GEOPHYSICAL RESEARCH LETTERS

Supporting Information for "Physical Insights from the Multidecadal Prediction of North Atlantic Sea Surface Temperature Variability Using Explainable Neural Networks"

Glenn Liu[1,3], Peidong Wang[2], Young-Oh Kwon[3]

[1]MIT-WHOI Joint Program in Oceanography/Applied Ocean Science and Engineering

[2]Department of Earth, Atmospheric, and Planetary Sciences, Massachussetts Institute of Technology, Cambridge, MA 02139

[3]Physical Oceanography Department, Woods Hole Oceanographic Institution

Introduction

For the supporting information, we provide details on the hyperparameters of the fully-connected neural network (FNN) used in this project (Table S1). We compare the performance between FNNs and convolutional neural networks (CNNs) (Fig. S1). Additional figures are also provided for different cases discussed in the main text. They demonstrate that the the main conclusions are not sensitive to these different cases.

Table S1. Neural Network Architecture and Training Hyperparameters used in the Fully-Connected Neural Network (FNN)

Number of layers	4
Neurons per layer	128
Activation Function	Rectified Linear Unit (ReLU)
Dropout Percentage*	50%
Max Epochs	50
Early Stoppping	5 Epochs of Increasing Loss
Mini Batch Size	32
Optimizer	Adam
Learning Rate	1 x 10^{-3}

*Dropout layer included prior the the last layer.

