

1 **Physical Insights from the Multidecadal Prediction of**  
2 **North Atlantic Sea Surface Temperature Variability**  
3 **Using Explainable Neural Networks**

4 **Glenn Liu<sup>1,3</sup>, Peidong Wang<sup>2</sup>, Young-Oh Kwon<sup>3</sup>**

5 <sup>1</sup>MIT-WHOI Joint Program in Oceanography/Applied Ocean Science and Engineering

6 <sup>2</sup>Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology,  
7 Cambridge, MA 02139

8 <sup>3</sup>Physical Oceanography Department, Woods Hole Oceanographic Institution

9 **Key Points:**

- 10 • Neural networks outperform persistence forecasts in predicting extreme states of  
11 North Atlantic sea surface temperature out to 25 years  
12 • An explainable neural network technique reveals successful predictions rely consis-  
13 tently on the Transition Zone Region  
14 • Predictions by neural networks trained on model output captures the phasing of  
15 multidecadal variability on an observation-based dataset

---

Corresponding author: Glenn Liu, [glennliu@mit.edu](mailto:glennliu@mit.edu)

**Abstract**

North Atlantic sea surface temperatures (NASST), particularly in the subpolar region, are among the most predictable locations in the world's oceans. However, the relative importance of atmospheric and oceanic controls on their variability at multidecadal timescales remain uncertain. Neural networks (NNs) are trained to examine the relative importance of oceanic and atmospheric predictors in predicting the NASST state in the Community Earth System Model 1 (CESM1). In the presence of external forcings, oceanic predictors outperform atmospheric predictors, persistence, and random chance baselines out to 25-year leadtimes. Layer-wise relevance propagation is used to unveil the sources of predictability, and reveal that NNs consistently rely upon the Gulf Stream-North Atlantic Current region for accurate predictions. Additionally, CESM1-trained NNs do not need additional transfer learning to successfully predict the phasing of multidecadal variability in an observational dataset, suggesting consistency in physical processes driving NASST variability between CESM1 and observations.

**Plain Language Summary**

North Atlantic sea surface temperatures, particularly in the subpolar region, are among the most predictable locations in the world's oceans. However, it remains uncertain if processes in the atmosphere or ocean are more important for driving temperature fluctuations in this region occurring over multiple decades. We use a machine learning approach and train a neural network to predict the sea surface temperature state from climate model outputs, given snapshots of atmospheric or oceanic variables. Ocean variables lead to more accurate predictions relative to atmospheric variables and standard prediction baselines out to 25 years ahead if processes that drive the trends in climate, such as human-induced warming, are present in the data. These successful predictions arise consistently from the same region near the Gulf Stream-North Atlantic Current region. Despite being trained on climate models, the neural networks can predict the timing of observed positive and negative states of real-world sea surface temperatures, suggesting that there is potential for using model output to train neural networks at predicting the actual North Atlantic sea surface variability.

**1 Introduction**

Sea surface temperature (SST) anomalies averaged over the North Atlantic region exhibit alternating warm and cold periods on decadal timescales, known as the Atlantic Multidecadal Variability (AMV, or Atlantic Multidecadal Oscillation). The societal relevance of predicting AMV is underscored by linkages to multidecadal variations across multiple Earth system processes both within and beyond the North Atlantic (Zhang et al., 2019; Ruprich-Robert et al., 2021, and references therein). However, the dominant driver of AMV remains highly contested; leading contenders include ocean dynamics (Kim et al., 2018; Zhang et al., 2019; Arzel et al., 2022), atmospheric dynamics (Clement et al., 2015; Cane et al., 2017), and variations in external forcing (L. N. Murphy et al., 2021; Klavans et al., 2022). Each of these drivers imply different timescales of predictability, and the short observational record further complicates the disentanglement of their contributions.

Yet the subpolar North Atlantic (SPNA), the center of action for AMV, is considered among the most predictable locations for SST and ocean heat content across all ocean basins, with skill extending to decadal timescales (Buckley et al., 2019; Yeager, 2020). Mean wintertime mixed-layer depths reach over 1000 meters within the SPNA, resulting in large heat capacity that translates to long persistence and memory of SST anomalies (Deser et al., 2003; Holte et al., 2017). The SPNA encompasses key deep-water formation sites of the Atlantic Meridional Overturning Circulation (AMOC), and has been linked to multi-year to

64 multi-decadal predictability, both locally and in other regions such as the tropical Atlantic  
 65 (Dunstone et al., 2011; Menary et al., 2015).

66 Current state-of-the-art approaches for decadal prediction of the climate system are  
 67 often computationally intensive and highly sensitive to initial conditions, or constrained  
 68 by assumptions of linearity in simplified models such as the Linear Inverse Model (Zanna,  
 69 2012; Huddart et al., 2017; Smith et al., 2019; Meehl et al., 2022). An alternative pathway  
 70 emerges from neural networks (NN) and their ability to capture nonlinear processes and  
 71 transformations (Hornik et al., 1989; Toms et al., 2020). NNs have successfully outperformed  
 72 dynamical forecasts of El Niño-Southern Oscillation (ENSO) at interannual timescales (Ham  
 73 et al., 2019) and detecting transitions between positive and negative states of the Pacific  
 74 Decadal Oscillation (Gordon et al., 2021). Furthermore, recent developments of techniques  
 75 such as Layer-wise Relevance Propagation (LRP) provide a way to peer into the “black  
 76 box” of the NNs and identify the critical features for skillful predictions (Toms et al., 2020;  
 77 Gordon et al., 2021; Wang et al., 2022). In this work, we investigate the potential of applying  
 78 NNs to predicting NASST and use LRP to examine the relative importance of atmospheric  
 79 and oceanic sources of predictability across multiple timescales.

