Skillful Multiyear Sea Surface Temperature Predictability in CMIP6 Models and Historical Observations

Frances V. Davenport[1,2], Elizabeth A. Barnes[2], and Emily M. Gordon[2,3]

[1]Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, CO.

[2]Department of Atmospheric Science, Colorado State University, Fort Collins, CO.

[3]Department of Earth System Science, Stanford University, Stanford, CA.

Corresponding author: Frances Davenport (f.davenport@colostate.edu)

Key Points

- Neural networks can learn predictable signals of internal sea surface temperature variability at 1-3, 1-5, and 3-7 year lead times
- Neural networks trained on climate model output can skillfully predict sea surface temperature variability in historical observations
- Neural network skill in predicting observed sea surface temperature variability depends on the climate model used for training

Abstract

We use neural networks and large climate model ensembles to explore predictability of internal variability in sea surface temperature anomalies on interannual (1-3 year) and decadal (1-5 and 3-7 year) timescales. We find that neural networks can skillfully predict SST anomalies at these lead times, especially in the North Atlantic, North Pacific, Tropical Pacific, Tropical Atlantic and Southern Ocean. The spatial patterns of SST predictability vary across the nine climate models studied. The neural networks identify “windows of opportunity” where future SST anomalies can be predicted with more certainty. Neural networks trained on climate models also make skillful SST predictions in historical observations, although the skill varies depending on which climate model the network was trained. Our results highlight that neural networks can identify predictable internal variability within existing climate datasets and show important differences in how well patterns of SST predictability in climate models translate to the real world.

Plain Language Summary

We train neural networks (a machine learning model) to predict sea surface temperature between 3 and 7 years in the future. The neural networks are trained using data from existing climate model simulations. The regions where neural networks make the most accurate predictions depend on which climate model is used for training. The neural networks also make accurate predictions using historical observations, which means some of the patterns learned from the climate models also apply to the real climate system. However, there are unique differences between prediction accuracy in climate models and observations, which suggests directions for future research.

1 Introduction

Skillful predictions of regional climate variability on multiyear to decadal timescales would provide valuable information for near-term societal decision making and adaptation (Findell et al., 2023; Kushnir et al., 2019). While this goal remains a significant challenge, a number of studies have shown potential for predicting patterns of internal climate variability, particularly those related to large-scale ocean variability. For example, some patterns of ocean variability thought to have predictable components on three- to-ten year timeframes include the El-Nino Southern Oscillation (ENSO), Atlantic Multidecadal Variability (AMV), and the Pacific Decadal Oscillation (PDO)(Cassou et al., 2018; Meehl et al., 2009; Van Oldenborgh et al., 2012). These oceanic patterns can also lead to predictability of important processes over land, including rainfall over the Sahel (Martin & Thorncroft, 2014), North American precipitation (Enfield et al., 2001), Atlantic Hurricane frequency (Smith et al., 2010), late winter precipitation over Western Europe (Simpson et al., 2019), and North American and European summer temperatures (Sutton & Hodson, 2005).

Many recent insights into multiyear climate prediction come from initialized decadal hindcast experiments, where model simulations are initialized to match historical observations as closely as possible, and then run for up to a decade (e.g. Delgado-Torres et al., 2022; Meehl et al., 2021; Yeager et al., 2018). The hindcast simulation can then be verified against what actually occurrs in the observations. Higher prediction skill is achieved when more ensemble members are included in a hindcast experiment, with often at least 10, and sometimes as many as 40, ensemble members used (Meehl et al., 2021). The computational expense associated with these experiments thus poses a considerable challenge for decadal prediction. Initialized simulations are also subject to model drift, which occurs when a simulation that has been initialized to match observations drifts towards it's own model climatology. How exactly initialized forecasts should be corrected to account for this drift presents an additional challenge for decadal prediction (Meehl et al., 2022; Risbey et al., 2021).

