Supporting Information for ”Incorporating Uncertainty into a Regression Neural Network Enables Identification of Decadal State-Dependent Predictability”

Emily M. Gordon[1], Elizabeth A. Barnes[1]

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

Contents of this file

Introduction The text in this document (Text S1) is a description of explainable AI (XAI), and provides a discussion of XAI findings which support the conclusions in the main text. This text references Figures S2 and S3 which are the XAI analyses of Figures

---

3 and 4 from the main document, respectively. In Figure 1 we provide plots showing the Atlantic multi-decadal variability (AMV) and interdecadal Pacific oscillation (IPO) patterns calculated in the CESM2 long control run.

IPO Index Calculation We calculate the IPO index using the method outlined by Henley et al. (2015) 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:

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

The resulting pattern from projecting the IPO index onto global SSTs is plotted in Figure S1, with the boxes in these calculations outlined in purple.

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, Ebert-Uphoff, and Barnes (2021) and Mamalakis, Barnes, and Ebert-Uphoff (2022).

The explainability composite maps for each region investigated in the main text is provided in Figure S4-S5. 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 S4 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 in the main text). 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 (Mamalakis et al., 2021)).

In Figure S5 we look at composite explainability maps for confident predictions of negative SST anomaly in the North Pacific ocean (green dot, same as Figure 4 in the main text). 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 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.

References

Henley, B. J., Gergis, J., Karoly, D. J., Power, S., Kennedy, J., & Folland, C. K. (2015, December). A Tripole Index for the Interdecadal Pacific Oscillation. *Clim. Dyn.*, *45*(11), 3077–3090. Retrieved from `https://doi.org/10.1007/s00382-015-2525-1` doi: 10.1007/s00382-015-2525-1

Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2022, February). Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *ArXiv*. Retrieved from `http://arxiv.org/abs/2202.03407`

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021, March). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *ArXiv*. Retrieved from `http://arxiv.org/abs/2103.10005`