## 80 2 Methods and Data

### 81 2.1 Datasets

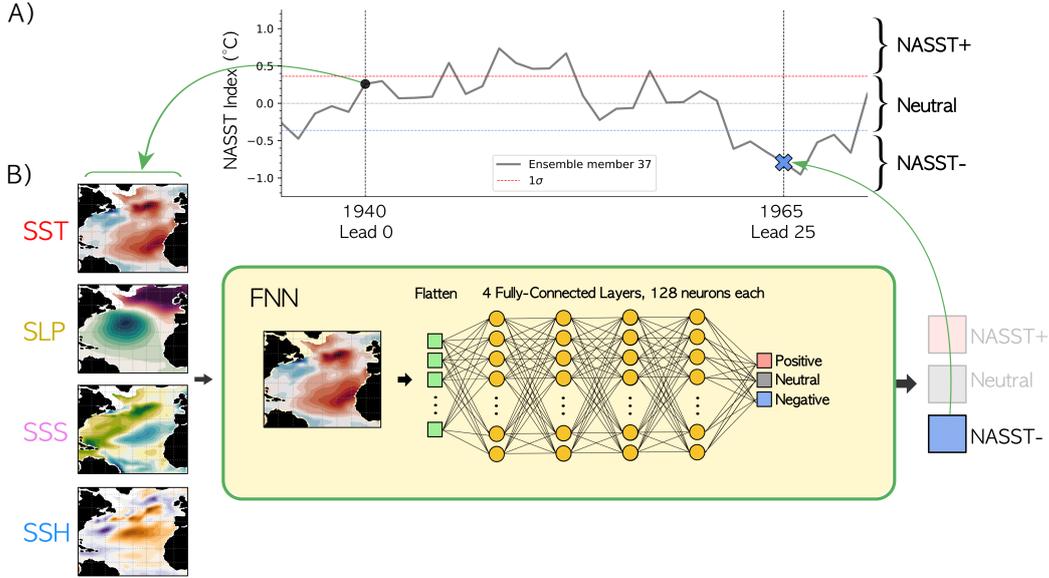
82 We use the Community Earth System Model 1 (CESM1) Large Ensemble Simulations  
 83 (LENS) based on a fully-coupled global climate model with nominal 1-degree resolution  
 84 (Kay et al., 2015). We focus on a single model to investigate if NNs can learn the physics  
 85 of NASST variability, without confounding factors and biases that arise from cross-model  
 86 comparisons. CESM1 LENS features 42 members under the same external forcing but  
 87 with slightly different atmospheric initial conditions, representing a comprehensive range  
 88 of intrinsic climate variability. We use the historical period common across all ensemble  
 89 members (1920 to 2005), totaling of 3,612 years of data for training, validation, and testing  
 90 of the NNs.

91 To investigate if the predictability learned from CESM1 translates to a realistic dataset,  
 92 we test the NNs on an observational dataset, the Hadley Center Sea Ice and Sea Surface  
 93 Temperature (HadISST) that includes monthly data between 1870 and 2022 at 1-degree  
 94 resolution (Rayner et al., 2003). Since the NNs require inputs of the same size, we re-grid  
 95 HadISST to match the CESM1 resolution using bilinear interpolation.

### 96 2.2 Prediction Objective

97 The input features are 2-D annual mean snapshots of atmospheric and/or oceanic pre-  
 98 dictors (discussed in Section 2.3) over the North Atlantic (80 to 0°W, 0 to 65°N), and  
 99 the output prediction is the state of NASST (either positive, negative, or neutral) a given  
 100 number of years later (Fig. 1). The NASST index is the area-weighted, annual mean SST  
 101 anomaly over the North Atlantic, essentially the unfiltered AMV Index (Ting et al., 2009).  
 102 Considering recent work that suggests the importance of external forcing in driving AMV  
 103 (L. N. Murphy et al., 2021; Klavans et al., 2022), we also examine differences in predictabil-  
 104 ity of NASST *with* and *without* external forcings such as the anthropogenic warming trend,  
 105 defined by the 42-member ensemble mean (referred to as *forced* and *unforced*, respectively).

106 We focus on predicting extreme NASST states due to its strong scientific and societal  
 107 impacts. A 1-standard deviation ( $\sigma$ ) threshold is used to separate the NASST into posi-  
 108 tive, negative, and neutral states (similar results are obtained using tercile thresholds). The  
 109 threshold was selected to be high enough to distinguish extreme NASST anomalies, but  
 110 low enough to permit sufficient samples for training. To prevent biases towards predict-  
 111 ing a specific class simply due to its frequency of occurrence, following standard practice



**Figure 1.** Schematic diagram of the NN prediction of NASST state using an example NASST-event in 1965 from ensemble member 37 of CESM1 LENS (Panel A). The snapshot of a selected predictor from 25 years prior (1940) is given to a FNN (Panel B), which outputs a prediction of the NASST state.

112 (Drummond et al., 2003; Buda et al., 2018; Gordon et al., 2021), we subsample the CESM1  
 113 output during training and validation so that there are equally 300 events per NASST state.

114 **2.3 Atmospheric and Oceanic Predictors**

115 To evaluate the importance of atmospheric versus oceanic drivers for NASST variability,  
 116 we train networks to predict the NASST state given 2-D annual mean anomalies of the 4  
 117 following predictors:

- 118 1. **SST**, also used to calculate the NASST indices.
- 119 2. **Sea level pressure (SLP)**, an atmospheric predictor reflecting the state of the  
 120 dominant atmospheric modes of variability in the region, e.g., the North Atlantic  
 121 Oscillation (NAO)(Hurrell & Deser, 2010; Ruprich-Robert & Cassou, 2015).
- 122 3. **Sea surface salinity (SSS)**, an oceanic predictor that is not directly damped by  
 123 heat fluxes to the atmosphere, allowing for the investigation of redistribution and  
 124 damping by ocean circulation and its connections with NASST variability (Zhang,  
 125 2017).
- 126 4. **Sea surface height (SSH)**, an oceanic predictor used to infer geostrophic circulation  
 127 with connections to variations in the strength of subpolar gyre (Koul et al., 2020).  
 128 SSH is also related to subsurface ocean heat content with potential for long-term  
 129 predictability (Buckley et al., 2019; Yeager, 2020).