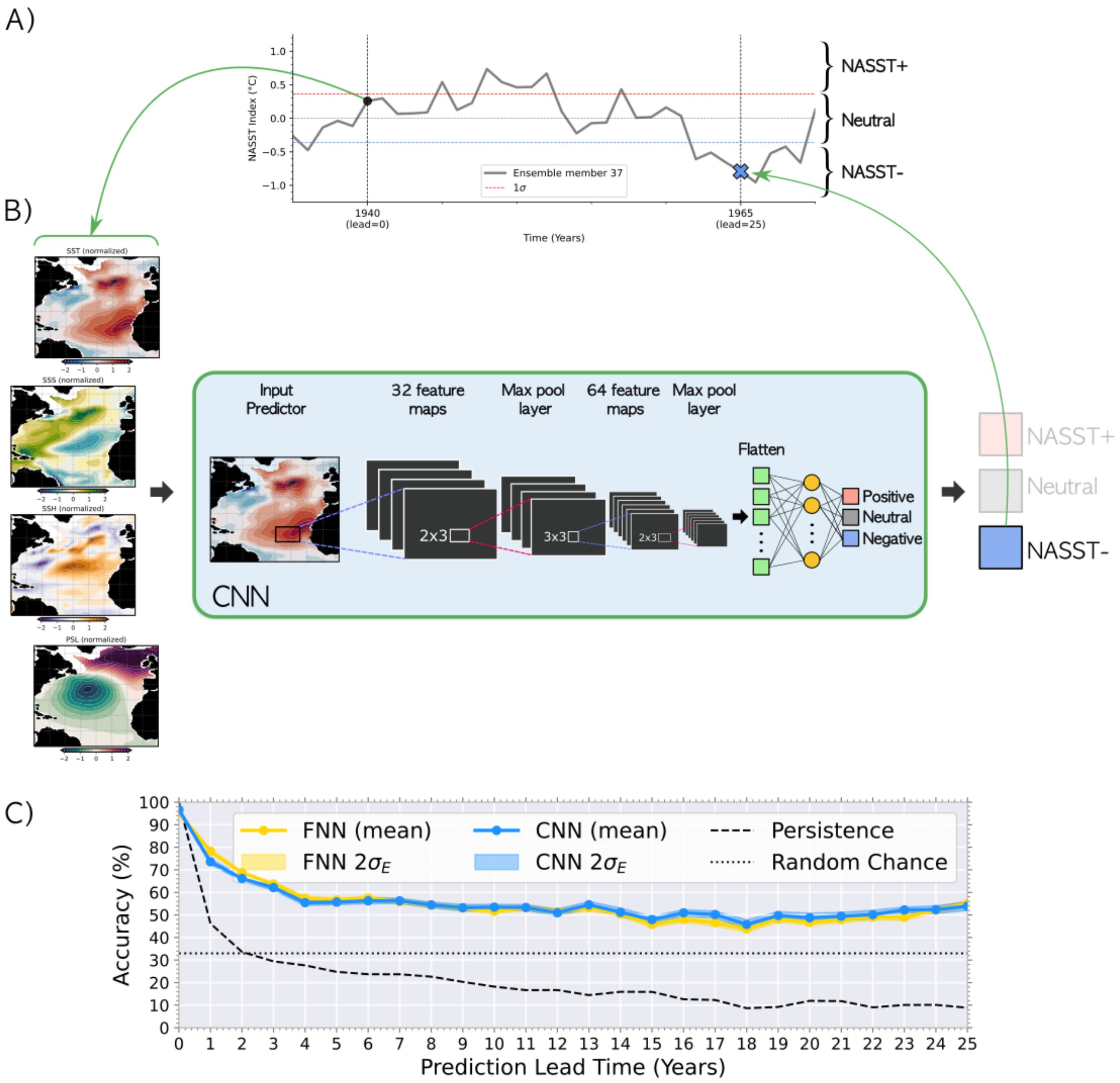

Figure S1. Schematic diagram of an example AMV Prediction problem (A) for the 2-layer CNN (B). The comparison in positive and negative North Atlantic Sea Surface Temperature (NASST+, NASST-) test accuracy between the FNN (yellow) and CNN (blue) for an SST predictor (C), with the random chance (dotted) and persistence baselines (black). Both networks perform similarly regardless of predictor, and their means are largely within the 95% standard error across initialized networks (shading).

