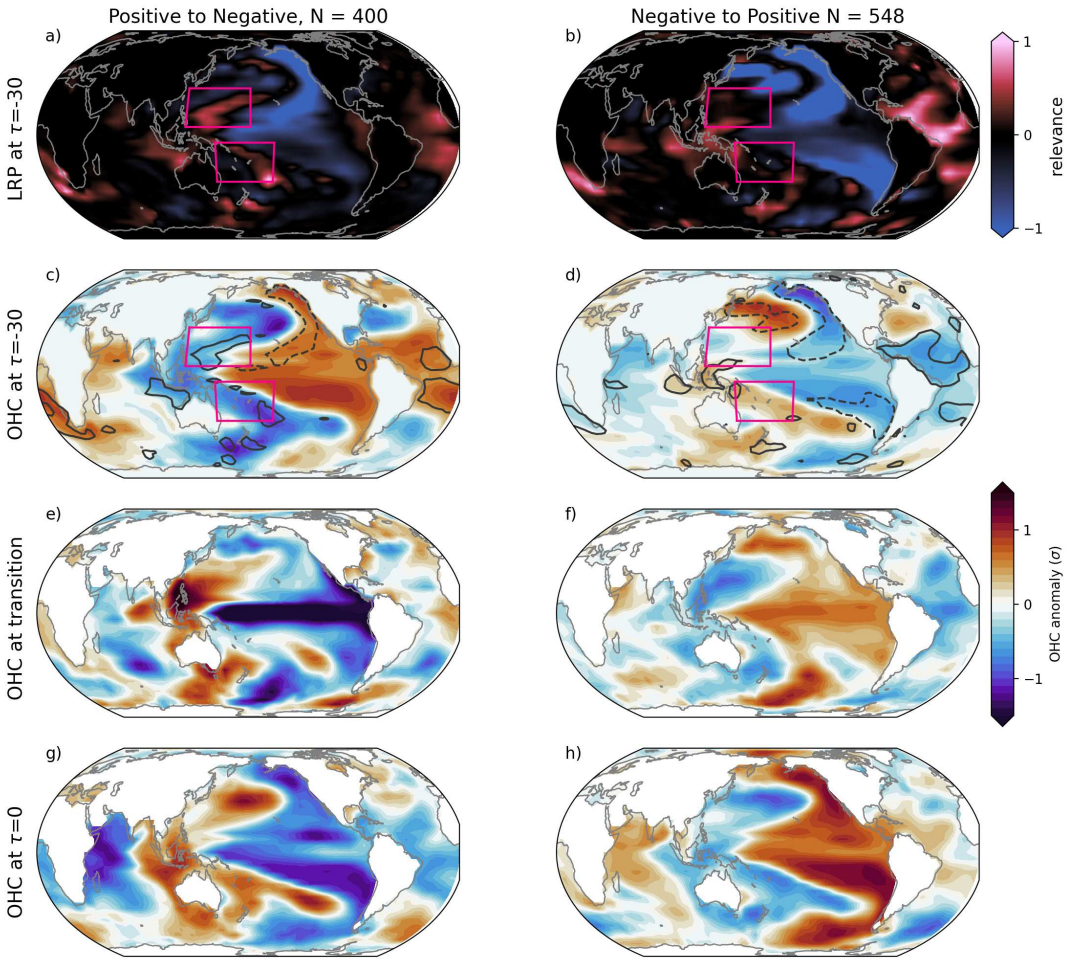

Figure A.7: As Figure 2.4 but only for correct transition predictions where the PDO phase length preceding AND following a PDO transition are > 30 months.

Appendix B: Supporting Information for Chapter 3

B.1 Introduction

The text in this document (section B.3) is a description of explainable AI (XAI), and provides a discussion of XAI findings which support the conclusions in Chapter 3. This text references Figures B.4 and B.5 which are the XAI analyses of Figures 3.3 and 3.4, respectively. In Figure B.1 we provide plots showing the Atlantic multi-decadal variability (AMV) and interdecadal Pacific oscillation (IPO) patterns calculated in the CESM2 long control run.

B.2 IPO Index Calculation

We calculate the IPO index using the method outlined by Henley et al. [88] and we detail here. From the deseasoned SST data we calculate the area averaged monthly SST anomalies in three boxes in the Pacific Ocean:

1. 25°N to 45°N and 140°E to 145°W
2. 10°S to 10°N and 170°E to 90°W
3. 50°S to 15°S and 150°E to 160°W

Using the numbering above, the index is calculated from the following equation:

$$\mathrm{IPO} = \mathrm{Box2} - 0.5 * (\mathrm{Box1} + \mathrm{Box3}) \tag{B.1}$$

The resulting pattern from projecting the IPO index onto global SSTs is plotted in Figure B.1, with the boxes in these calculations outlined in purple.