130 These predictors are observable from the ocean surface, and are thus more likely to  
 131 have longer records into the future with satellite observations, providing potential for appli-  
 132 cation to operational predictions of climate. We tested additional predictors from CESM1,  
 133 including net air-sea heat flux, barotropic streamfunction, mixed-layer depth, heat and salt

134 content, and wind stress and its curl. None of these predictors yielded significantly better  
 135 performance, so we focus on the above four variables.

136 Each predictor is cropped to the domain used to compute the NASST index. Ocean  
 137 variables are re-gridded to match atmospheric grid using bilinear interpolation. We exclude  
 138 regions over land and where the ice fraction exceeds 5% so that the NNs are given the same  
 139 areas for each predictor. We normalize each predictor by dividing by  $1\sigma$  across the time,  
 140 space, and ensemble dimensions, ensuring comparable variability between predictors and  
 141 equal numerical contribution during the training process (Singh & Singh, 2020). Multiple  
 142 NNs are trained with each of the above mentioned predictors separately. NNs that include  
 143 all predictors as input did not yield improved skill, but rather indicate equivalent accuracy  
 144 to the best predictor at each leadtime (not shown).

## 145 2.4 Network Architecture and Training Procedure

146 To separately investigate the dependency in timescale and predictor, each NN is trained  
 147 to predict the NASST state at a specific leadtime ( $t=0$  to 25 years) given one predictor at  
 148 a time. We withhold 10 members of CESM1 LENS for testing, and split the remaining  
 149 32 members into training (90%) and validation (10%) subsets. We initialize 100 different  
 150 networks to account for randomness in the training process, totaling 10,400 networks (26  
 151 leadtimes  $\times$  4 predictors  $\times$  100 initialized networks). The training and validation sets are  
 152 shuffled and resampled for each training iteration, ensuring that the results are not sensitive  
 153 to a particular subset. Each network is trained for 50 epochs, but the training process is  
 154 stopped if the validation loss increases for 5 consecutive epochs to prevent over-fitting. All  
 155 discussed results are from the withheld testing set.

156 We explored combinations of architectures and hyperparameters for convolutional neural  
 157 networks (CNNs) and fully-connected neural networks (FNNs). Both architectures  
 158 yielded comparable performance (Fig. S1C). Our preliminary results with more complex  
 159 networks did not produce significantly better results, but full exploration of other architec-  
 160 tures is left for future work. Since our objective is not to tune network hyperparameters to  
 161 maximize accuracy, but rather to gain physical insight on drivers of NASST variability by  
 162 examining inter-predictor differences, we focus on the simpler FNN in this study containing  
 163 4 layers with 128 neurons each.

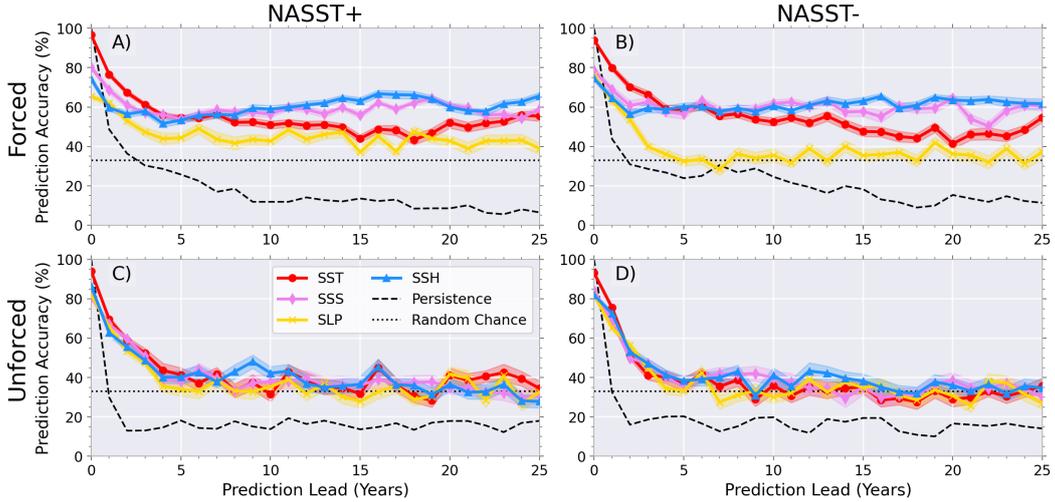
## 164 2.5 Prediction Baselines

165 We compare the accuracy of the trained NNs to two baselines. Since each class is evenly  
 166 sampled during the training, there is a 33% chance that a given class will occur, which we  
 167 set as the *random chance baseline*. We additionally examine the other extreme using the  
 168 standard *persistence baseline* that assumes uninterrupted continuation of the current state  
 169 (A. H. Murphy, 1992), which gives a stronger baseline than a damped persistence. For  
 170 example, if the system is at NASST+ at the starting time ( $t=0$  years), we assume it will  
 171 also be NASST+ for the target leadtime.

## 172 3 Higher skill from oceanic predictors at multidecadal leadtimes in the 173 presence of external forcing

174 We focus on the prediction skill for NASST+ and NASST- events (Fig. 2). For the  
 175 predictions of Neutral events, the NNs had low accuracy equivalent to random chance. This  
 176 is expected due to the challenge of predicting cases at the class boundaries or events with a  
 177 weaker signal (Batista et al., 2004).