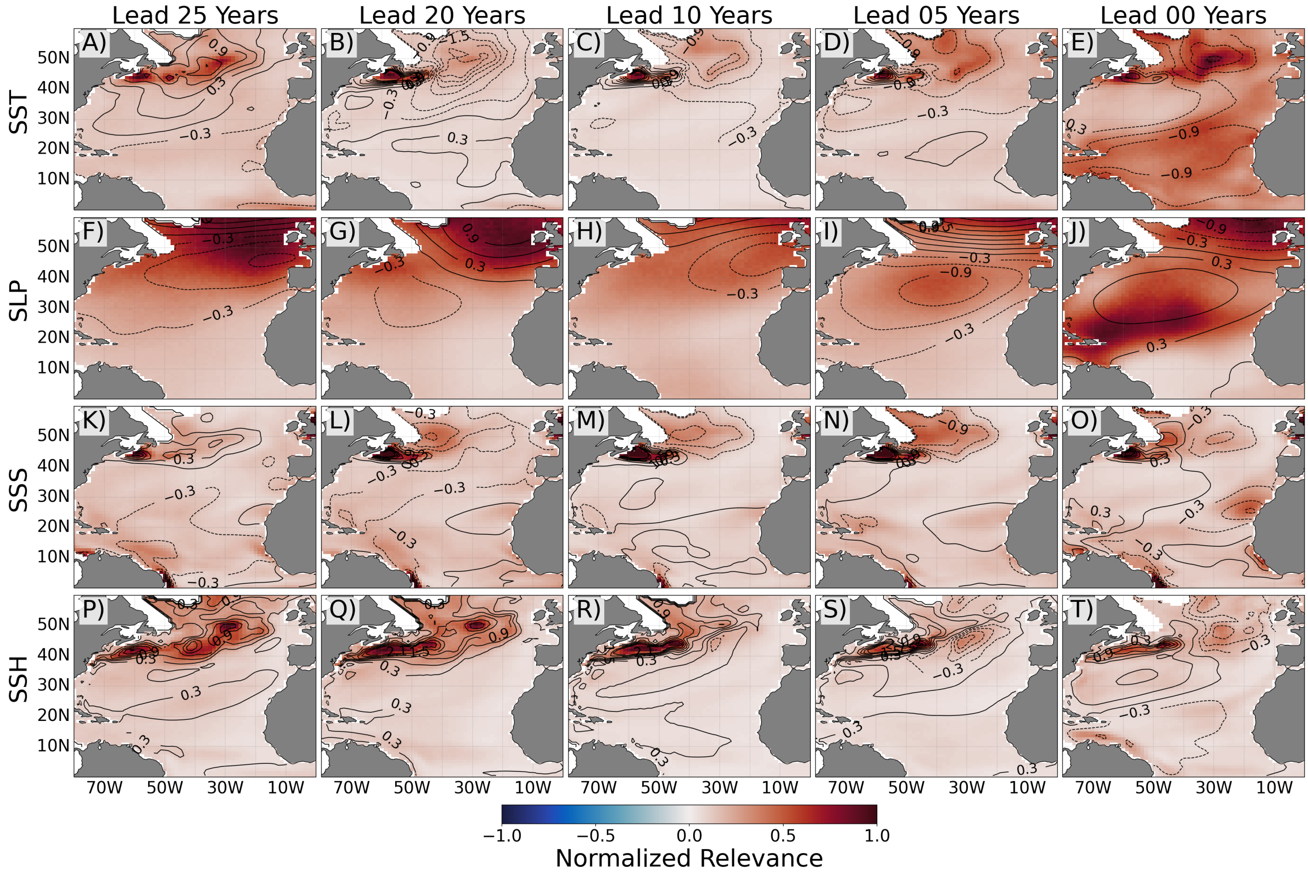

Figure S2. Same as Figure 3, but for “correct” NASST- predictions for the top 50 performing networks. The regions of high relevance, i.e., sources of predictability, resemble that of NASST+, though there are small differences. The AMV maximum in the central subpolar gyre is more distinctly outline for SST at leadtime 0 (Panel E). Additionally, the NN focuses on anomalies closer to the Azores High at 5-year leadtimes, rather than directly to the Iceland low as in the NASST+ case (Panel I).

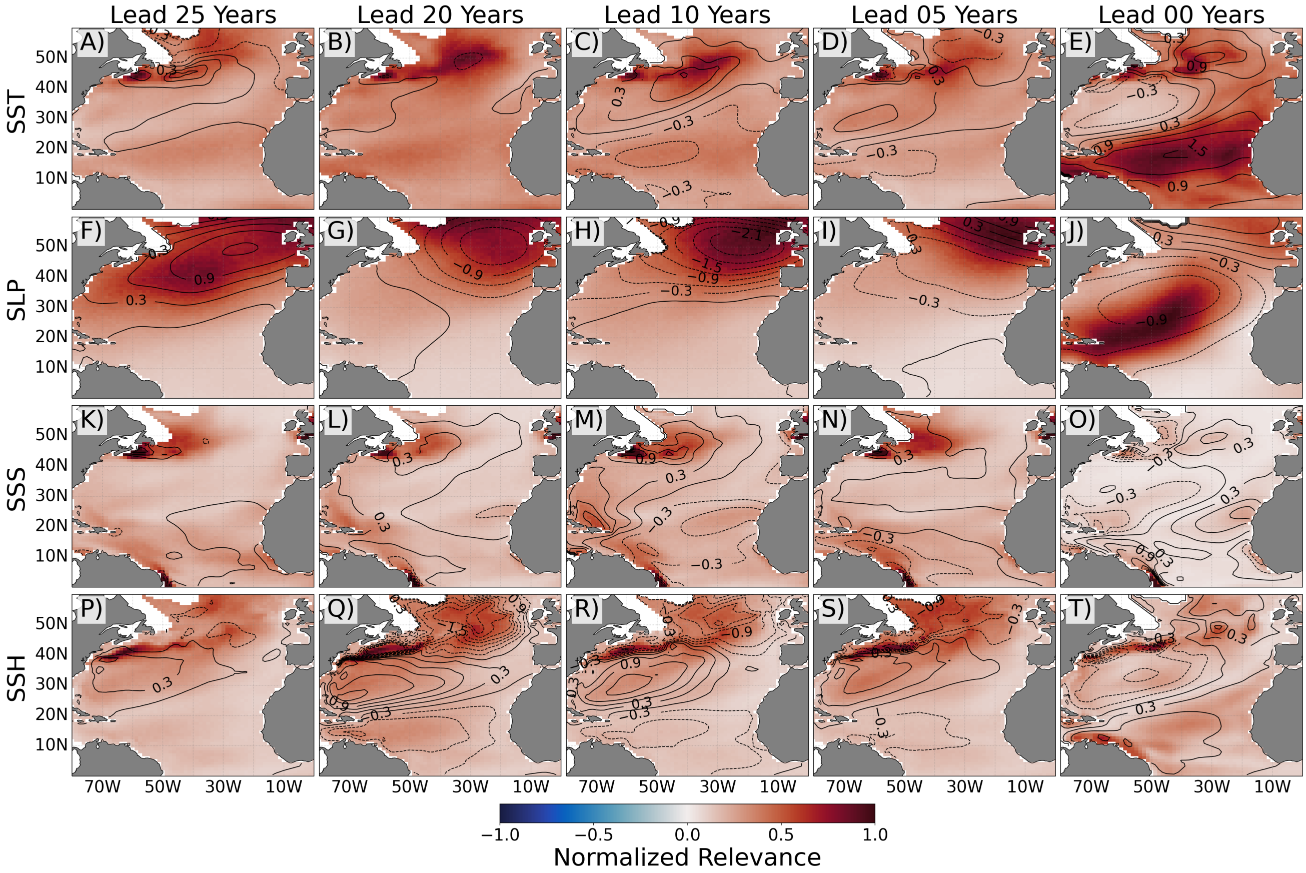

Figure S3. Same as Fig. 3, but for the unforced case where the external forcing was removed. The regions of maximum relevance resemble that of the forced NASST+ predictions. This similarity between both forced and unforced cases suggests that the NASST predictability is sourced from similar regions and dynamics, though further work is needed to explicitly investigate the responsible processes.