B.3 Neural Network Explainability

To support our results, we use neural network explainability techniques (explainable AI or XAI) to examine the decision-making process of the ANNs. The underlying goal of

the XAI methods used here is to provide an indication of how each input pixel contributed to a neural network's prediction. The methods we use here are attribution methods, in particular we use three methods, Integrated Gradient, LRP-Z (which is the same as Input times Gradient for networks with ReLU activation) and LRP-epsilon. All of these methods assign each input pixel a relevance, where positive relevance indicates that a pixel contributed to positively to an output node of interest and vice versa. For comprehensive discussion of XAI with application to climate science, and best practices, see Mamalakis et al. [61] and Mamalakis et al. [112].

The explainability composite maps for each region investigated in Chapter 3 is provided in Figure B.4-B.5. Each of the first three columns is a different method (Gradient, Input times Gradient, LRP-epsilon from left to right). We use an epsilon value of 0.01, and apply Gaussian smoothing to each explainability map to assist with visualization. Each row is a different OHC level (OHC to 100 m, OHC to 300 m, OHC to 700 m from top to bottom). The right-most column in each is the composite OHC input which acts as a reference to how the relevance patterns correspond to the physical input maps.

In Figure B.4 we look at the composite explainability maps for confident predictions of positive SST anomaly in the North Atlantic ocean (green dot, same as in Figure 3.3). For all three methods, red regions contributed to the neural network's positive prediction. It appears the positive OHC anomaly in the North Atlantic Ocean contributed to the positive SST prediction, especially at the lowest level of the ocean (OHC to 700m). All XAI methods show the same patterns, reducing the likelihood for spurious relevance (although not eliminating it, see [61]).

In Figure B.5 we look at composite explainability maps for confident predictions of negative SST anomaly in the North Pacific ocean (green dot, same as Figure 3.4). Here, the blue regions imply regions that contributed to neural network's negative prediction. Here, relevance highlights that the negative anomaly in the Kuroshio region in the upper layers, coupled with the positive anomaly in the off equatorial Pacific in lowest layers

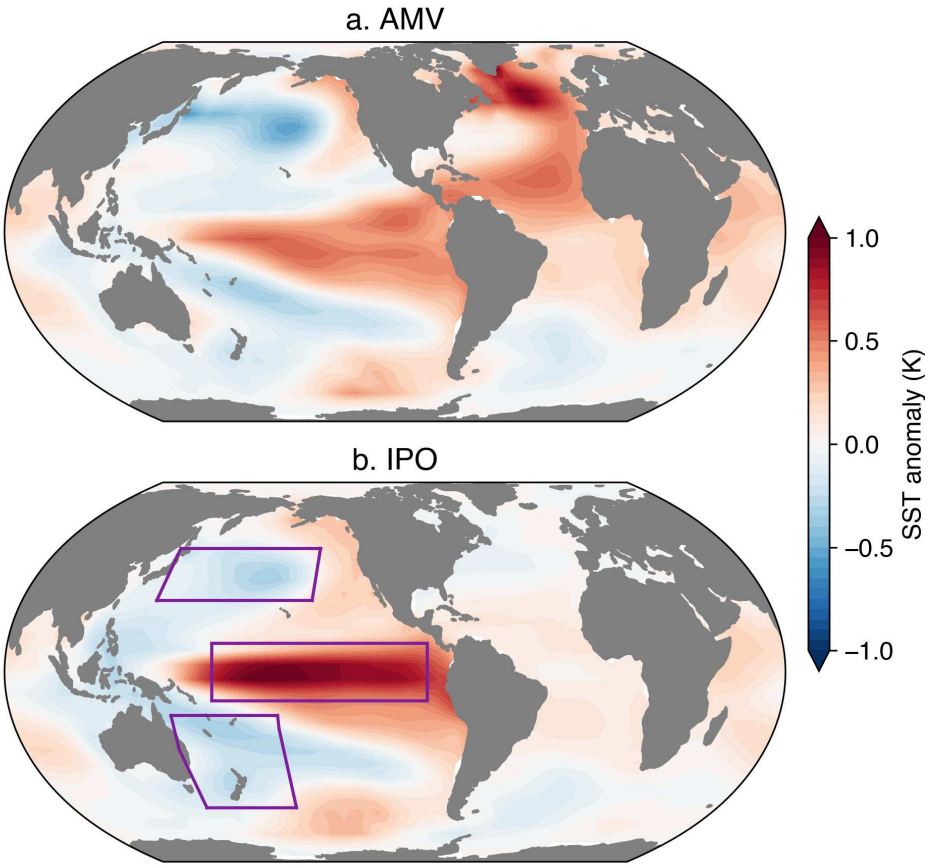

Figure B.1: Patterns of large scale SST variability in CESM2 calculated using the methods discussed in Section 3.2.3. AMV index projected onto global SSTs. b. IPO index projected onto global SSTs.

most contributed to the negative prediction. This anomaly pattern is indicative of the IPO's positive phase. Again the highlighted relevances are consistent across explainability methods.