178 In the forced case (Fig. 2A-B), NNs outperform both persistence and random chance  
 179 baselines regardless of the predictor. The atmospheric variable, SLP, has similar-to-worse  
 180 accuracy at all leadtimes compared to SST. While this is unsurprising, considering the



**Figure 2.** The mean accuracy by leadtime for predicting NASST+ and NASST- states for NNs trained with each predictor. X-axis is the prediction leadtime from 0 to 25 years. Shading indicates the 95% standard error of 100 NNs for each predictor. NNs trained with oceanic predictors SSH (blue) and SSS (pink) outperform those trained with SST (red) and SLP (yellow) at long leadtimes in the forced case (A-B). For the unforced case (C-D), performance is similar to the random chance baselines after 5-10 years (C-D).

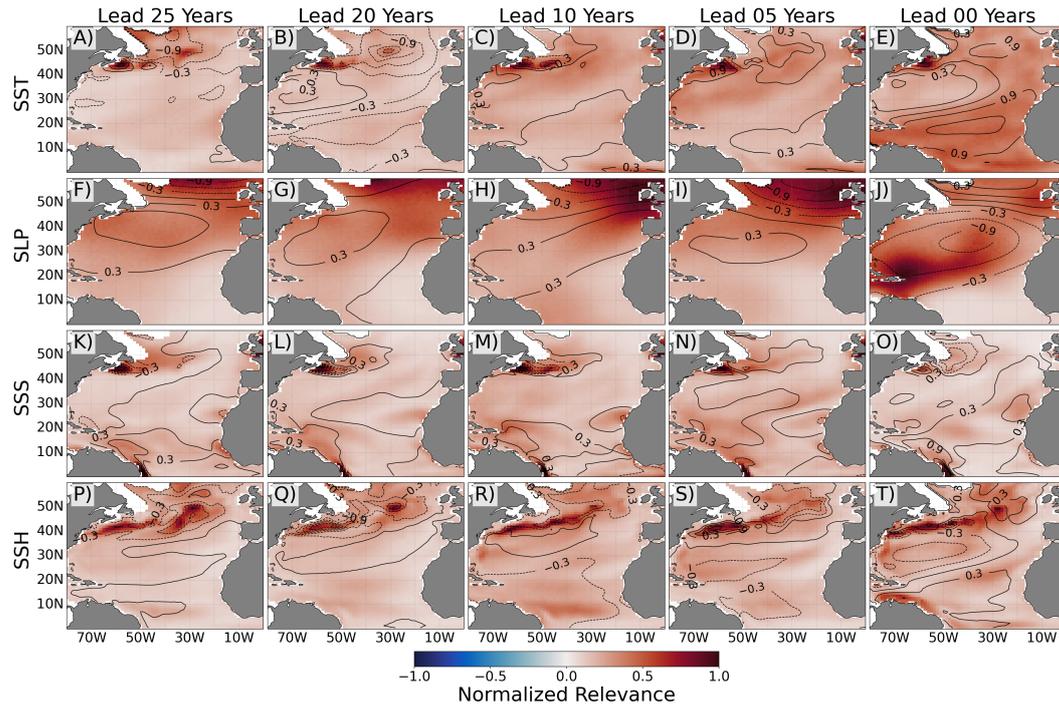
181 short persistence timescales of the atmosphere in the extratropics, on the order of weeks  
 182 (Frankignoul & Hasselmann, 1977), the NN still outperforms the persistence forecast and  
 183 the random chance baseline for predicting NASST+ at all the leadtimes.

184 While SST appears to be a better predictor at earlier leadtimes, NNs trained by both  
 185 oceanic predictors (SSS and SSH) achieve consistently higher accuracy than SST at decadal  
 186 and longer leadtimes (Fig. 2A-B). Prolonged predictability from SSS could arise from ab-  
 187 sence of strong, direct damping by turbulent heat fluxes that exists in SSTs, allowing for  
 188 more persistent SSS anomalies (Mignot & Frankignoul, 2003; Zhang, 2017). Similarly, sub-  
 189 surface heat content information present in SSH is shielded from damping by surface heat  
 190 fluxes, leading to more persistence and potential predictability relative to SST (Deser et al.,  
 191 2003; Buckley et al., 2019).

192 The increased predictability from oceanic variables is dependent upon the presence of  
 193 external forcings. After removing the ensemble mean from the predictors and NASST index  
 194 and repeating the training procedure, all NNs exhibit performance comparable to random  
 195 chance after 5-10 years with minimal inter-predictor difference. This suggests both the  
 196 importance of considering external forcing for climate prediction on multidecadal timescales  
 197 and its differing impact on predictability derived from oceanic variables.

#### 198 4 Consistent source of long-term predictability in the Transition Zone

199 We investigate the source of predictability for each predictor using LRP to examine  
 200 the network’s decision-making process (Böhle et al., 2019). LRP back-propagates the “rel-  
 201 evance” for given sample’s prediction from the final output node to the input layer of the  
 202 NN. The total relevance is conserved during this process through a series of propagation  
 203 rules, resulting in a “heatmap” of relevance indicating each pixel’s contribution to the net-  
 204 work’s final decision (Montavon et al., 2019; Samek et al., 2021). Previous works compared



**Figure 3.** Composite relevance values (color) for "correct" NASST+ predictions of the top 50 performing networks for 0- to 25-year leadtimes, for the predictors from SST (A-E), SLP (F-J), SSS (K-O) and SSH (R-T), respectively. Relevance values are normalized for each composite. SSS relevance values were doubled to aid interpretability. Contours are the respective composites of standardized predictors for the given leadtime.

205 such relevance maps with known patterns of physical processes for predicting Pacific climate  
 206 variability for possible correspondences (Toms et al., 2020; Gordon et al., 2021).