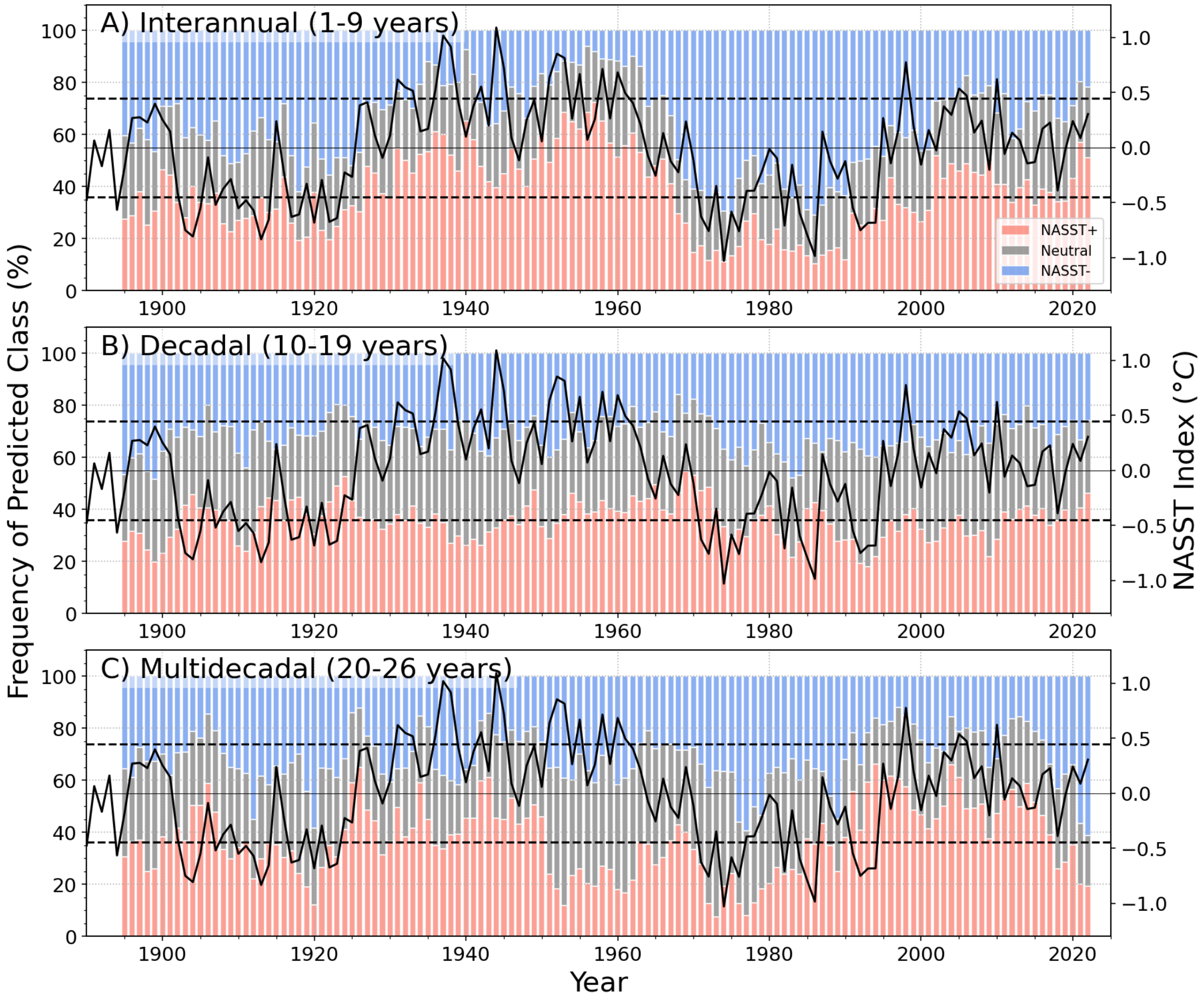

Figure S4. Same as Figure 4, but for NNs trained with unforced CESM1 data predicting the NASST Index from HadISST, detrended by removing a cubic fit. The result is not sensitive to the detrending method.