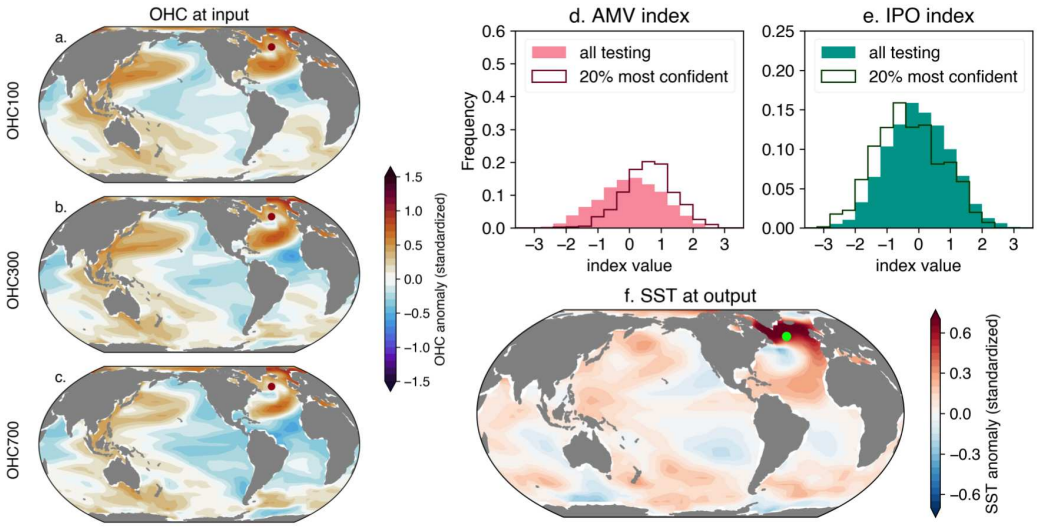

Figure B.2: As Figure 3.3 but for the training and validation data.

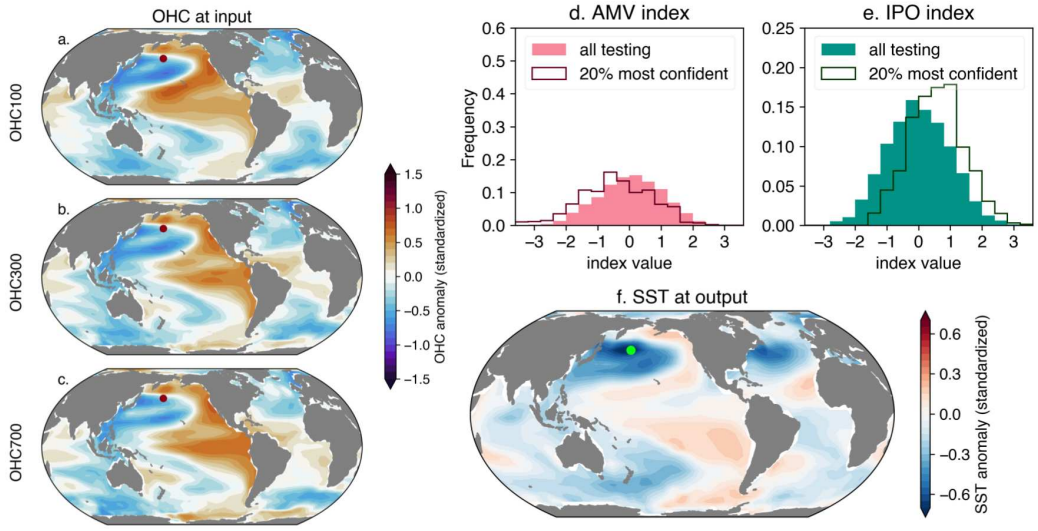

Figure B.3: As Figure 3.4 but for the training and validation data.