207 Since LRP produces the relevance map for a single sample, we examine the overall  
 208 learned source of predictability by compositing relevances across *correct* predictions for the  
 209 top 50 performing NNs of NASST+ and NASST-. The composites are normalized prior  
 210 to visualization to have values between 0 and 1, though the raw output relevance is of  
 211 order  $10^{-4}$ . We show relevance composites for key leadtimes between 0-25 years overlaid on  
 212 composites of input predictors at corresponding leadtimes (Fig. 3) for the forced NASST+  
 213 cases. Results are broadly consistent in unforced and for NASST- cases (Fig. S2-S3).

214 For instantaneous predictions (leadtime 0), the relevance maps resemble known patterns  
 215 associated with AMV and its drivers. For example the SST relevance map (Fig. 3E)  
 216 captures the canonical horseshoe pattern of AMV (Zhang et al., 2019). Furthermore, the  
 217 maximum relevance south of Newfoundland in SST, SSH, and SSS is collocated with the  
 218 SPNA-Gulf Stream dipole associated with AMV-related SSTs and major ocean circulation  
 219 features (Zhang, 2008; Nigam et al., 2018; Oelsmann et al., 2020; Gu & Gervais, 2022).  
 220 Interestingly, a second relevance maxima for SSS is present near the Amazon River outflow  
 221 region, though further investigation is needed to determine if this is a model-dependent  
 222 feature and its physical mechanisms. Overall, these aspects lend confidence that the NN  
 223 has learned to rely upon regions that vary strongly with AMV and its associated ocean  
 224 drivers.

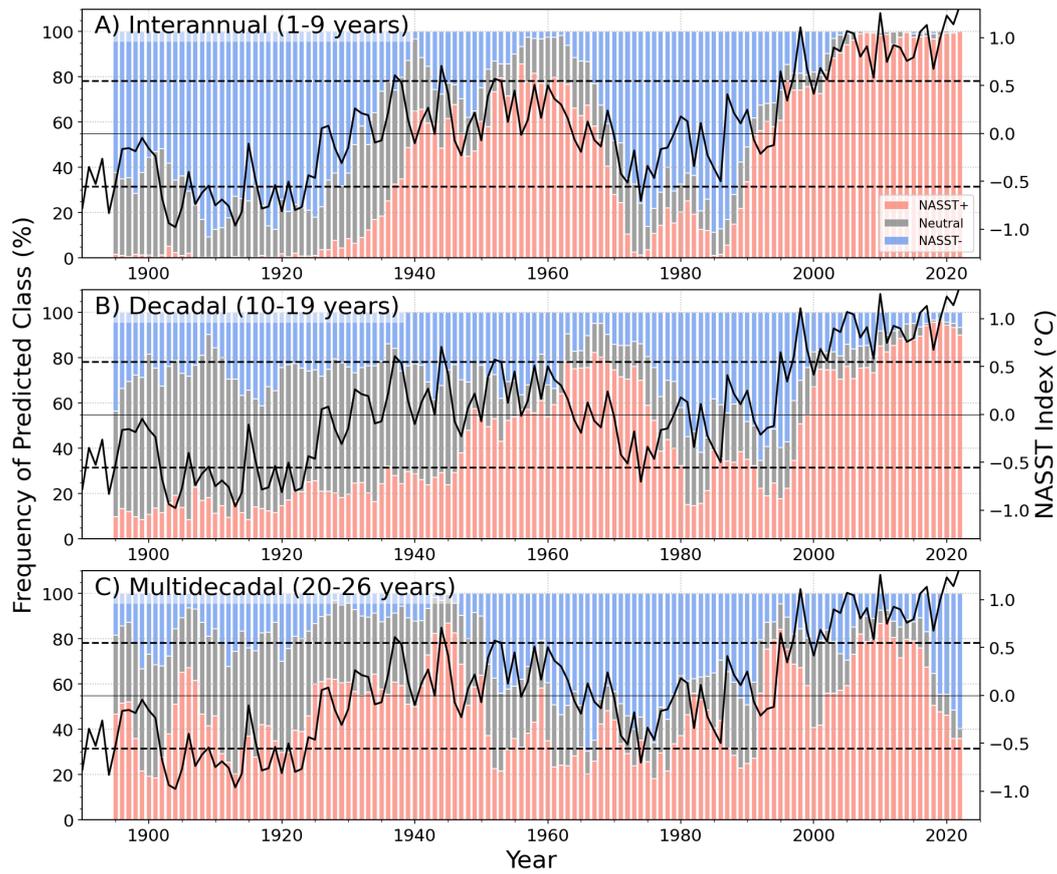
225 Patterns associated with atmospheric drivers of NASST variability also emerge in rel-  
 226 evance maps at leadtimes longer than 5 years (Fig. 3F-I). Successful predictions by SLP-  
 227 trained NNs rely upon negative SLP anomalies near the Icelandic Low in the northeastern  
 228 Atlantic, a center of action for NAO (Hurrell & Deser, 2010; Deser et al., 2010). This learned  
 229 reliance on the NAO-NASST linkage without additional input is encouraging, suggesting  
 230 that additional predictability beyond the persistence baseline achieved by SLP-trained NNs  
 231 may arise from large-scale air-sea interaction in this region and resulting ocean circulation  
 232 anomalies.

233 The Transition Zone between the subpolar and subtropical gyres emerges as a consis-  
 234 tently important region for predicting NASST regardless of leadtime for oceanic predictors  
 235 (Fig. 3K-T) (Buckley & Marshall, 2016). This region is influenced by AMOC and its as-  
 236 sociated fingerprint in surface and subsurface temperatures (Zhang, 2008). Relevance over  
 237 this region remains high irrespective of the class (NASST+ or NASST-) or the presence of  
 238 external forcing (Fig. S3). Since the NNs can derive multidecadal predictability of NASST  
 239 by focusing on a region strongly influenced by AMOC, this result highlights the poten-  
 240 tial importance of ocean dynamics for determining the state of both forced and unforced  
 241 NASST.