More recently, data-driven or machine learning (ML) based approaches have been used to explore multiyear climate predictability (e.g. Gordon et al., 2021; Qin et al., 2022; Toms et al., 2021). In these studies, a statistical or ML model is trained to predict a climate variable or pattern of interest using existing climate datasets. Because of the need for large amounts of training data, many (although not all) prior studies have focused on multiyear predictability within large climate model simulations. For example, Toms et al. (2021) and Gordon et al. (2021) both use 1,200 years or more from the pre-industrial control run of the Community Earth System Model Version 2 (CESM2) to analyze predictability of land surface temperatures and the PDO, respectively.

A clear benefit of ML-based approaches is the potential to learn about predictability of the climate system from existing coupled atmosphere-ocean general circulation model (GCM) simulations, reducing the need for additional initialized simulations. However, as with any approach that relies on GCM simulations, the trained ML models are subject to any biases present in the underlying simulations. A few studies have explored whether ML models trained on GCMs can make accurate predictions in observations. For example, Labe and Barnes (2022) show that a neural network trained on CESM2 can predict observed global warming slowdowns. Ham et al. (2019) show skillful predictions of observed ENSO variability with up to 17 month lead times using a neural network trained on simulations from different GCMs. These studies show potential for using ML models to predict observed climate variability, but whether or not multiyear predictability in climate models reflects predictability of the real climate system more broadly is still an open question.

Here, we analyze the predictability of sea surface temperature (SST) using neural networks and historical simulations from the Coupled Model Intercomparison Project Phase 6 (CMIP6) archive (Eyring et al., 2016). We focus specifically on predicting internal variability of SSTs at interannual (1-3 year) and decadal (1-5 and 3-7 year) timescales, and apply our analysis globally. In order to have sufficient training data, we analyze GCMs that have at least 30 historical simulations. After evaluating SST predictability within each GCM, we analyze whether the information learned by the neural networks can lead to accurate SST predictions when tested on historical observations. Our goal is (i) to provide an overview and comparison of patterns of SST predictability across different GCMs in the CMIP6 archive and (ii) to identify regions where the SST predictability learned from GCMs provides the most skillful predictions of the real ocean.

2 Materials and Methods

2.1 CMIP6 data

We analyze monthly SST data from nine GCMs that have at least 30 historical simulations in the CMIP6 archive: *ACCESS-ESM1-5* (Ziehn et al., 2020), *CanESM5* (Swart et al., 2019), *CNRM-CM6-1* (Voldoire et al., 2019), *GISS-E2-1-G* (Kelley et al., 2020), *IPSL-*

CM6A-LR (Boucher et al., 2020), *MIROC-ES2L* (Hajima et al., 2020), *MIROC6* (Tatebe et al., 2019), *MPI-ESM1-2-LR* (Mauritsen et al., 2019), and *NorCPM1* (Bethke et al., 2021). The historical simulations span 1850-2014, giving a total of 4,950 model-years for each GCM.

Before neural network training, we preprocess the data for each GCM. First, we regrid all climate model output to a common 5°x5° latitude-longitude grid. We analyze latitudes between 65S to 65N. We calculate 12-month, 36-month and 60-month average SSTs at each grid point. From each time series (12-month, 36-month and 60-month averages), we subtract the ensemble-mean for each year at each grid point. By removing the ensemble mean response to external forcing, we focus our analysis on learning predictable components of internal climate variability. Once the ensemble mean is removed, we calculate the mean and standard deviation of SSTs at each grid point and use these to calculate standardized SST anomalies at each grid point at each timestep. Lastly, we calculate tercile limits at each grid point that are used to classify each SST anomaly as negative (bottom third), neutral (middle third), and positive (top third). The tercile limits are calculated separately for each simulation because some simulations are consistently cooler or warmer than the ensemble mean over the historical simulation period. Calculating the terciles separately creates a balanced number of negative, neutral, and positive anomalies within each simulation.

2.2 Neural network architecture and training

We train convolutional neural networks (CNNs) to predict SST anomalies using the GCM output (Figure 1). The CNN takes four global maps of prior SSTs as input. These maps correspond to SSTs averaged over 0-1 years, 1-2 years, 2-3 years, and 3-8 years prior. While variables such as ocean heat content may also be useful predictors, we only use sea surface temperature so that we can test the CNN using globally available sea surface temperature observations (see Section 2.4). For each set of input maps, the CNN predicts the SST anomaly at a given location (one grid cell) at a given time in the future. Each prediction is the relative likelihood of three categories: positive SST anomaly (the top tercile of historical anomalies), neutral anomaly (middle tercile), or negative anomaly (bottom tercile).