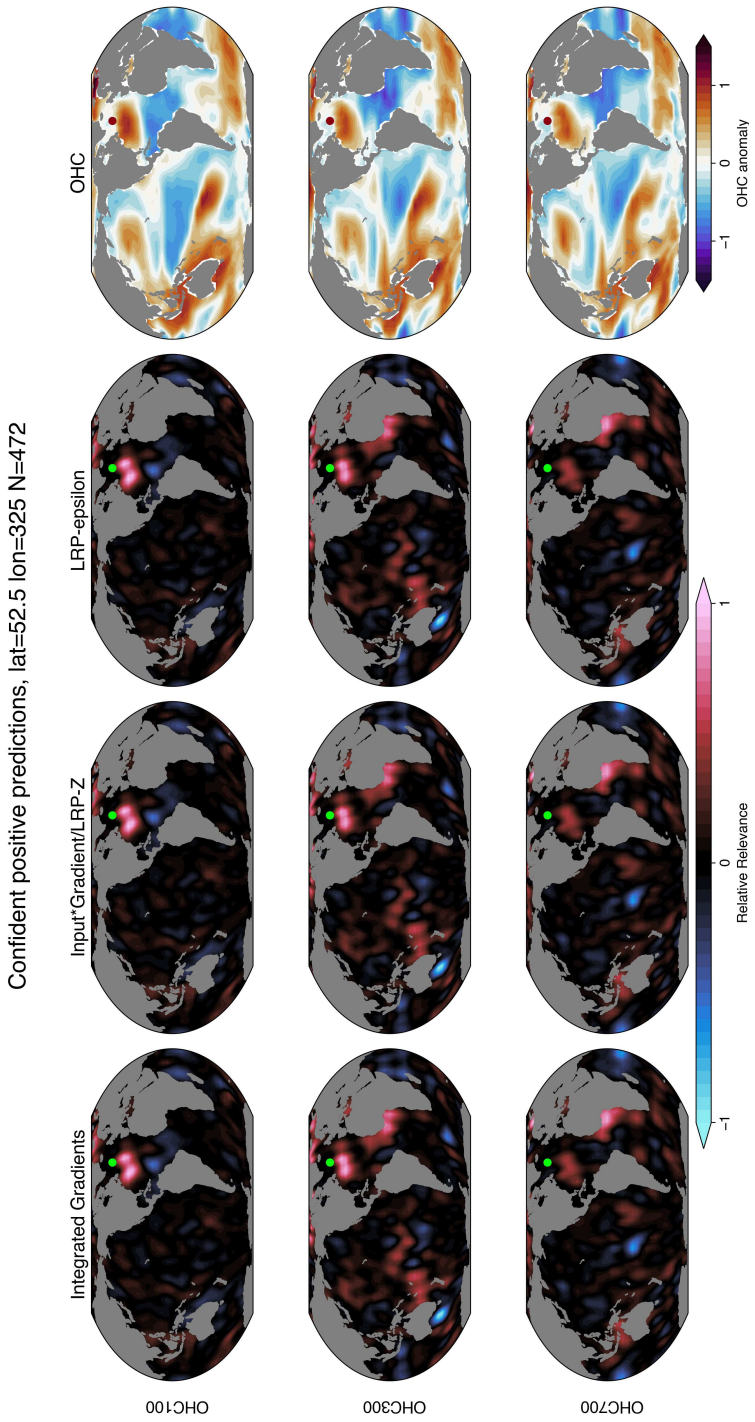

Figure B.4: Composite explainability maps for predictions in Figure 3.3. Each of the first three columns is a different technique (Integrated Gradients, Input times Gradient, LRP-epsilon from left to right), while each row is a different ocean layer (OHC to 100 m, OHC to 300 m, OHC to 700 m from top to bottom). The right-most column is the composite OHC input (the same as Fig 3.3a-c).

