Oceanic harbingers of Pacific Decadal Oscillation predictability in CESM2 detected by neural networks

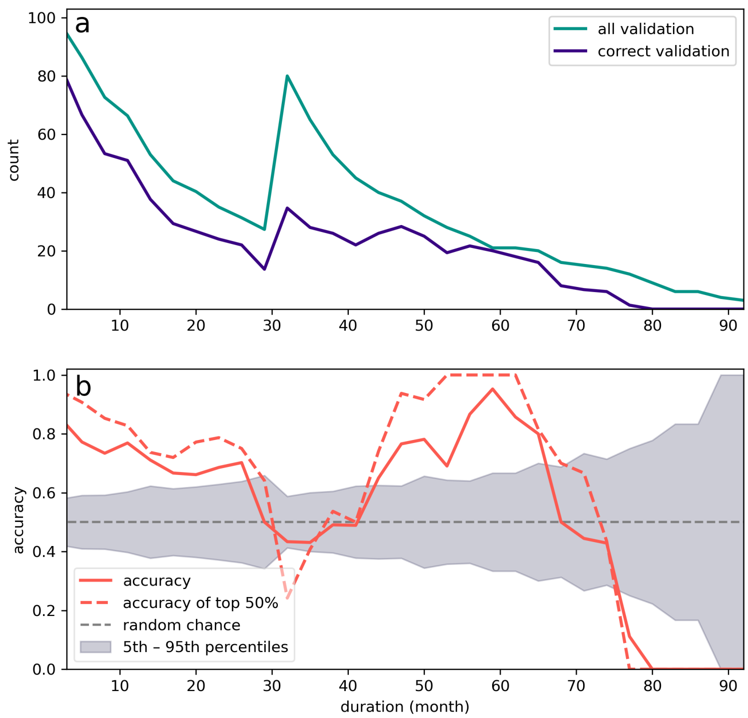
Emily M Gordon[1,1], Elizabeth A Barnes[1,1], and James Wilson Hurrell[1,1]

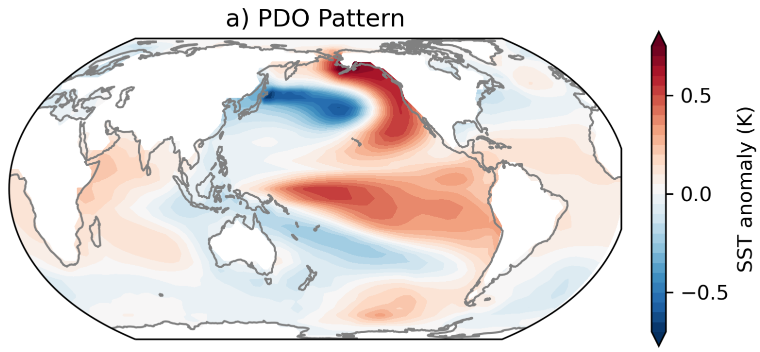
[1]Colorado State University

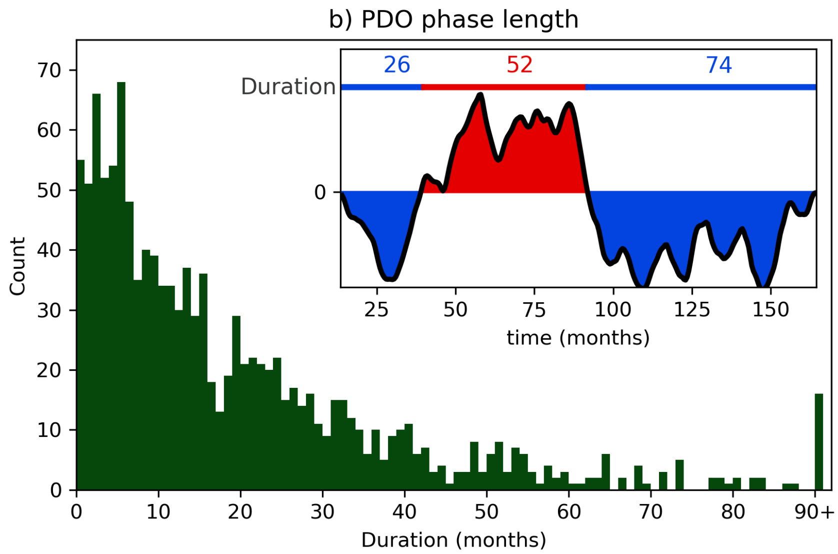
November 30, 2022

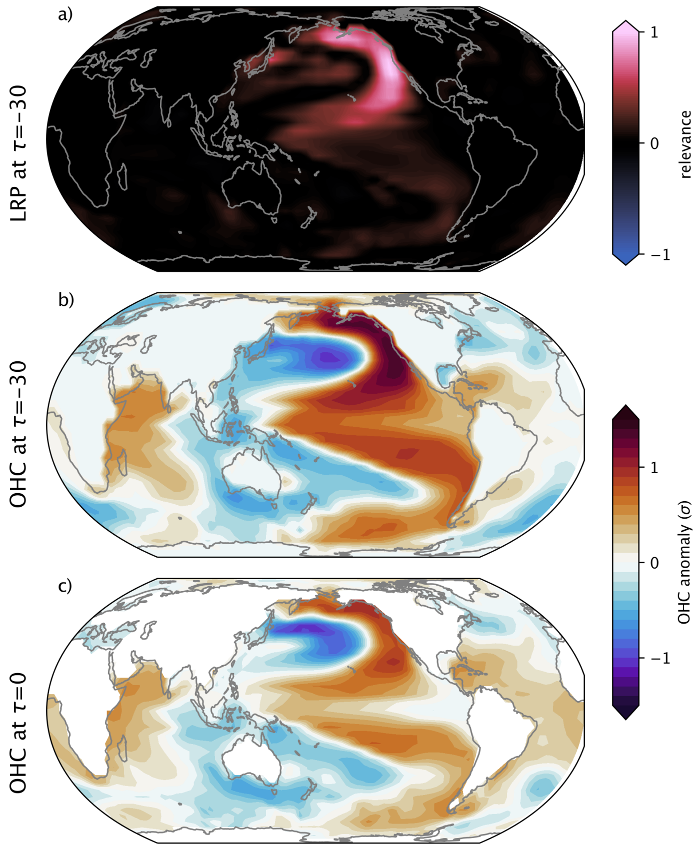
Abstract

Predicting Pacific Decadal Oscillation (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. We show that artificial neural networks (ANNs) predict PDO persistence and transitions on the interannual timescale. Using layer-wise relevance propagation to investigate the ANN predictions, we demonstrate that the ANNs utilize oceanic patterns that have been previously linked to predictable PDO behavior. For PDO transitions, ANNs recognize a build-up of ocean heat content in the off-equatorial western Pacific 12-27 months before a transition occurs. The results support the continued use of ANNs in climate studies where explainability tools can assist in mechanistic understanding of the climate system.

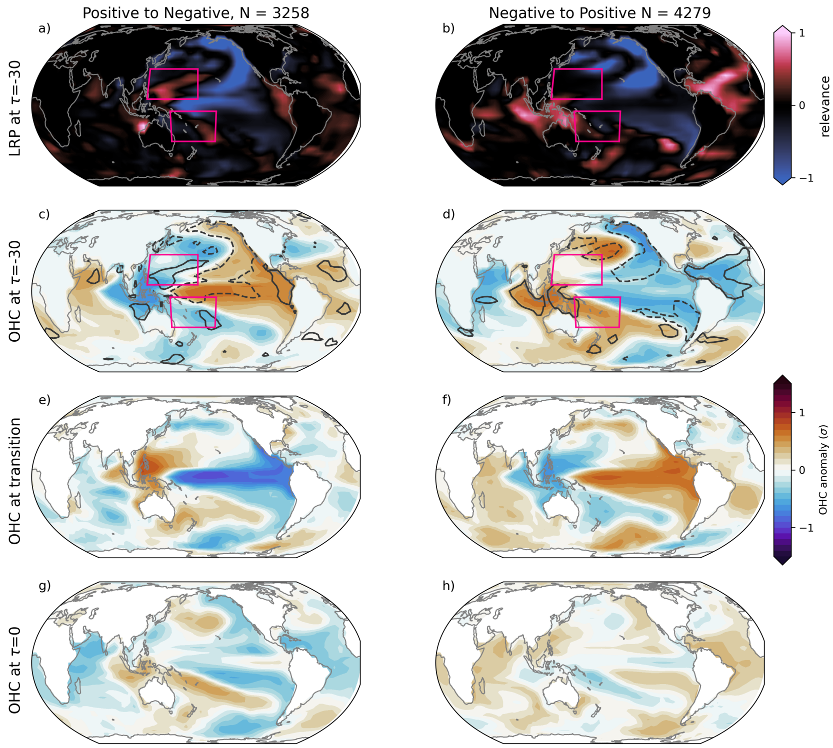

a
all validation
correct validation
count
b
accuracy
accuracy of top 50%
random chance
5th – 95th percentiles
accuracy
duration (month)


a) PDO Pattern
0.5
0.0
−0.5
SST anomaly (K)


b) PDO phase length
Count
Duration (months)
0
10
20
30
40
50
60
70
80
90+
Duration
26
52
74
time (months)
25
50
75
100
125
150



a)
LRP at τ=−30
b)
OHC at τ=−30
c)
OHC at τ=0
1
0
−1
relevance
1
0
−1
OHC anomaly (σ)