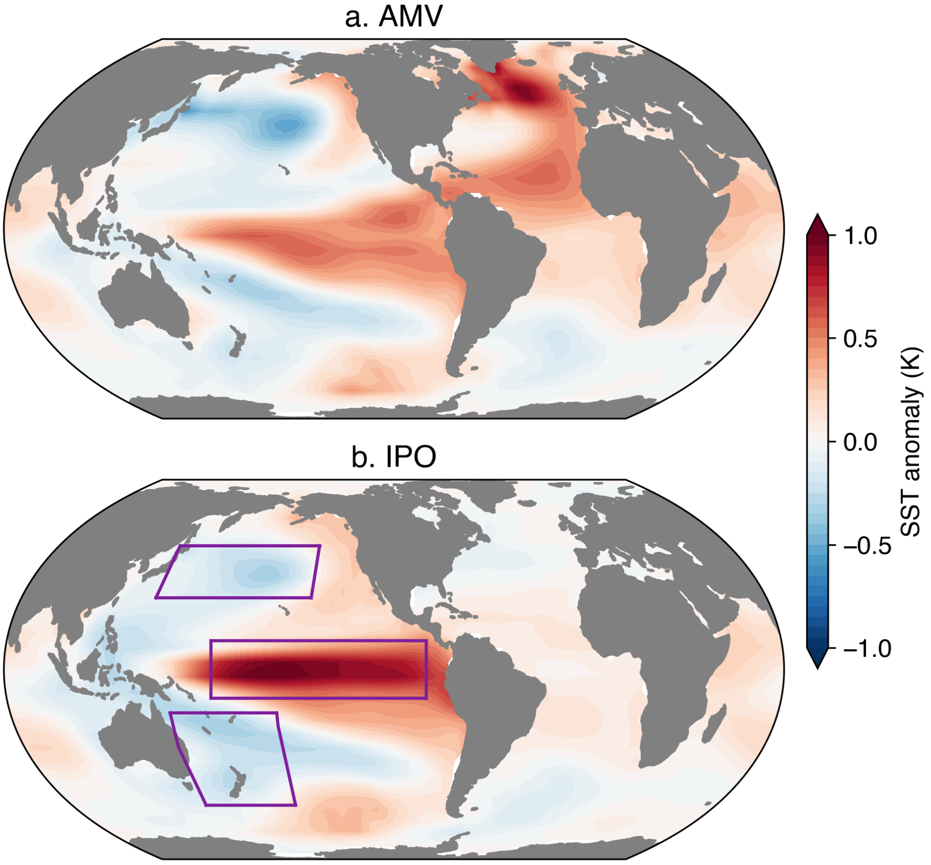

Figure S1. Patterns of large scale SST variability in CESM2 calculated using the methods discussed in Section 2.3 in Main a. AMV index projected onto global SSTs. b. IPO index projected onto global SSTs.

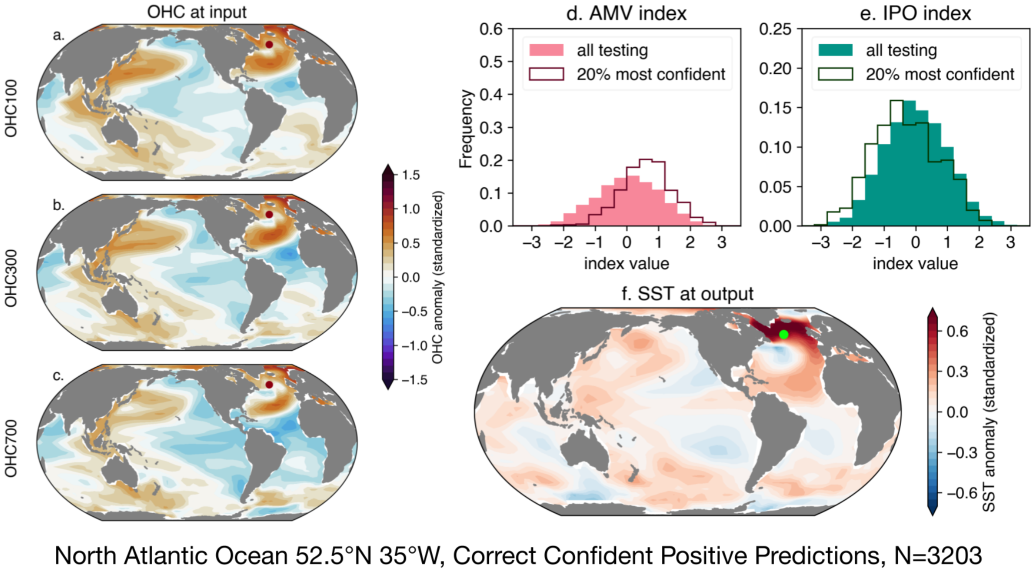

Figure S2. As Figure 3 in the main document but for the training and validation data.