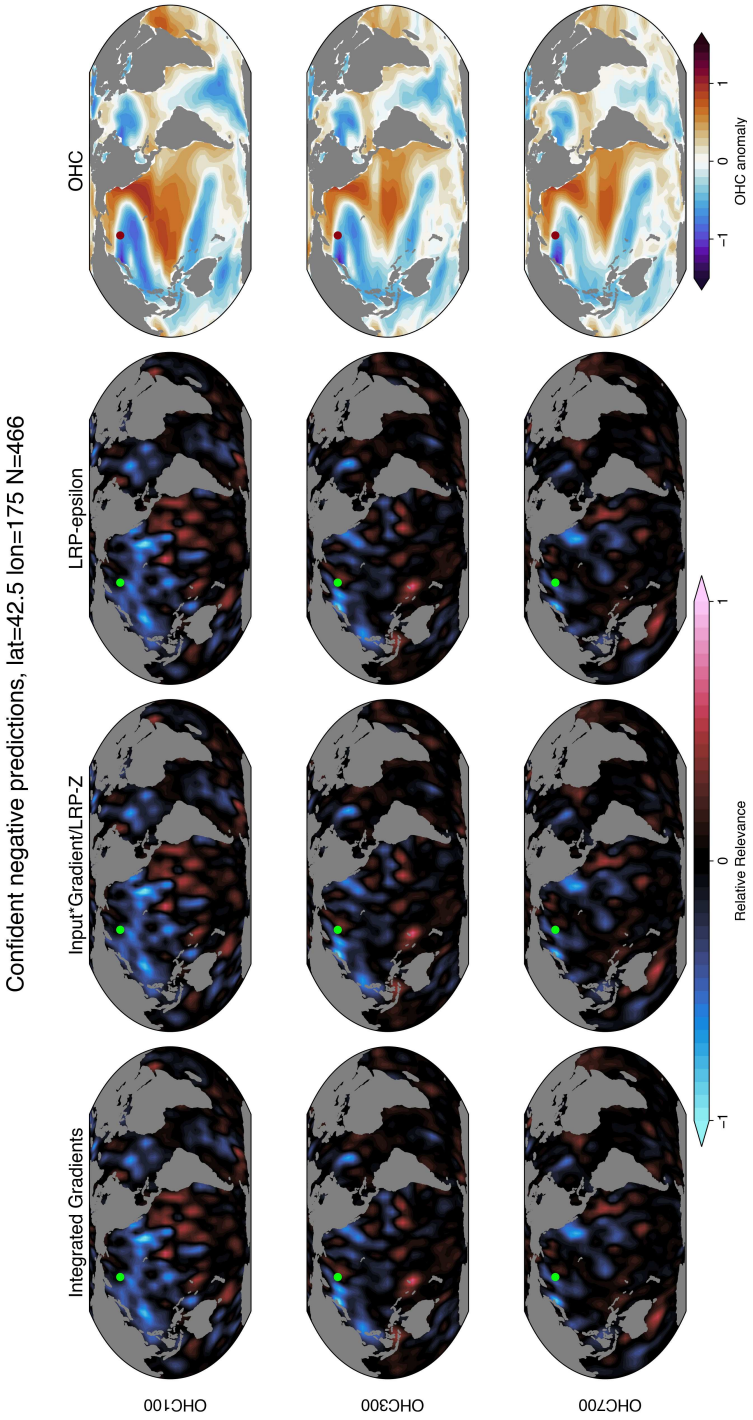

Figure B.5: As Figure B.4 but for North Pacific predictions in Figure 3.4.

Appendix C: Supporting Information for Chapter 4

C.1 Detailed Information About Training ANNs

The IV_Network is two SST maps (5184 pixels) densely connected to a hidden layer of 20 nodes, followed by a hidden layer of 40 nodes. This is densely connected to the output layer of 3 nodes. The hidden layers use rectified linear unit (ReLU) activation while the final layer has no activation. We also remove all bias terms from the IV_Network because this forces a zero input to be mapped to zero output (as we define zero IV input to be zero anomaly from the forced response). The EF_Network consists of an input map of the forced response (2592 pixels) connected to a single hidden layer of 10 nodes with ReLU activation, which is connected to an output layer of three nodes with no activation. The architectures differ between the two networks because we expect the internal variability task to be more complicated, whereas there is limited predictability that can be garnered from the forced response.

The outputs of each network are summed pairwise with no bias term and after summation, the softmax activation is added to convert the predictions to probability-like. We call the fully coupled system the Combined Network. The Combined Network is trained with categorical cross-entropy loss on the final predictions so that each component network must learn to correctly scale its contribution to the output. During training, we implement dropout at a rate of 0.2 between the hidden layers in the IV_Network. We also use ridge regularization of 0.01 on the inputs of the EF_Network and 0.0001 on the inputs to the IV_Network. The Combined Network is trained with Adam optimizer and an initial learning rate of 0.1 which steps down to 0.01 after 30 epochs and 0.001 after a further 30 epochs.

We use class weighting during training because we train on the full 1960-2100 period but the terciles are chosen from 2020-2050, meaning that the classes are not necessarily balanced in the training data. We weight each class by its relative size in the training data

i.e. if the number of samples in each of the classes is 100/95/95 then they are weighted 1/1.05/1.05. If the sample weight is greater than 1.1 in this formulation then it is reduced by 10% to prevent any sample weights from becoming too large, e.g. if the classes are 100/95/80 then sample weighting would be 1/1.05/1.13 instead of 1/1.05/1.25 (since 1.25 $\times$ 0.9 is 1.13). This was found to be effective to prevent the ANNs from defaulting to a single class on unbalanced data. The neural networks are trained for a maximum of 1000 epochs with early stopping if the validation loss does not decrease for 100 epochs. This combination of inputs, outputs, architecture, and hyperparameters was found to perform flexibly across all grid boxes.

C.2 Example ANN predictions

We provide three examples of ANN predictions in Figure C.1. These predictions are all real predictions made by networks used in the study. All three examples are correct predictions. In Figure C.1a, the IV_Network and EF_Network both predict the upper tercile node, by assigning the highest weight to this node. After the softmax activation (which conserves relative ranking) the upper tercile is therefore the prediction of the ANN.

In Figure C.1b, the IV_Network and EF_Network make different predictions, with the IV_Network predicting the lower tercile, and the EF_Network predicting the upper tercile. However, once the networks are combined and softmax applied, the lower tercile has the highest weight meaning that it is the Combined Network's prediction. In this case, the inclusion internal variability was the reason the Combined Network made the correct prediction, demonstrating how internal variability can provide future predictability, and these networks can be used to examine this.