Positive to Negative, N = 3258
Negative to Positive N = 4279
a)
b)
c)
d)
e)
f)
g)
h)
LRP at τ=-30
OHC at τ=-30
OHC at transition
OHC at τ=0
relevance
1
0
−1
OHC anomaly (σ)
1
0
−1

Oceanic harbingers of Pacific Decadal Oscillation predictability in CESM2 detected by neural networks

E. M. Gordon[1], E. A. Barnes[1], J. W. Hurrell[1]

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

Key Points:

- Artificial neural networks (ANNs) predict Pacific Decadal Oscillation (PDO) persistence and transitions in CESM2.
- Explainable AI unveils regions used by ANNs for predicting the PDO on interannual timescales.
- Predictable PDO transitions can be preceded by a heat build up in off-equatorial western Pacific.

Corresponding author: E. M. Gordon, `emily.m.gordon95@gmail.com`

Abstract

Predicting Pacific Decadal Oscillation (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. We show that artificial neural networks (ANNs) predict PDO persistence and transitions with lead times of 12 months onward. Using layer-wise relevance propagation to investigate the ANN predictions, we demonstrate that the ANNs utilize oceanic patterns that have been previously linked to predictable PDO behavior. For PDO transitions, ANNs recognize a build-up of ocean heat content in the off-equatorial western Pacific 12-27 months before a transition occurs. The results support the continued use of ANNs in climate studies where explainability tools can assist in mechanistic understanding of the climate system.

Plain Language Summary

The Earth's oceans are capable of storing large amounts of heat with spatial patterns of ocean heat lasting for decades at a time. One such pattern is called the Pacific Decadal Oscillation (PDO). As these patterns indicate how heat is distributed over the globe, they are associated with increased predictability of extreme weather events as well as being an important factor for marine ecosystems. Predicting when the PDO will shift from one pattern to the other has proven a tricky proposition in climate science as mechanisms from the atmosphere and the ocean both play a role. Here we show that artificial intelligence can predict PDO transitions over 12 months in advance. We also investigate the predictions and show that they are related to known physical mechanisms — our models are making the right predictions for the right reasons. We leverage past knowledge, and the new discoveries from artificial intelligence to speculate how ocean patterns can lead to PDO predictability.

1 Introduction

The Pacific Decadal Oscillation (PDO; Mantua et al., 1997; Zhang et al., 1997) is recognised as one of the most important sources of predictability on decadal timescales (Cassou et al., 2018). 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 (Newman et al., 2016, 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 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 (Cassou et al., 2018). This is especially true in predicting PDO transitions, i.e. when the PDO shifts from one phase to the other. Stochastic models (Deser et al., 2003; Newman et al., 2003; Schneider & Cornuelle, 2005), linear inverse models (LIMs; Newman, 2007; Alexander et al., 2008; Dias et al., 2019), atmosphere-only models (Farneti et al., 2014) and fully coupled climate models (Meehl & Hu, 2006; Meehl et al., 2014) 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 (2006) 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. (2016) 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; Gent et al., 2011) 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 (Farneti et al., 2014; Lu et al., 2021).

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 (Newman, 2007). Artificial neural networks (ANNs), a form of unsupervised 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 (Barnes et al., 2020; Ham et al., 2019; Toms et al., 2020; Mayer & Barnes, 2021). Specifically in the case of oceanic predictability, Ham et al. (2019) 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 (2021) 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 (Newman et al., 2016). 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 (Toms et al., 2020).

2 Data and Methods

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; Danabasoglu et al., 2020) pre-industrial control run for the Coupled Model Intercomparison Project, Phase 6 (CMIP6; Eyring et al., 2016). The presence of realistic ENSO and PDO variability in CESM2 was demonstrated by Capotondi et al. (2020). 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 (Fasullo & Nerem, 2016). Both OHC and SST are interpolated to a $4°\times4°$ 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.

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 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. This is because PDO transitions are considered to be influenced by interannual variability associated with e.g. ENSO (Meehl et al., 2016, 2021) and we want our ANNs to be able to account for these processes. The distribution of phase durations in CESM2 is shown in Figure 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 (Capotondi et al., 2020). 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 (Li et al., 2019), hence high frequency PDO variability may become more relevant in future climate scenarios.

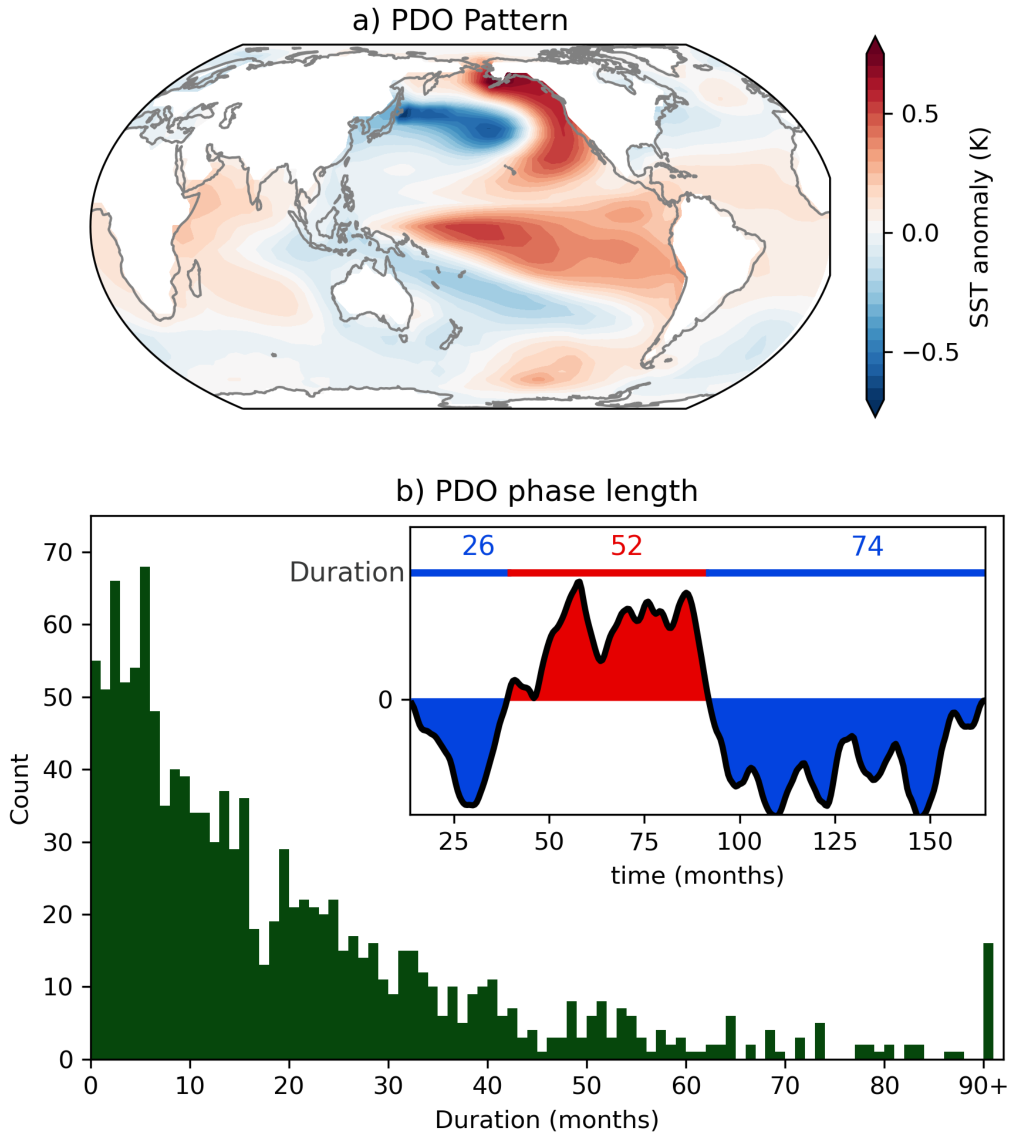


Figure 1. a) North Pacific PC 1 projected onto global de-seasoned SST. b) Histogram showing distribution of PDO phase lengths in CESM pre-industrial control run. Inset: slice of PDO index showing PDO phase length as number of months between phase changes.