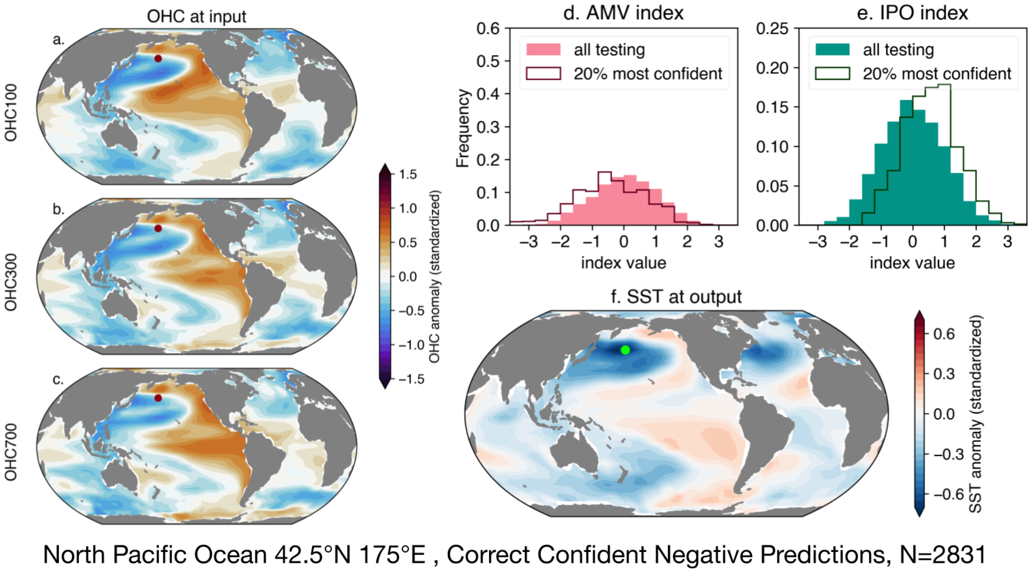

Figure S3. As Figure 4 in the main document but for the training and validation data.

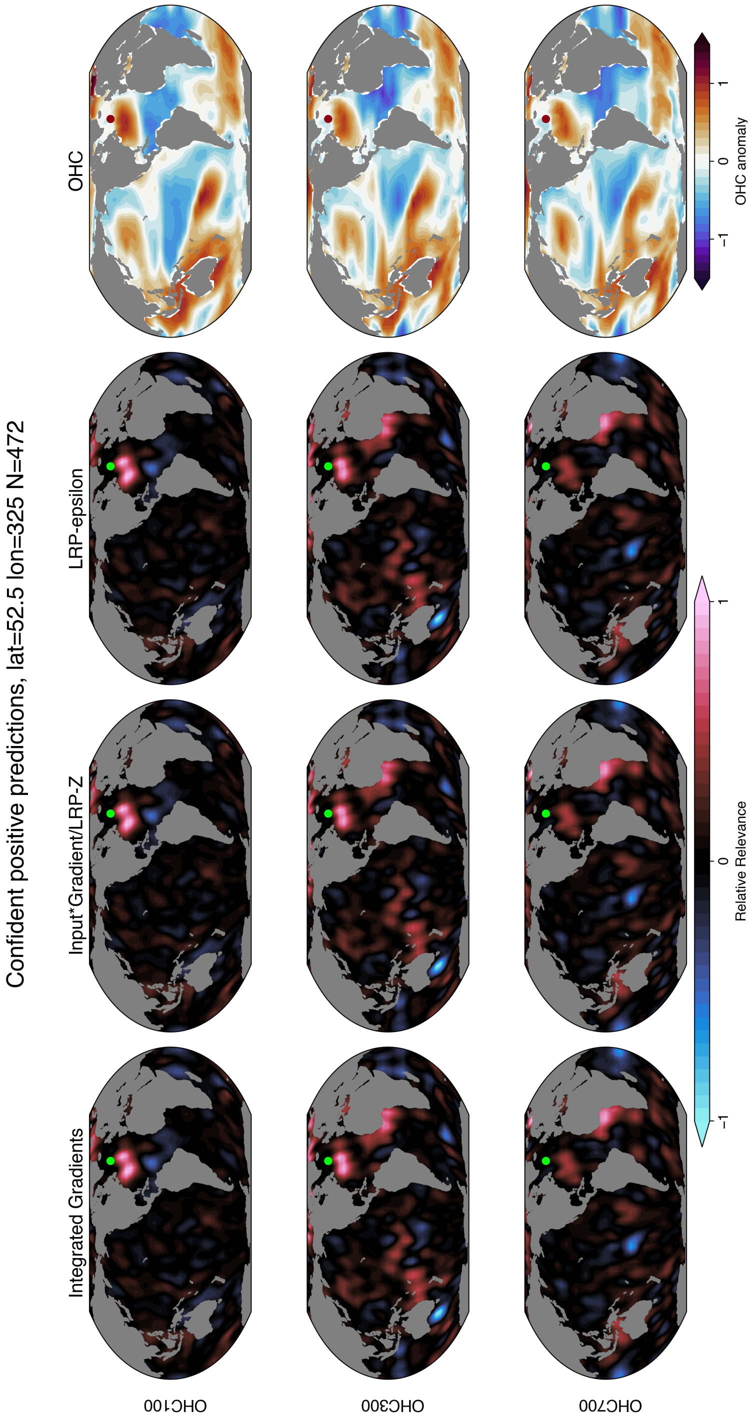

Figure S4. Composite explainability maps for predictions in Figure 3 of the main text. Each of the first three columns is a different technique (Integrated Gradients, Input times Gradient, LRP-epsilon from left to right), while each row is a different ocean layer (OHC to 100 m, OHC to 300 m, OHC to 700 m from top to bottom). The right-most column is the composite OHC input (the same as Fig 3a-c).