Lastly (and most complexly, bear with us), in Figure C.1c, the IV_Network and EF_Network make different predictions. This time however, the combinations work against each other to such an extent that after summation neither is the highest node, but instead the middle tercile is the winning prediction. In this example, the no single component

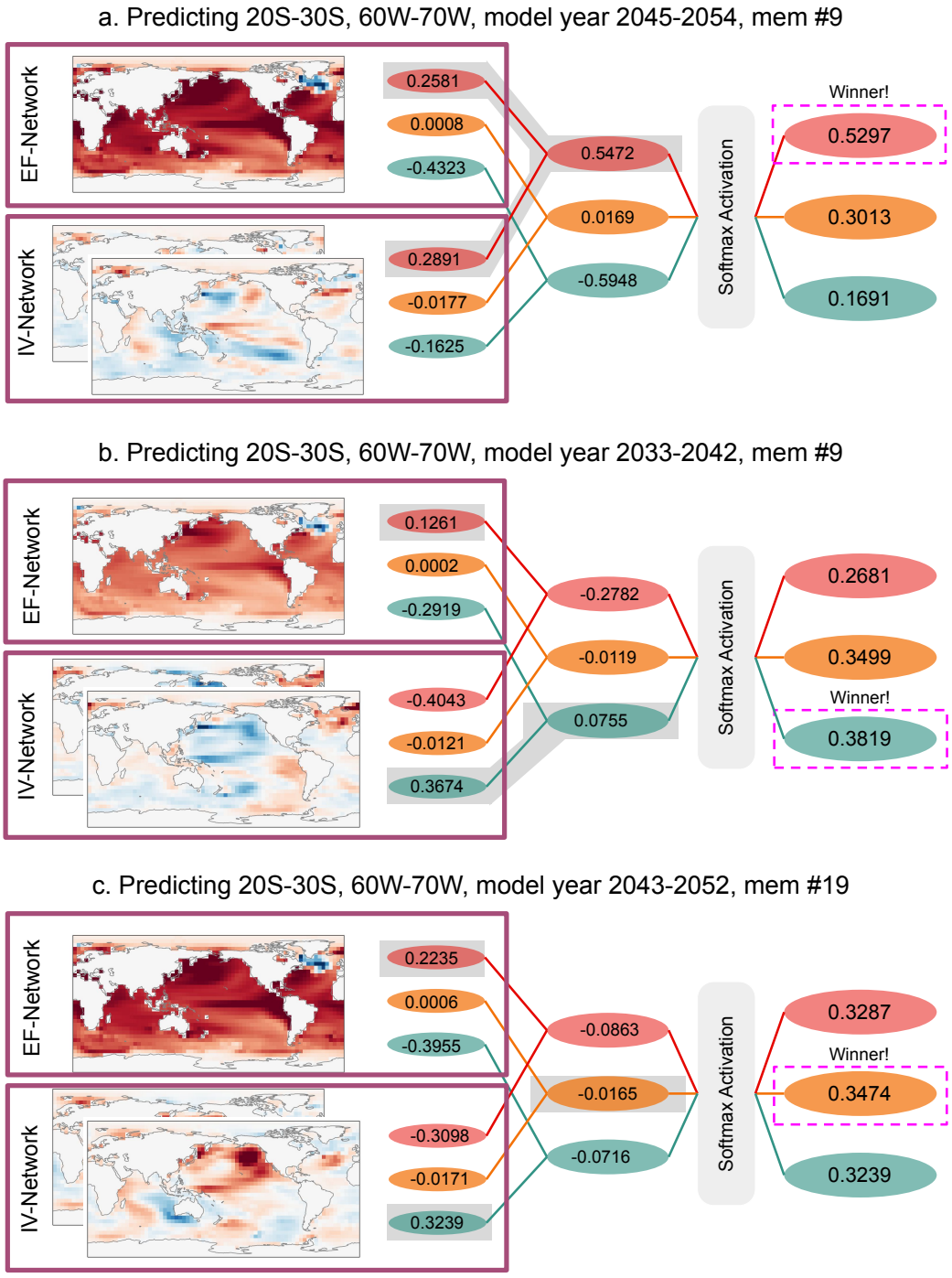

Figure C.1: Schematic of three example predictions of future SST trend at 20S-30S 60W-70W. Each panel demonstrates input maps, their predictions by the individual neural networks, and the final prediction before and after the softmax activation is added for actual samples from the testing set. Prediction year and ensemble member is given in each panel title. Grey shading indicates the winning prediction at each step in the neural networks. Pink box indicates the final output by the Combined Network.