2.2 Artificial Neural Network

We use a single layer artificial neural network (ANN) 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 30 months. An overview of neural networks is provided in the supplement 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 $4° \times 4°$ 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 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. 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 the supplement text. 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; Bach et al., 2015). 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. (2020); Mamalakis et al. (2021) and used in studies by e.g. Mayer and Barnes (2021); Toms et al. (2021); Sonnewald and Lguensat (2021). 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.

3 Results

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. 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. 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 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 3a, while Figures 3b and 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. 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. 3a align with the positive horse-shoe shape in 3b. This suggests that the ANNs recognize large positive OHC anomalies in the North Pacific ocean as being an indicator that the PDO will persist on the interannual timescale, and this is consistent with our physical understanding.

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 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.

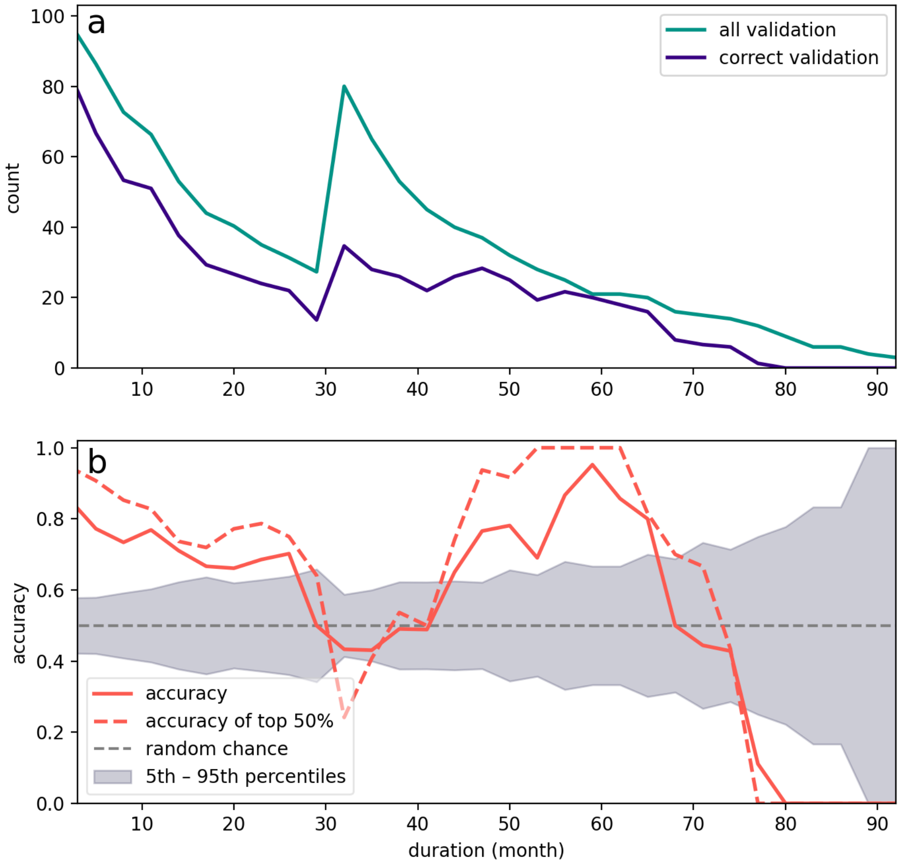


Figure 2. a: Average distribution of phase duration in the validation data for the three ANNs, green shows all the validation data and blue is number correctly predicted by the ANN with data binned into 3 month averages. b: Red line is accuracy of each phase duration bin (blue divided by green from above), 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.

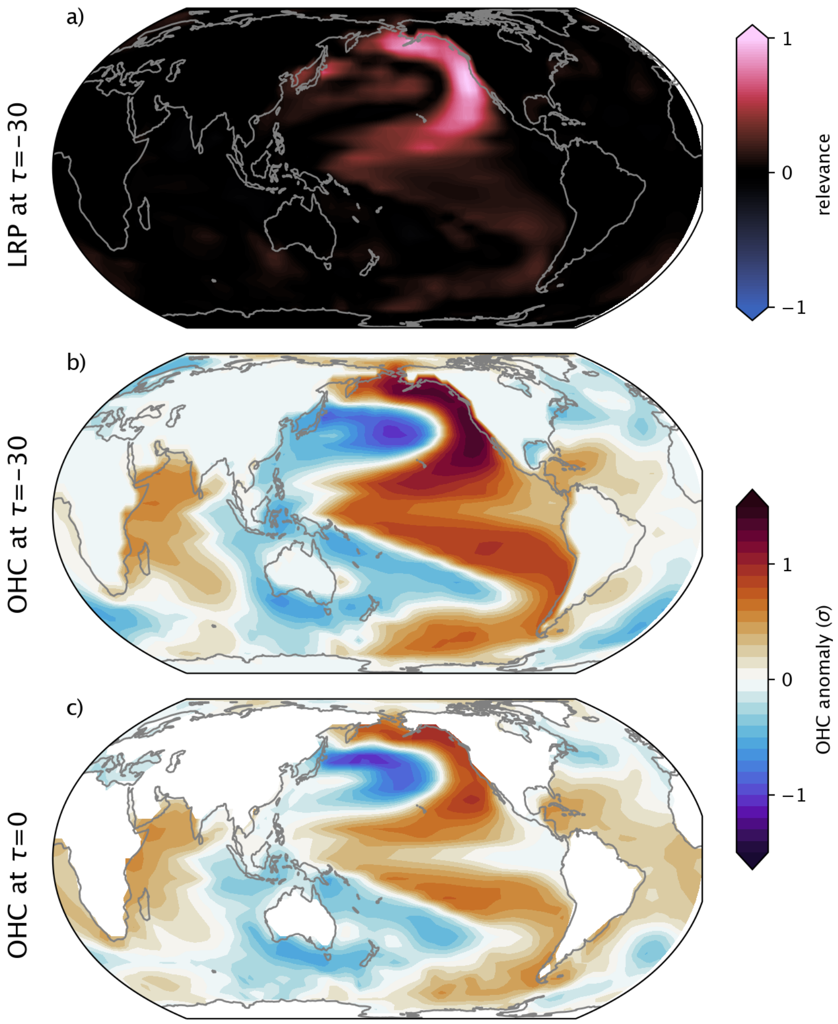


Figure 3. Composite maps when ANN correctly and confidently predicts persistence. a) Composite mean of LRP maps at final input month (τ=-30). Red areas correspond to positive relevance and blue to negative relevance. b) Composite mean of OHC input maps at τ=-30. Color scale is OHC anomaly in units of standard deviation σ at each grid-point. c) Composite mean of OHC at predicted month, color scale as in b).