## 242 **5 CESM1-trained neural networks predict the multidecadal oscillation** 243 **of observed NASST states**

244 Does the NNs' skill for NASST prediction apply beyond the CESM1 model world?  
 245 Because of the limited observational record of SSH, SSS, and SLP, we test if NNs trained  
 246 on CESM1 *SSTs* can successfully predict the NASST state in HadISST. Accounting for  
 247 reductions due to the 25 year leadtime, there remains 128 years of data between 1895 to 2022.  
 248 The  $1\sigma$  threshold ( $0.55^\circ\text{C}$ ) yielded 29 (17) NASST+ (NASST-) events. The distribution is  
 249 skewed due to the warming trend. Due to the limited samples, we do not perform transfer  
 250 learning for the HadISST dataset and the accuracy values were noisy, particularly at long  
 251 leadtimes. Therefore, we focus broadly on the frequency of predictions by class (Fig. 4).

252 The frequency of predictions by class across all NNs aligns with the multidecadal oscil-  
 253 lation of the NASST in HadISST, including larger frequency of NASST- pre-1925, between  
 254 1960-1990, and the intervening warm periods. This is true particularly for interannual and



**Figure 4.** Frequency of predicted class of each target year aggregated for interannual (1-9 years) (A), decadal (10-19 years) (B), and multidecadal (20-25 years) (C) lead times for the HadISST (in colored bars). Blue/red/gray bars are the frequency of the negative/positive/neutral NASST predictions. The NASST Index from HadISST (solid-black line) and  $1\sigma$  thresholds (dashed-black lines) are shown for reference.

255 multidecadal leadtimes (Fig. 4A,C), with shifted phasing at decadal leadtimes (Fig. 4B).  
 256 The same results are recovered for the unforced case, though the multidecadal phasing of  
 257 predictions is nearly absent for the decadal leadtimes (Fig. S4). These are surprising results  
 258 for two main reasons: The first is that the NN is not simply predicting the anthropogenic  
 259 warming trend (e.g. monotonically increasing NASST+ predictions in time), but instead  
 260 has successfully learned the non-linear, oscillatory behavior of the observed NASST index.  
 261 The second is that the weights *not* have been re-adjusted to HadISST, revealing that NNs  
 262 trained on potentially biased CESM1 output maintain their ability to predict the phasing  
 263 of observed multidecadal climate variability. Overall, this suggests promise for applying  
 264 NNs trained on model output to predicting the general details and trajectory of non-linear  
 265 multidecadal climate variability in corresponding observational datasets such as HadISST.

## 266 6 Discussion and Summary

267 We investigated the potential of applying NNs to multidecadal prediction of NASST  
 268 variability and using LRP to understand the relative contributions of oceanic and atmo-  
 269 spheric drivers. Three main conclusions of this work are:

- 270 1. NNs trained with oceanic variables can predict NASST+ and NASST- states on  
 271 multidecadal timescales, outperforming persistence and random chance baselines in  
 272 the presence of external forcing.
- 273 2. The Transition Zone emerges as consistent region from where NNs derive predictive  
 274 skill, regardless of prediction leadtime, NASST state, and the presence of the external  
 275 forcing, suggesting a connection to ocean dynamics such as AMOC.
- 276 3. NNs trained on CESM1 were able to predict the multidecadal phasing of observed  
 277 NASST states without weight readjustment, suggesting promise for training NNs  
 278 using model output for multidecadal prediction of observed climate.

279 While increased predictive skill from oceanic variables highlights the importance of  
 280 ocean dynamics for multidecadal NASST variability, we find that this depends upon the  
 281 presence of external forcing. There is little difference in skill between the predictors in  
 282 unforced case, suggesting that external forcing differently impacts predictability derived  
 283 from oceanic and atmospheric variables. A possible explanation is the larger heat capacity  
 284 of the ocean allows for the integration of the externally forced signal, leading to increased  
 285 predictability on multidecadal and longer timescales (Frankignoul & Hasselmann, 1977).

286 The high-relevance over the Transition Zone region is remarkably consistent across  
 287 timescales in both unforced and forced cases. This region corresponds to the maximum  
 288 loading in the AMOC fingerprint, suggesting that the dynamics driving both internal and  
 289 external NASST variability are collocated and linked to ocean dynamics (Zhang, 2008).  
 290 Predictability arising from a stationary feature in a single region, rather than smaller-  
 291 scale features that propagate across the domain, might also explain why the simpler FNN  
 292 performed comparably to CNNs; For predicting NASST, the absolute position of the feature  
 293 is more important than its translation invariance, erasing the advantage conferred by the  
 294 CNN architecture (Barnes et al., 2022).

295 A cautionary note is that higher accuracy from networks trained with oceanic predictors  
 296 could be a model dependent feature. Our results are focused on NNs trained with CESM1,  
 297 a coarse-resolution model with biases in the separation of the Gulf Stream and position of  
 298 the North Atlantic Current (Kirtman et al., 2012). Since our relevance maps reveal that  
 299 NNs depend upon this region for skillful predictions of NASST state, verifying the model  
 300 dependence of this aspect by training NNs with other model large ensembles, reanalyses, or  
 301 observational datasets is an important future endeavor. Considering connections between  
 302 biases in mean state and decadal variability over the SPNA, exploring correspondences

303 between the resultant relevance maps and biases in ocean circulation may unveil further  
 304 hints on the importance of ocean dynamics for NASST predictability (Menary et al., 2015).

## 305 Open Research Section

306 Datasets for this research are available in these in-text data citation references: (Kay  
 307 et al., 2015), (Rayner et al., 2003). The monthly output from the CESM1 Large Ensemble  
 308 is publicly available from the National Center for Atmospheric Research’s Climate Data  
 309 Gateway on the Earth System Grid ([https://www.cesm.ucar.edu/community-projects/  
 310 lens/data-sets/](https://www.cesm.ucar.edu/community-projects/lens/data-sets/)). Monthly variables TS, LANDFRAC, ICEFRAC, SSS, PSL, and SSH  
 311 were used for this study, and further specific instructions on accessing the output for CESM1  
 312 is detailed at this link: ([https://www.cesm.ucar.edu/community-projects/lens/data-  
 313 -sets](https://www.cesm.ucar.edu/community-projects/lens/data-sets/)). The HadISST dataset can be downloaded directly from their website ([https://  
 314 www.metoffice.gov.uk/hadobs/hadisst/](https://www.metoffice.gov.uk/hadobs/hadisst/)).