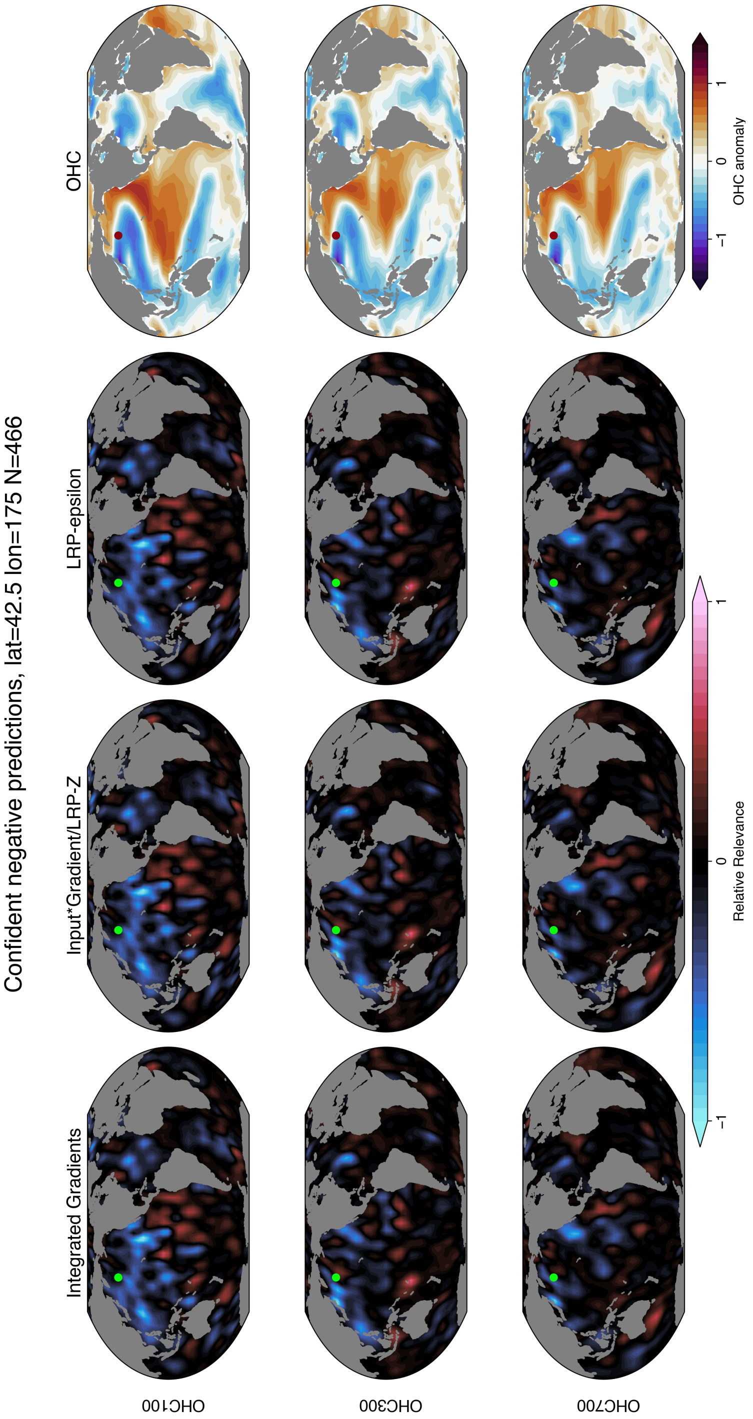

Figure S5. As Figure S4 but for North Pacific predictions in Figure 4 in the main text.