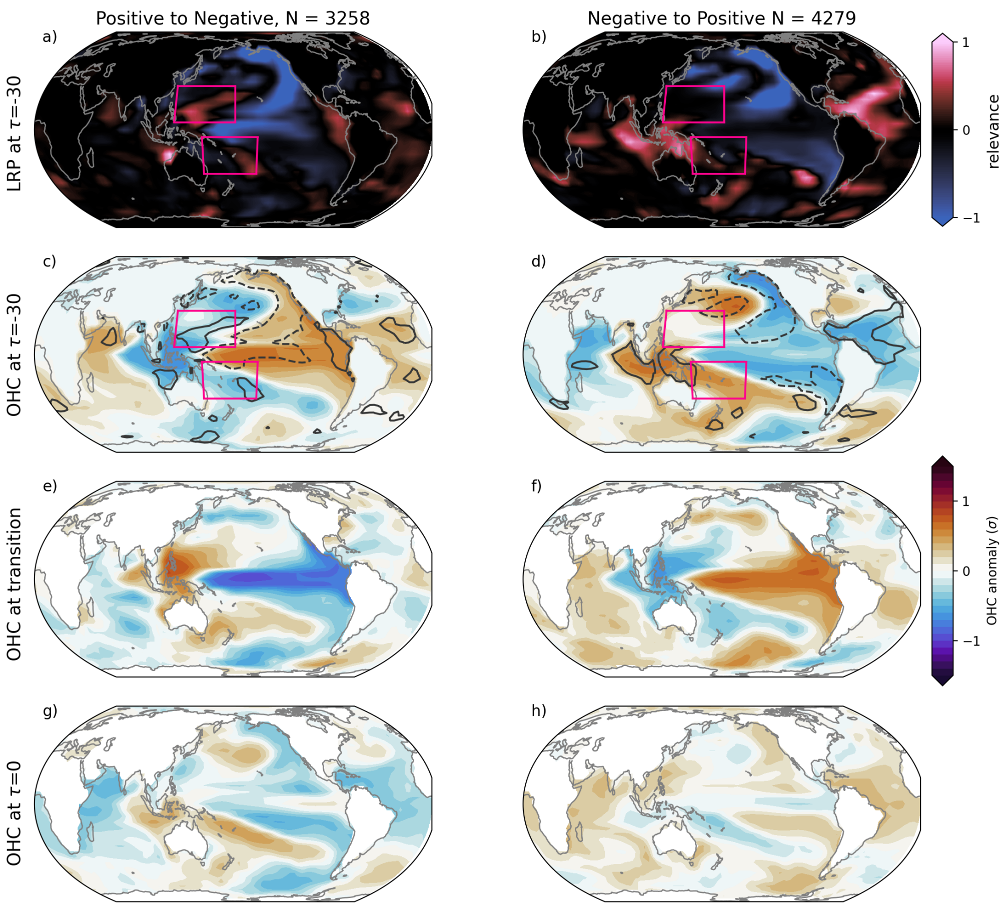


Figure 4. Composite maps of correct and confident predictions of PDO transition when transition occurs 12-27 months after input. Left column is positive to negative transitions, and right column is negative to positive transitions. Number of samples in each column is included in the title. Panels a) 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.

Figure 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 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 4a and 4b are the LRP maps for the final input map (month $\tau = -30$) with Figure 4c and 4d the corresponding OHC. We highlight the strongest relevance regions from the LRP maps by superimposing LRP contours (Fig. 4a and 4b) onto the OHC (Fig. 4c and 4d), with solid lines contours outlining highest 5% relevance values. Similarly, dashed contours encircle regions with the lowest 5% relevance values. Furthermore, we include pink squares in Fig. 4a–d to emphasize the regions where a build-up of OHC has been suggested in the literature to precede a PDO transition (Meehl et al., 2016). Lastly, to track the OHC evolution throughout the transition process, panels 4e and 4f show the OHC when the transition occurs, and 4g and 4h the OHC at month $\tau = 0$. Note in Figure S3-S4 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. 4c), while large positive anomalies precede negative to positive transitions in the southern off-equatorial western Pacific (Fig. 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. 4c, the ANNs have recognized the northern region of heat content build up, with high relevance in the LRP composite in Fig. 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. 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 (Meehl et al., 2016, pink boxes in Fig. 4b) 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. 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 supplement figure S6 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. S6 (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. 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. 4e and 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. (2016), 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. 4g and 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 S6). Conversely, there are other signals detected in the relevance maps (Figs 4a and 4b), and in addition the OHC anomalies are not consistently strong in the off-equatorial regions (Fig. 4d) which suggests that there are likely mechanisms other than that proposed by Meehl et al. (2016) that contribute to PDO transitions. 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.

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. (2016) 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. (2021) 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. (2021), 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 S7). We use 2.5 years here as a balance between sample size and long duration phases.

The maps in Figures 3 and 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 4. Figures S5-S6 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 (Kucharski et al., 2016; Chikamoto et al., 2020; Johnson et al., 2020; Meehl et al., 2020) 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 (Ham et al., 2013) whereas La Nina events are thought to be largely triggered by a preceding El Nino event. In Figure 4b, the neural networks concentrate relevance in the Atlantic basin preceding the El Nino event (and PDO transition) in Figure 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.

Acknowledgments

EMG is partially funded by Fulbright New Zealand. EAB is supported, in part, by NSF CAREER AGS-1749261 under the Climate and Large-scale Dynamics program.

The authors declare that they have no conflicts of interest.

Analysis was carried out in Python 3.7 and 3.9, ANNs were developed using TensorFlow (Abadi et al., 2016), while LRP visualizations were created with iNNvestigate (Alber et al., 2019). Colormaps were used from CMasher (van der Velden, 2020). Regridding was achieved using Climate Data Operators (CDO; Schulzweida, 2019).

Thanks to John Fasullo at the National Center for Atmospheric Research (NCAR) for diagnosing the OHC from CESM2. We would like to acknowledge high-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

Data Availability: CESM2 pre-industrial control output for CMIP6 (doi:10.22033/ESGF/CMIP6.7627) is freely available from Earth System Grid https://esgf-node.llnl.gov/projects/cmip6/ .

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Zheng, X. (2016, November). Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265–283). Savannah, GA: USENIX Association. Retrieved from `https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi`

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., … Kindermans, P.-J. (2019). innvestigate neural networks! *Journal of Machine Learning Research*, *20*(93), 1-8. Retrieved from `http://jmlr.org/papers/v20/18-540.html`

Alexander, M. A., Matrosova, L., Penland, C., Scott, J. D., & Chang, P. (2008, January). Forecasting Pacific SSTs: Linear Inverse Model Predictions of the PDO. *Journal of Climate*, *21*(2), 385–402. Retrieved 2021-05-19, from `http://journals.ametsoc.org/view/journals/clim/21/2/2007jcli1849.1.xml` (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/2007JCLI1849.1

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015, 07). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), 1-46. Retrieved from `https://doi.org/10.1371/journal.pone.0130140` doi: 10.1371/journal.pone.0130140

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020, September). Indicator patterns of forced change learned by an artificial neural network. *J. Adv. Model. Earth Syst.*, *12*(9). Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2020MS002195 doi: 10.1029/2020ms002195

Capotondi, A., Deser, C., Phillips, A. S., Okumura, Y., & Larson, S. M. (2020). ENSO and Pacific Decadal Variability in the Community Earth System Model Version 2. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2019MS002022. Retrieved 2021-05-04, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002022 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS002022) doi: https://doi.org/10.1029/2019MS002022

Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I.-S., & Caltabiano, N. (2018, March). Decadal Climate Variability and Predictability: Challenges and Opportunities. *Bull. Amer. Meteor. Soc.*, *99*(3), 479–490. Retrieved 2020-09-29, from http://journals.ametsoc.org/bams/article/99/3/479/70287/Decadal-Climate-Variability-and-Predictability (Publisher: American Meteorological Society) doi: 10.1175/BAMS-D-16-0286.1

Chikamoto, Y., Johnson, Z. F., Wang, S.-Y. S., McPhaden, M. J., & Mochizuki, T. (2020). El Niño–Southern Oscillation Evolution Modulated by Atlantic Forcing. *Journal of Geophysical Research: Oceans*, *125*(8), e2020JC016318. Retrieved 2021-04-07, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JC016318 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020JC016318) doi: https://doi.org/10.1029/2020JC016318

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., ... Strand, W. G. (2020). The Community Earth System Model Version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001916. Retrieved 2021-05-03, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001916 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001916) doi: https://doi.org/10.1029/2019MS001916

Deser, C., Alexander, M. A., & Timlin, M. S. (2003, January). Understanding the Persistence of Sea Surface Temperature Anomalies in Midlatitudes. *J. Clim.*, *16*(1), 57–72. Retrieved from http://journals.ametsoc.org/jcli/article/

16/1/57/29818/Understanding-the-Persistence-of-Sea-Surface doi: 10.1175/1520-0442(2003)016⟨0057:UTPOSS⟩2.0.CO;2

Dias, D. F., Subramanian, A., Zanna, L., & Miller, A. J. (2019, March). Remote and local influences in forecasting Pacific SST: a linear inverse model and a multimodel ensemble study. *Clim Dyn*, *52*(5), 3183–3201. Retrieved 2021-05-20, from https://doi.org/10.1007/s00382-018-4323-z doi: 10.1007/s00382-018-4323-z