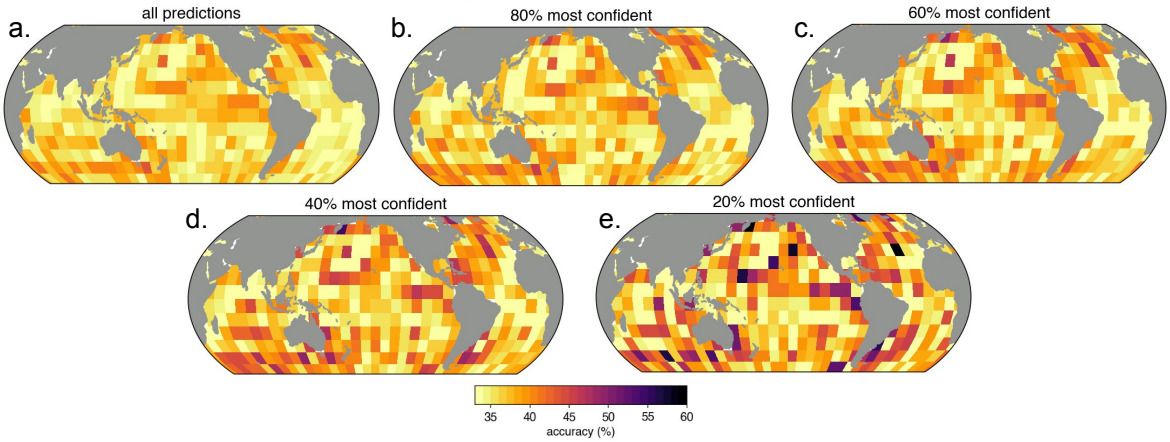

Figure C.2: Combined Network accuracy on testing data for the full training set and 4 different confidence levels (see panel headings).

network made the correct prediction, demonstrating the power of this technique to make synergistic predictions of the future.

C.3 ANN confidence levels

Figure C.2 demonstrates how neural network accuracy generally increases with increasing confidence in prediction, implying that the ANNs have learned which samples are more predictable than others. In Chapter 4 we include only the accuracy across all samples (Figure 2a = Figure C.2a) and the accuracy for the 20% most confident (Figure 2b = Figure C.2e). Here show the accuracy for three additional confidence cut-offs (80%, 60%, 40% = Fig. C.2b, Fig. C.2c, Fig. C.2d). This indicates that while 80% was chosen in Chapter 4 as the cut-off, other confidence levels can be used to show the existence of windows of opportunity.

C.4 K-means Robustness

Here we demonstrate several key metrics for checking the robustness of our K-means clustering in Section 3.2 . Firstly, we note that the optimal choice of number of clusters can be tricky and vary based on the random state of the K-means algorithm. One way of deciding on choice of cluster number is using the silhouette score, which maximizes

for the optimal number of clusters. Figure C.3a demonstrates the silhouette score for initial state used in our cluster analysis, showing that six is the optimal number of clusters. Secondly, we note that although six is found to be best for the combination of the data and random state, this does not necessarily mean that all data points are well described by the learned centroids. We therefore plot both the distance from the cluster centroid at each grid point (Figure C.3b), and correlation coefficient with the cluster centroid with each grid point (Figure C.3). In both panels, lighter colors indicate grid points which are best described to the cluster center, either by closest distance (Figure C.3b) or highest correlation (Figure C.3c). This demonstrates that the general regions we choose for close analysis in Chapter 4 (North Pacific, North Atlantic, and Tropical Pacific) are also best described by their cluster centroids.

C.5 Composite Nino3.4, PDO and North Atlantic subpolar gyre evolution

We plot some climate indices to relate states of internal variability in Figure 4 to large scale variability. We calculate the PDO index (Figure C.4a and C.4d) by first removing the ensemble mean from each member. We then calculate the first EOF of North Pacific (20N-60N, 110E-120W) annual mean SST for the entire ensemble for the entire run (1850-2100). The corresponding principal component timeseries is defined as the PDO index. We define the North Atlantic subpolar gyre SST (Figure C.4b and C.4e) as the area weighted annual mean SST from 40N-60N, 80W-20W with the ensemble mean removed. Nino3.4 (Figure C.4c and C.4f) is calculated as the area weighted annual mean SST over 5N-5S, 170W-120W taken from each ensemble member with the ensemble mean removed.