315 Software for this work is available on Zenodo (DOI: [https://doi.org/10.5281/zenodo  
 316 .8342739](https://doi.org/10.5281/zenodo.8342739)), and the corresponding linked github repository ([https://github.com/glennliu265/  
 317 predict\\_nasst](https://github.com/glennliu265/predict_nasst)). The data will be The Pytorch-LRP Software is available in from the in-  
 318 text data citation reference: (Böhle et al., 2019) and can be found in the following repository  
 319 (<https://github.com/moboehle/Pytorch-LRP>).

## 320 Acknowledgments

321 GL is supported by the Department of Defense through the National Defense Science and En-  
 322 gineering Graduate Fellowship Program. GL and Y-OK gratefully acknowledge the support  
 323 by the U.S. Department of Energy Office of Science Biological and Environmental Research  
 324 as part of the Regional and Global Model Analysis program area (DE-SC0019492). Y-OK is  
 325 also supported by National Science Foundation Division of Atmospheric and Geospace Sci-  
 326 ences Climate and Large-scale Dynamics program (AGS-2055236). PW acknowledges grant  
 327 2128617 from the Atmospheric Chemistry Division of the National Science Foundation and  
 328 support of VoLo foundation.

## 329 References

- 330 Arzel, O., Huck, T., Hochet, A., & Mussa, A. (2022). Internal ocean dynamics contribution  
 331 to north atlantic interdecadal variability strengthened by ocean–atmosphere thermal  
 332 coupling. *Journal of Climate*, *35*(24), 4605–4624.
- 333 Barnes, E. A., Barnes, R. J., Martin, Z. K., & Rader, J. K. (2022). This looks like that  
 334 there: Interpretable neural networks for image tasks when location matters. *Artificial  
 335 Intelligence for the Earth Systems*, *1*(3), e220001.
- 336 Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several  
 337 methods for balancing machine learning training data. *ACM SIGKDD explorations  
 338 newsletter*, *6*(1), 20–29.
- 339 Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-wise relevance propa-  
 340 gation for explaining deep neural network decisions in mri-based alzheimer’s disease  
 341 classification. *Frontiers in aging neuroscience*, *11*, 194.
- 342 Buckley, M. W., DelSole, T., Lozier, M. S., & Li, L. (2019). Predictability of north atlantic  
 343 sea surface temperature and upper-ocean heat content. *Journal of Climate*, *32*(10),  
 344 3005–3023.
- 345 Buckley, M. W., & Marshall, J. (2016). Observations, inferences, and mechanisms of the  
 346 atlantic meridional overturning circulation: A review. *Reviews of Geophysics*, *54*(1),  
 347 5–63.
- 348 Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance  
 349 problem in convolutional neural networks. *Neural networks*, *106*, 249–259.
- 350 Cane, M. A., Clement, A. C., Murphy, L. N., & Bellomo, K. (2017). Low-pass filtering, heat

- flux, and atlantic multidecadal variability. *Journal of Climate*, *30*(18), 7529–7553.
- Clement, A., Bellomo, K., Murphy, L. N., Cane, M. A., Mauritsen, T., Rädcl, G., & Stevens, B. (2015). The atlantic multidecadal oscillation without a role for ocean circulation. *Science*, *350*(6258), 320–324.
- Deser, C., Alexander, M. A., & Timlin, M. S. (2003). Understanding the persistence of sea surface temperature anomalies in midlatitudes. *Journal of Climate*, *16*(1), 57–72.
- Deser, C., Alexander, M. A., Xie, S.-P., & Phillips, A. S. (2010). Sea surface temperature variability: Patterns and mechanisms. *Annual review of marine science*, *2*, 115–143.
- Drummond, C., Holte, R. C., et al. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets ii* (Vol. 11, pp. 1–8).
- Dunstone, N., Smith, D., & Eade, R. (2011). Multi-year predictability of the tropical atlantic atmosphere driven by the high latitude north atlantic ocean. *Geophysical Research Letters*, *38*(14).
- Frankignoul, C., & Hasselmann, K. (1977). Stochastic climate models, part ii application to sea-surface temperature anomalies and thermocline variability. *Tellus*, *29*(4), 289–305.
- 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*(21), e2021GL095392.
- Gu, Q., & Gervais, M. (2022). Diagnosing two-way coupling in decadal north atlantic sst variability using time-evolving self-organizing maps. *Geophysical Research Letters*, *49*(8), e2021GL096560.
- Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year enso forecasts. *Nature*, *573*(7775), 568–572.
- Holte, J., Talley, L. D., Gilson, J., & Roemmich, D. (2017). An argo mixed layer climatology and database. *Geophysical Research Letters*, *44*(11), 5618–5626.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.
- Huddart, B., Subramanian, A., Zanna, L., & Palmer, T. (2017). Seasonal and decadal forecasts of atlantic sea surface temperatures using a linear inverse model. *Climate Dynamics*, *49*, 1833–1845.
- Hurrell, J. W., & Deser, C. (2010). North atlantic climate variability: the role of the north atlantic oscillation. *Journal of marine systems*, *79*(3-4), 231–244.
- Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., . . . others (2015). The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349.
- Kim, W. M., Yeager, S. G., & Danabasoglu, G. (2018). Key role of internal ocean dynamics in atlantic multidecadal variability during the last half century. *Geophysical Research Letters*, *45*(24), 13–449.
- Kirtman, B. P., Bitz, C., Bryan, F., Collins, W., Dennis, J., Hearn, N., . . . others (2012). Impact of ocean model resolution on cesm climate simulations. *Climate dynamics*, *39*, 1303–1328.
- Klavans, J. M., Clement, A. C., Cane, M. A., & Murphy, L. N. (2022). The evolving role of external forcing in north atlantic sst variability over the last millennium. *Journal of Climate*, *35*(9), 2741–2754.
- Koul, V., Tesdal, J.-E., Bersch, M., Hátún, H., Brune, S., Borchert, L., . . . Baehr, J. (2020). Unraveling the choice of the north atlantic subpolar gyre index. *Scientific Reports*, *10*(1), 1–12.
- Meehl, G. A., Teng, H., Smith, D., Yeager, S., Merryfield, W., Doblas-Reyes, F., & Glanville, A. A. (2022). The effects of bias, drift, and trends in calculating anomalies for evaluating skill of seasonal-to-decadal initialized climate predictions. *Climate Dynamics*, *59*(11-12), 3373–3389.
- Menary, M. B., Hodson, D. L., Robson, J. I., Sutton, R. T., Wood, R. A., & Hunt, J. A.