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. Retrieved from https://gmd.copernicus.org/articles/9/1937/2016/ doi: 10.5194/gmd-9-1937-2016

Farneti, R., Molteni, F., & Kucharski, F. (2014, June). Pacific interdecadal variability driven by tropical–extratropical interactions. *Clim Dyn*, *42*(11), 3337–3355. Retrieved 2021-03-17, from https://doi.org/10.1007/s00382-013-1906-6 doi: 10.1007/s00382-013-1906-6

Fasullo, J. T., & Nerem, R. S. (2016, October). Interannual Variability in Global Mean Sea Level Estimated from the CESM Large and Last Millennium Ensembles. *Water*, *8*(11), 491. Retrieved from https://www.mdpi.com/2073-4441/8/11/491 doi: 10.3390/w8110491

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., . . . Zhang, M. (2011, October). The Community Climate System Model Version 4. *J. Clim.*, *24*(19), 4973–4991. Retrieved from https://journals.ametsoc.org/view/journals/clim/24/19/2011jcli4083.1.xml?tab_body=fulltext-display doi: 10.1175/2011JCLI4083.1

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019, September). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572. Retrieved 2021-05-06, from http://www.nature.com/articles/s41586-019-1559-7 (Number: 7775 Publisher: Nature Publishing Group) doi: 10.1038/s41586-019-1559-7

Ham, Y.-G., Kug, J.-S., & Park, J.-Y. (2013, August). Two distinct roles of Atlantic SSTs in ENSO variability: North Tropical Atlantic SST and Atlantic Niño. *Geophys. Res. Lett.*, *40*(15), 4012–4017. Retrieved from http://dx.doi.org/10.1002/grl.50729 doi: 10.1002/grl.50729

Johnson, Z. F., Chikamoto, Y., Wang, S.-Y. S., McPhaden, M. J., & Mochizuki, T. (2020, August). Pacific decadal oscillation remotely forced by the equatorial Pacific and the Atlantic Oceans. *Clim Dyn*, *55*(3), 789–811. Retrieved 2021-05-25, from https://doi.org/10.1007/s00382-020-05295-2 doi: 10.1007/s00382-020-05295-2

Kucharski, F., Ikram, F., Molteni, F., Farneti, R., Kang, I.-S., No, H.-H., … Mogensen, K. (2016, April). Atlantic forcing of Pacific decadal variability. *Clim Dyn*, *46*(7), 2337–2351. Retrieved 2021-05-25, from https://doi.org/10.1007/s00382-015-2705-z doi: 10.1007/s00382-015-2705-z

Li, S., Wu, L., Yang, Y., Geng, T., Cai, W., Gan, B., … Ma, X. (2019, December). The Pacific Decadal Oscillation less predictable under greenhouse warming. *Nat. Clim. Chang.*, *10*(1), 30–34. Retrieved from https://www.nature.com/articles/s41558-019-0663-x doi: 10.1038/s41558-019-0663-x

Lu, Z., Yuan, N., Yang, Q., Ma, Z., & Kurths, J. (2021). Early Warning of the Pacific Decadal Oscillation Phase Transition Using Complex Network Analysis. *Geophysical Research Letters*, *48*(7), e2020GL091674. Retrieved 2021-05-04, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL091674 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020GL091674) doi: https://doi.org/10.1029/2020GL091674

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021). *Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset.*

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., & Francis, R. C. (1997, June). A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production*. *Bull. Amer. Meteor. Soc.*, *78*(6), 1069–1080. Retrieved 2020-10-18, from http://journals.ametsoc.org/bams/article/78/6/1069/55942/A-Pacific-Interdecadal-Climate-Oscillation-with (Publisher: American Meteorological Society) doi: 10.1175/1520-0477(1997)078⟨1069:APICOW⟩2.0.CO;2

Mayer, K. J., & Barnes, E. A. (2021). Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network. *Geophysical Research Letters*, *n/a*(n/a), e2020GL092092. Retrieved 2021-05-04, from http://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL092092 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020GL092092) doi: https://doi.org/10.1029/2020GL092092

Meehl, G. A., & Hu, A. (2006, May). Megadroughts in the Indian Monsoon Region and Southwest North America and a Mechanism for Associated Multidecadal Pacific Sea Surface Temperature Anomalies. *Journal of Climate*, *19*(9), 1605–1623. Retrieved 2021-05-19, from `http://journals.ametsoc.org/view/journals/clim/19/9/jcli3675.1.xml` (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/JCLI3675.1

Meehl, G. A., Hu, A., Castruccio, F., England, M. H., Bates, S. C., Danabasoglu, G., ... Rosenbloom, N. (2020, December). Atlantic and pacific tropics connected by mutually interactive decadal-timescale processes. *Nat. Geosci.*, 1–7. Retrieved from `http://www.nature.com/articles/s41561-020-00669-x` doi: 10.1038/s41561-020-00669-x

Meehl, G. A., Hu, A., & Teng, H. (2016, June). Initialized decadal prediction for transition to positive phase of the Interdecadal Pacific Oscillation. *Nature Communications*, *7*(1), 11718. Retrieved 2021-05-04, from `http://www.nature.com/articles/ncomms11718` (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/ncomms11718

Meehl, G. A., Teng, H., & Arblaster, J. M. (2014, October). Climate model simulations of the observed early-2000s hiatus of global warming. *Nature Climate Change*, *4*(10), 898–902. Retrieved 2021-03-17, from `http://www.nature.com/articles/nclimate2357` (Number: 10 Publisher: Nature Publishing Group) doi: 10.1038/nclimate2357

Meehl, G. A., Teng, H., Capotondi, A., & Hu, A. (2021, May). The role of interannual ENSO events in decadal timescale transitions of the Interdecadal Pacific Oscillation. *Clim. Dyn.*. Retrieved from `https://doi.org/10.1007/s00382-021-05784-y` doi: 10.1007/s00382-021-05784-y

Nadiga, B. T. (2021). Reservoir Computing as a Tool for Climate Predictability Studies. *Journal of Advances in Modeling Earth Systems*, *13*(4), e2020MS002290. Retrieved 2021-05-04, from `http://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002290` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002290) doi: https://doi.org/10.1029/2020MS002290

Newman, M. (2007, June). Interannual to Decadal Predictability of Tropical and North Pacific Sea Surface Temperatures. *Journal of Climate*, *20*(11), 2333–2356. Retrieved 2021-05-20, from https://journals.ametsoc.org/view/journals/clim/20/11/jcli4165.1.xml (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/JCLI4165.1

Newman, M., Alexander, M. A., Ault, T. R., Cobb, K. M., Deser, C., Di Lorenzo, E., … Smith, C. A. (2016, June). The Pacific Decadal Oscillation, Revisited. *J. Climate*, *29*(12), 4399–4427. Retrieved 2020-09-03, from https://journals.ametsoc.org/jcli/article/29/12/4399/34340/The-Pacific-Decadal-Oscillation-Revisited (Publisher: American Meteorological Society) doi: 10.1175/JCLI-D-15-0508.1

Newman, M., Compo, G. P., & Alexander, M. A. (2003, December). ENSO-Forced Variability of the Pacific Decadal Oscillation. *Journal of Climate*, *16*(23), 3853–3857. Retrieved 2021-05-21, from http://journals.ametsoc.org/view/journals/clim/16/23/1520-0442_2003_016_3853_evotpd_2.0.co_2.xml (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/1520-0442(2003)016⟨3853:EVOTPD⟩2.0.CO;2

Schneider, N., & Cornuelle, B. D. (2005, November). The Forcing of the Pacific Decadal Oscillation. *Journal of Climate*, *18*(21). Retrieved 2021-05-21, from https://journals.ametsoc.org/view/journals/clim/18/21/jcli3527.1.xml (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/JCLI3527.1

Schulzweida, U. (2019, October). *Cdo user guide.* Retrieved from https://doi.org/10.5281/zenodo.3539275 doi: 10.5281/zenodo.3539275

Sonnewald, M., & Lguensat, R. (2021, July). Revealing the impact of global heating on North Atlantic circulation using transparent machine learning. *J. Adv. Model. Earth Syst.*. Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2021MS002496 doi: 10.1029/2021ms002496

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. Retrieved 2021-05-04, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002002 (_eprint:

https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS002002) doi: https://doi.org/10.1029/2019MS002002

Toms, B. A., Barnes, E. A., & Hurrell, J. W. (2021, June). Assessing decadal predictability in an earth-system model using explainable neural networks. *Geophys. Res. Lett.*, *48*(12). Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2021GL093842` doi: 10.1029/2021gl093842

van der Velden, E. (2020, February). CMasher: Scientific colormaps for making accessible, informative and 'cmashing' plots. *The Journal of Open Source Software*, *5*(46), 2004. doi: 10.21105/joss.02004

Zhang, Y., Wallace, J. M., & Battisti, D. S. (1997, May). ENSO-like Interdecadal Variability: 1900–93. *Journal of Climate*, *10*(5), 1004–1020. Retrieved 2020-12-16, from `https://journals.ametsoc.org/view/journals/clim/10/5/1520-0442_1997_010_1004_eliv_2.0.co_2.xml` (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/1520-0442(1997)010⟨1004:ELIV⟩2.0.CO;2

Supporting Information for "Oceanic harbingers of PDO predictability detected by neural networks"

E. M. Gordon[1], E. A. Barnes[1], J. W. Hurrell[1]

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

Contents of this file

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.