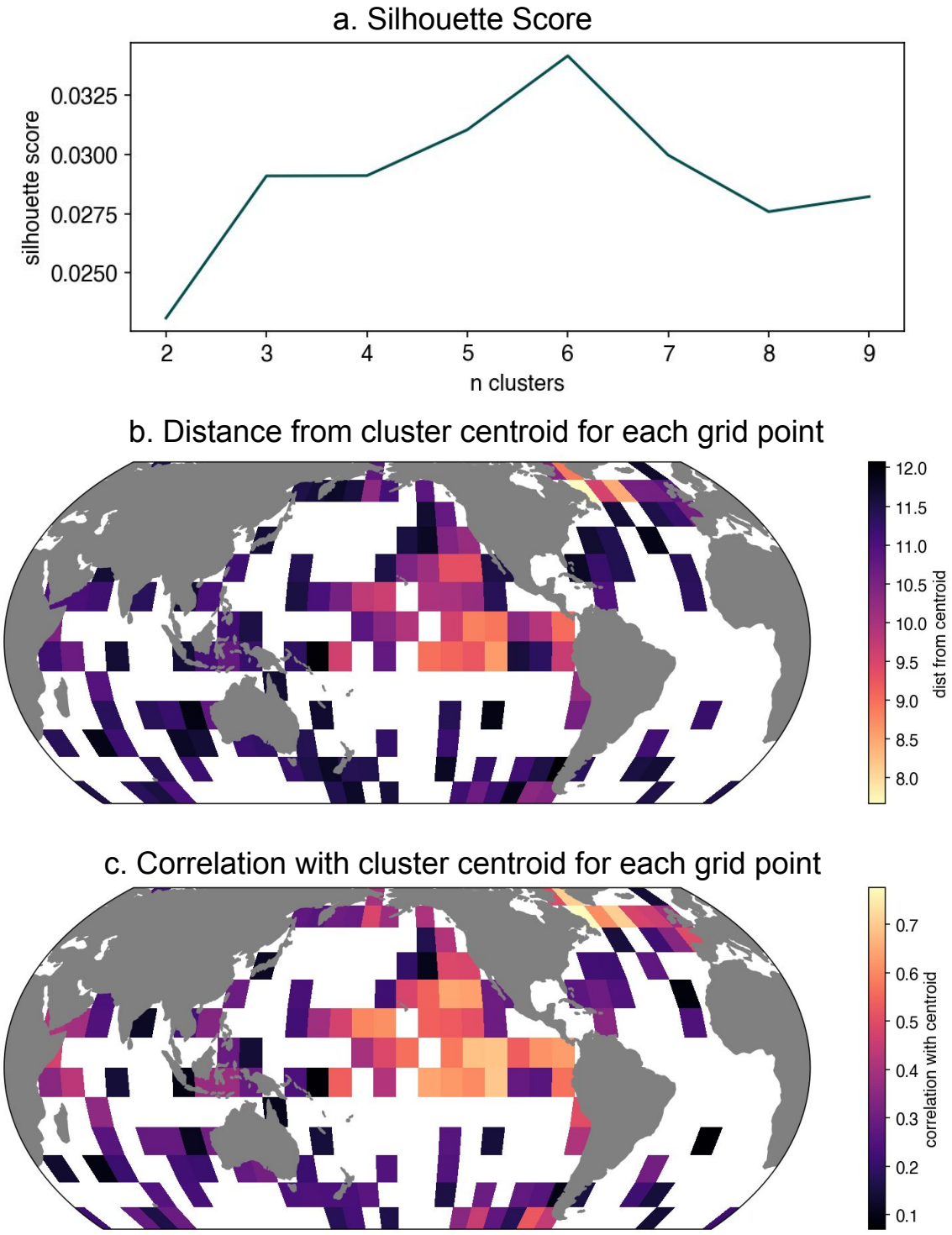

Figure C.3: (a) Silhouette score for 2-9 clusters. (b) Euclidean distance from the applicable centroid at each grid point. (c) Pearson correlation coefficient with applicable centroid at each grid point.

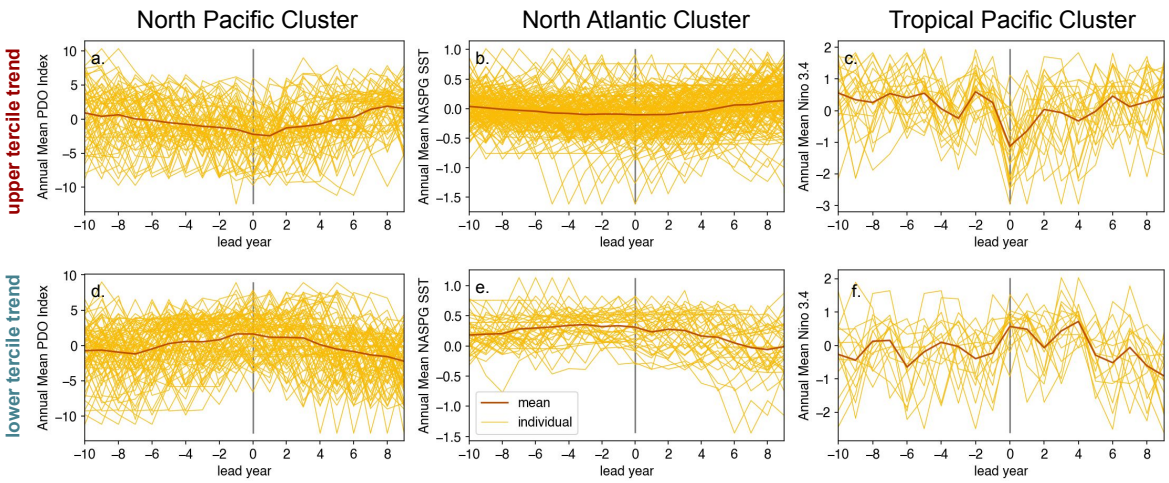

Figure C.4: a. Annual Mean Nino 3.4 index for 10 years before up to 10 years after predictions of positive SST trend in the Tropical Pacific cluster. b. Annual mean PDO index for predictions of upper tercile SST trend in the North Pacific cluster. c. Annual mean North Atlantic subpolar gyre SST for predictions of upper tercile SST trend in the North Atlantic cluser. d., e. and f. as a. b. and c. but for lower tercile predictions.