- 406 (2015). Exploring the impact of cmip5 model biases on the simulation of north atlantic  
407 decadal variability. *Geophysical Research Letters*, *42*(14), 5926–5934.
- 408 Mignot, J., & Frankignoul, C. (2003). On the interannual variability of surface salinity in  
409 the atlantic. *Climate dynamics*, *20*, 555–565.
- 410 Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-  
411 wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and*  
412 *visualizing deep learning*, 193–209.
- 413 Murphy, A. H. (1992). Climatology, persistence, and their linear combination as standards  
414 of reference in skill scores. *Weather and forecasting*, *7*(4), 692–698.
- 415 Murphy, L. N., Klavans, J. M., Clement, A. C., & Cane, M. A. (2021). Investigating the roles  
416 of external forcing and ocean circulation on the atlantic multidecadal sst variability  
417 in a large ensemble climate model hierarchy. *Journal of climate*, *34*(12), 4835–4849.
- 418 Nigam, S., Ruiz-Barradas, A., & Chafik, L. (2018). Gulf stream excursions and sectional de-  
419 tachments generate the decadal pulses in the atlantic multidecadal oscillation. *Journal*  
420 *of Climate*, *31*(7), 2853–2870.
- 421 Oelsmann, J., Borchert, L., Hand, R., Baehr, J., & Jungclaus, J. H. (2020). Linking ocean  
422 forcing and atmospheric interactions to atlantic multidecadal variability in mpi-esm1.  
423 *2. Geophysical Research Letters*, *47*(10), e2020GL087259.
- 424 Rayner, N., Parker, D. E., Horton, E., Folland, C. K., Alexander, L. V., Rowell, D., ...  
425 Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night  
426 marine air temperature since the late nineteenth century. *Journal of Geophysical*  
427 *Research: Atmospheres*, *108*(D14).
- 428 Ruprich-Robert, Y., & Cassou, C. (2015). Combined influences of seasonal east atlantic  
429 pattern and north atlantic oscillation to excite atlantic multidecadal variability in a  
430 climate model. *Climate Dynamics*, *44*, 229–253.
- 431 Ruprich-Robert, Y., Moreno-Chamarro, E., Levine, X., Bellucci, A., Cassou, C., Castruccio,  
432 F., ... others (2021). Impacts of atlantic multidecadal variability on the tropical  
433 pacific: a multi-model study. *npj climate and atmospheric science*, *4*(1), 33.
- 434 Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining  
435 deep neural networks and beyond: A review of methods and applications. *Proceedings*  
436 *of the IEEE*, *109*(3), 247–278.
- 437 Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classifica-  
438 tion performance. *Applied Soft Computing*, *97*, 105524.
- 439 Smith, D., Eade, R., Scaife, A., Caron, L.-P., Danabasoglu, G., DelSole, T., ... others  
440 (2019). Robust skill of decadal climate predictions. *Npj Climate and Atmospheric*  
441 *Science*, *2*(1), 13.
- 442 Ting, M., Kushnir, Y., Seager, R., & Li, C. (2009). Forced and internal twentieth-century  
443 sst trends in the north atlantic. *Journal of Climate*, *22*(6), 1469–1481.
- 444 Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural  
445 networks for the geosciences: Applications to earth system variability. *Journal of*  
446 *Advances in Modeling Earth Systems*, *12*(9), e2019MS002002.
- 447 Wang, P., Yuval, J., & O’Gorman, P. A. (2022). Non-local parameterization of atmospheric  
448 subgrid processes with neural networks. *Journal of Advances in Modeling Earth Sys-*  
449 *tems*, *14*(10), e2022MS002984.
- 450 Yeager, S. (2020). The abyssal origins of north atlantic decadal predictability. *Climate*  
451 *Dynamics*, *55*(7-8), 2253–2271.
- 452 Zanna, L. (2012). Forecast skill and predictability of observed atlantic sea surface temper-  
453 atures. *Journal of Climate*, *25*(14), 5047–5056.
- 454 Zhang, R. (2008). Coherent surface-subsurface fingerprint of the atlantic meridional over-  
455 turning circulation. *Geophysical Research Letters*, *35*(20).
- 456 Zhang, R. (2017). On the persistence and coherence of subpolar sea surface temperature and  
457 salinity anomalies associated with the atlantic multidecadal variability. *Geophysical*  
458 *Research Letters*, *44*(15), 7865–7875.
- 459 Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y.-O., Marsh, R., Yeager, S. G., ... Little,  
460 C. M. (2019). A review of the role of the atlantic meridional overturning circula-

461 tion in atlantic multidecadal variability and associated climate impacts. *Reviews of*  
462 *Geophysics*, 57(2), 316–375.