Text S1: 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, Barnes, and Ebert-Uphoff (2020).

Text S2: 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).

Text S3: 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 S1 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 the main. 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 S2 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.

References

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. Retrieved 2021-05-04, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002002 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS002002) doi: https://doi.org/10.1029/2019MS002002

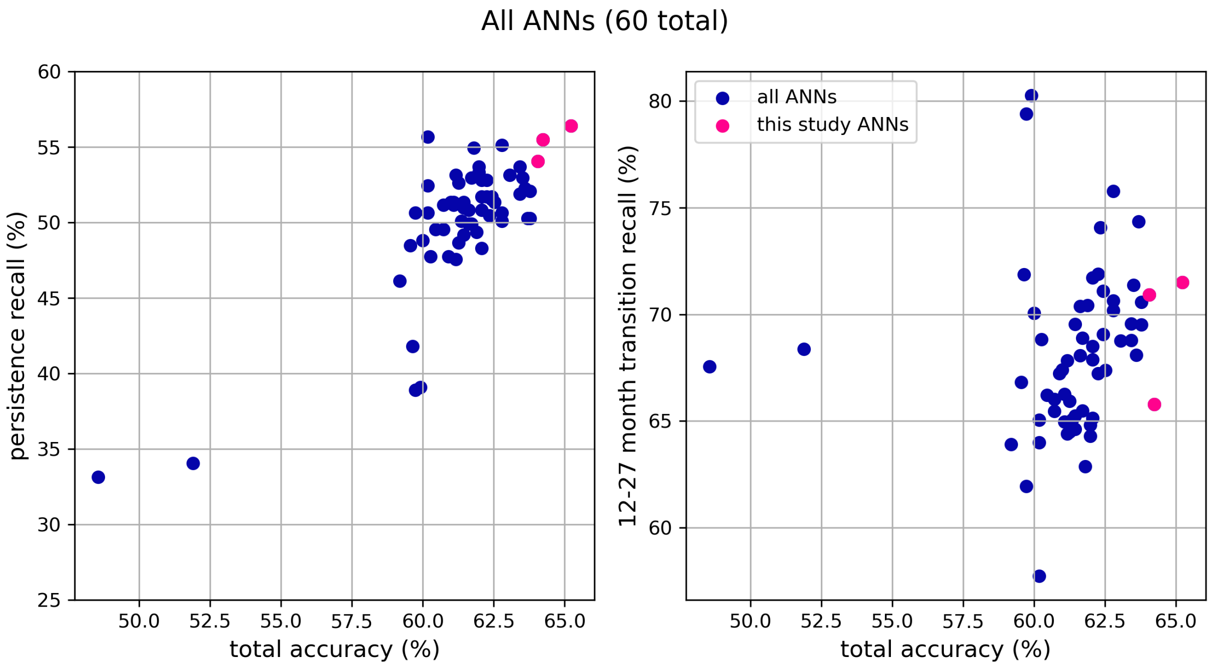


Figure S1. (left) Comparison of total accuracy (horizontal) and persistence recall (vertical) for all ANNs trained. Blue dots are all ANNs with pink dots representing the ANNs used in the study. (right) Comparison of total accuracy (horizontal) and 12-27 month transition recall.

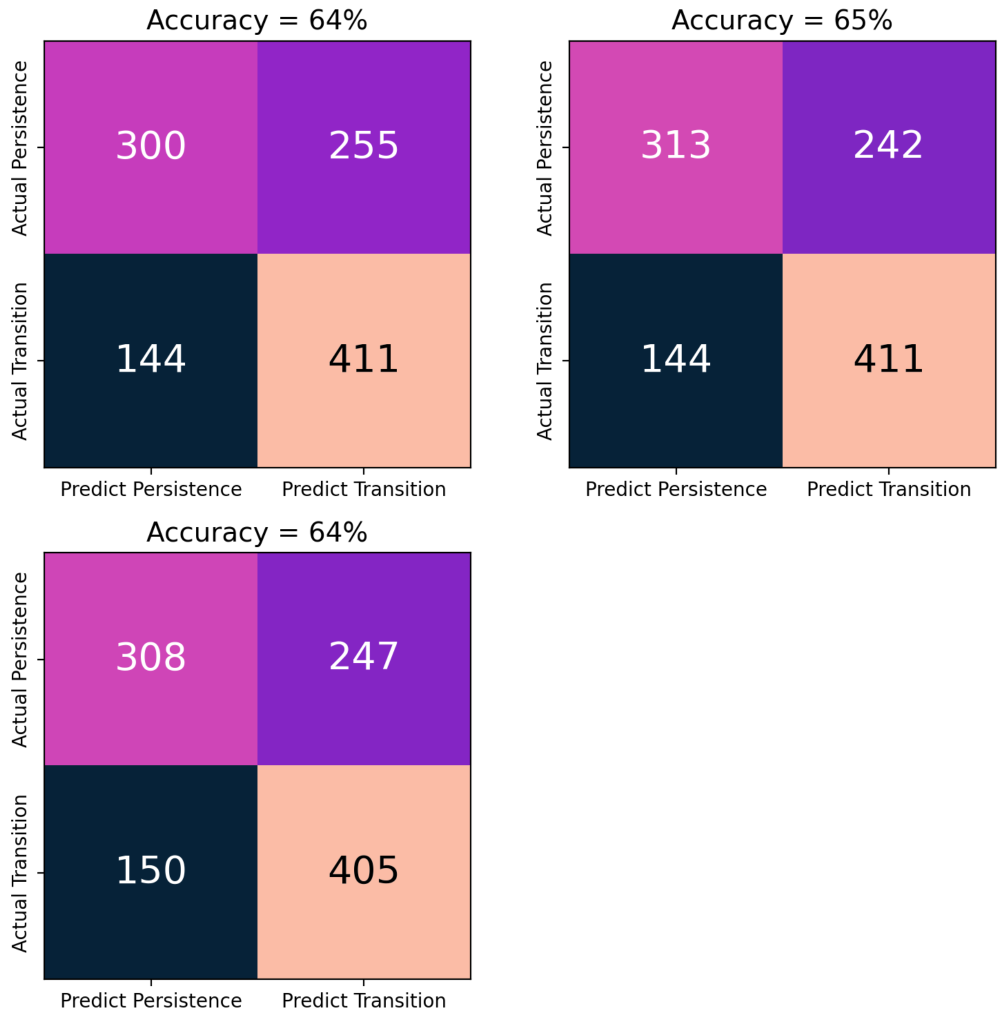


Figure S2. Confusion matrices for the 3 models used in this study. Vertical axis is the actual class and horizontal axis is the predicted class. Number of samples in each bin is printed in each square and total accuracy of each ANN in the title.

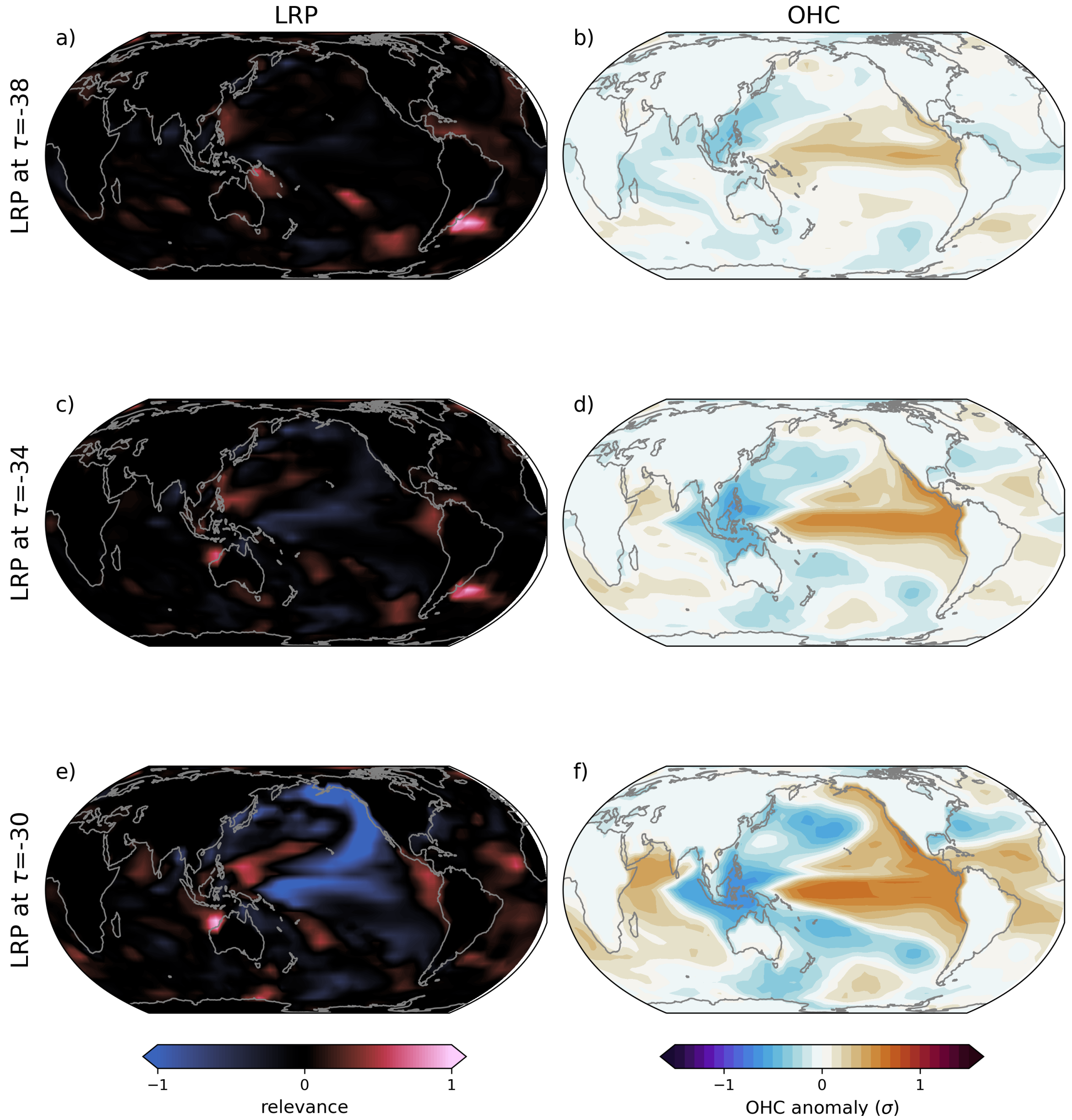


Figure S3. Left column: composite LRP maps for input maps where model correctly guesses transition from positive to negative occurs 12-27 months after final input. a) 38 months before output, c) 34 months before output, e) 30 months before output (and panel a in Figure 3). Right column: As left column but for composite OHC anomaly, with units of standard deviation at each grid point and color scale as in Figure 3.

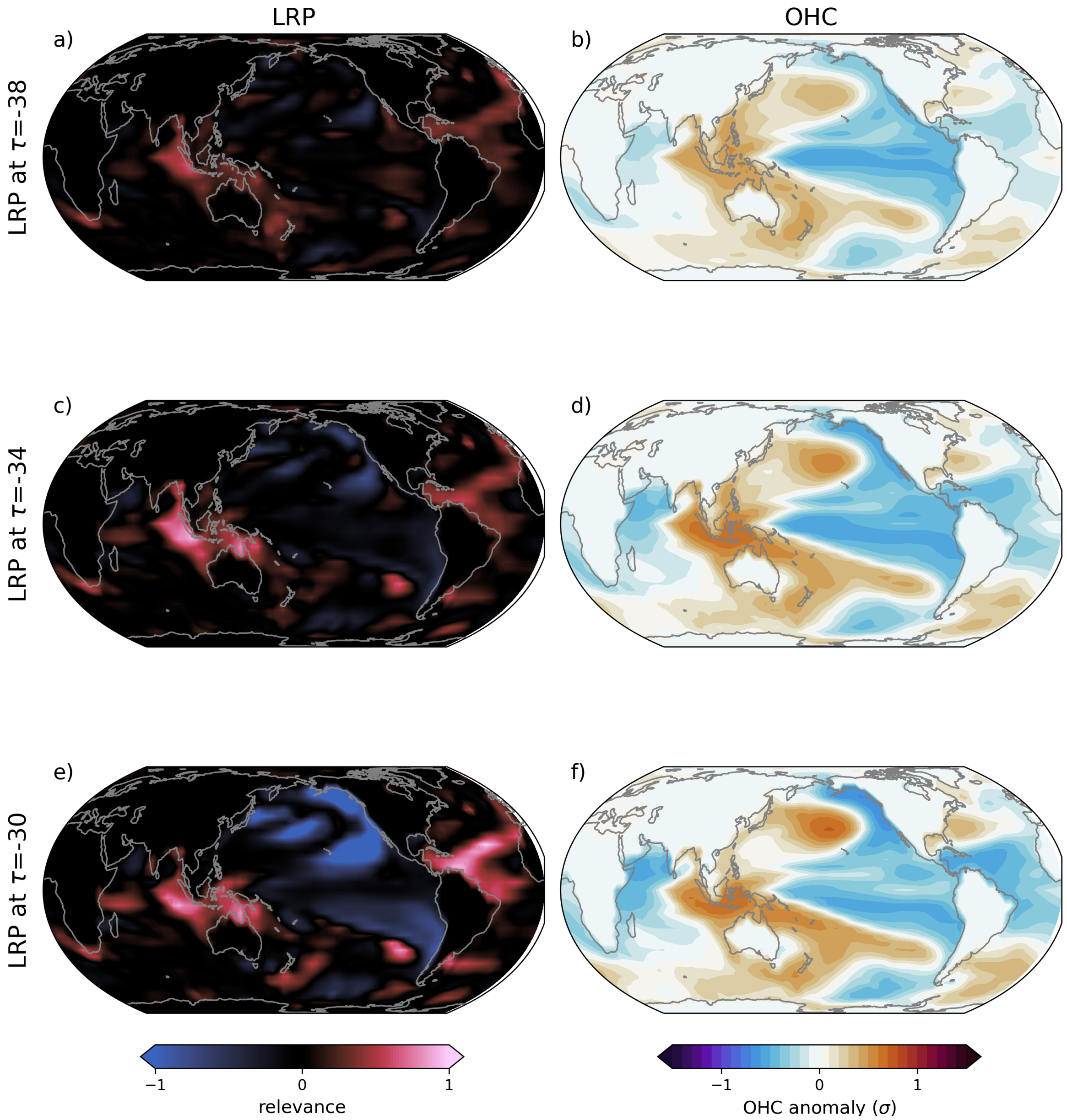


Figure S4. As Figure S4 but for negative to positive transitions.

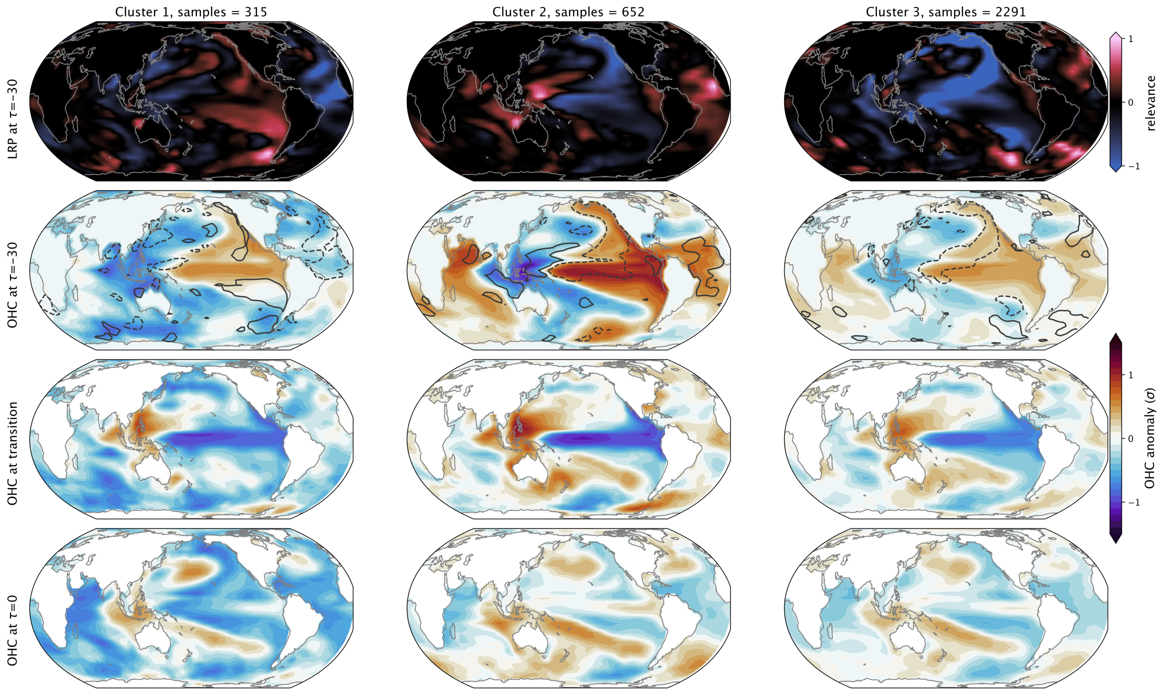


Figure S5. K-means of LRP maps when model correctly predicts positive to negative transition 12-27 months after input. Each column represents a cluster. Top row is LRP maps at month $\tau = -30$, second row is corresponding OHC with top and bottom 5% from the LRP contoured (dashed and dotted respectively as in Figure 3). The third row is OHC at the transition while the bottom row is OHC at month $\tau = 0$.

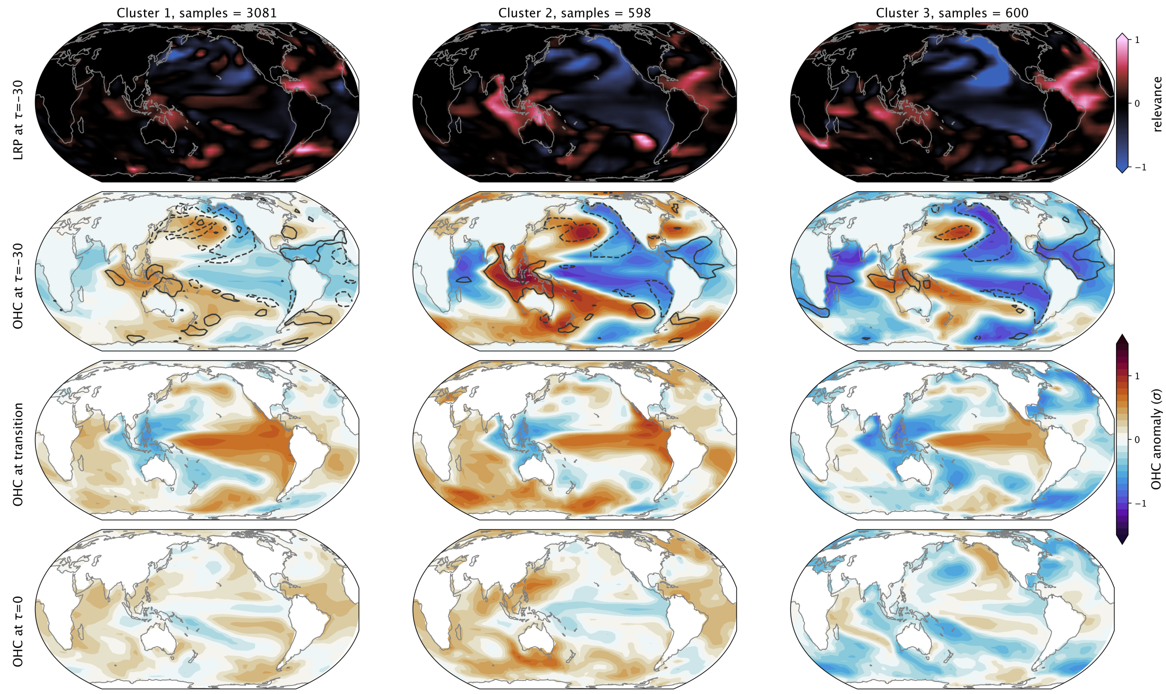


Figure S6. As Figure S5 but for negative to positive transitions

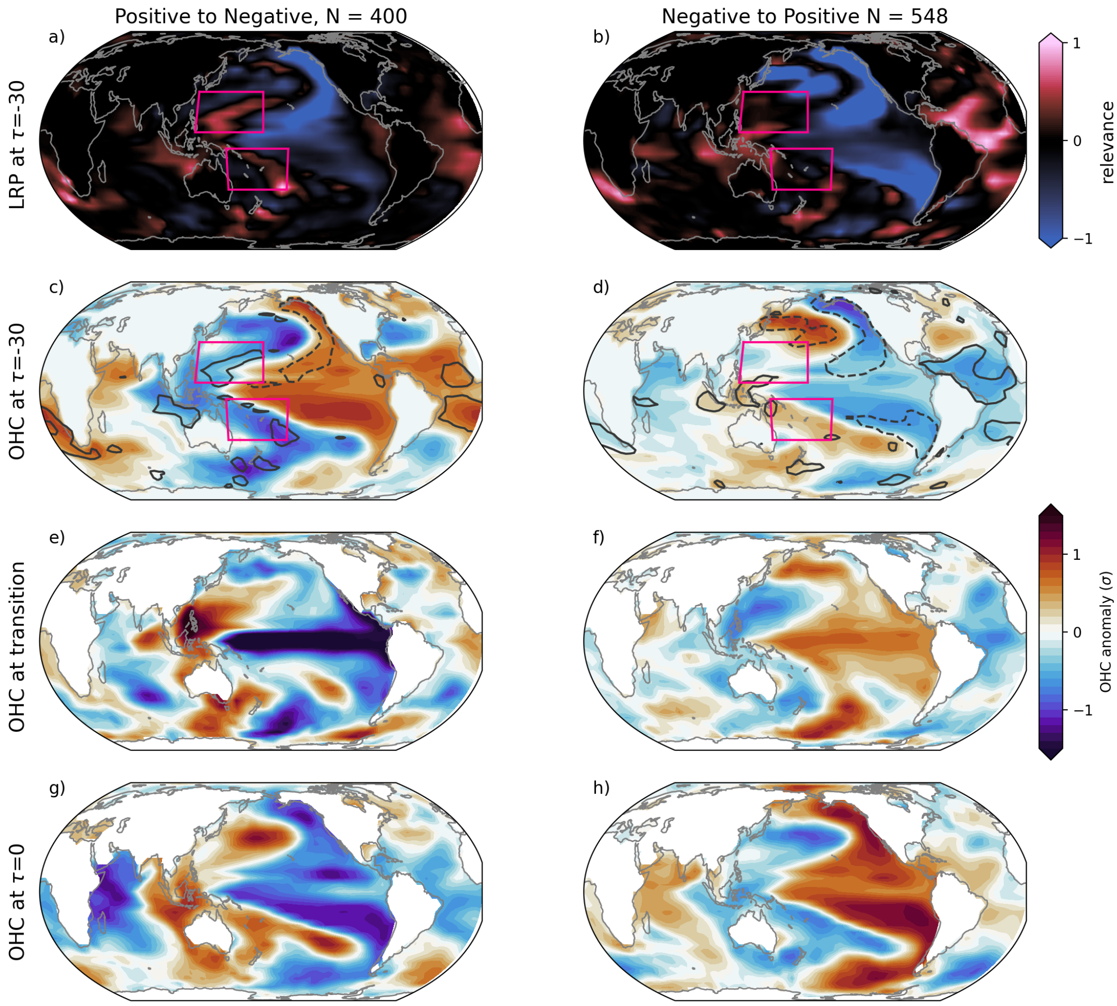


Figure S7. As Figure 4 in main but only for correct transition predictions where the PDO phase length preceding AND following a PDO transition are > 30 months.

Table S1. 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 main).
Output	Prediction of whether PDO transition occurs within 30 months of last input map.