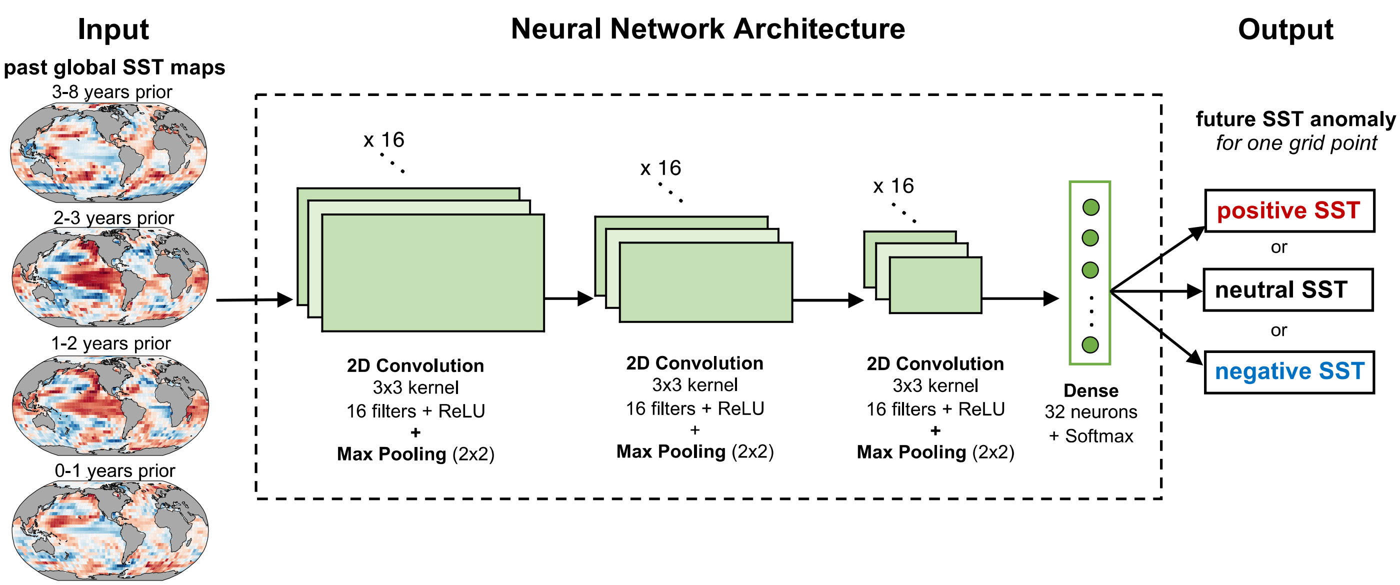

Figure 1. Overview of CNN architecture

We make SST predictions for three future time periods: years 1-3 (i.e. 36 month SST anomalies starting from the prediction date), years 1-5 (i.e. 60 month SST anomalies starting from the prediction date), and years 3-7 (i.e. 60 month SST anomalies starting 2 years after the prediction date). We train separate CNNs for each ocean grid cell, lead time, and GCM (over 30,000 CNNs in total).

We split the 30 historical simulations from each GCM into a training set of 22 simulations, a validation set of three simulations, and a test set of five simulations (*Supporting Information*, Table S1). We use hyperparameter tuning to select the CNN architecture shown in Fig. 1. Details of the hyperparameter tuning and CNN training are included in the *Supporting Information*.

2.3 Neural network accuracy and windows of opportunity

After training, we evaluate CNN performance on the testing data (five simulations per GCM). First, we calculate prediction accuracy across all testing data. We also examine whether the CNNs identify "*windows of opportunity*", or states of internal variability that are more predictable than others. We use the method from Mayer and Barnes (2021) and Gordon et al. (2023) to calculate accuracy for subsets of predictions with the highest "confidence", i.e. the samples where the CNN predicts a higher relative likelihood of one class versus the others. Higher prediction accuracy among more confident predictions indicates that the CNN has successfully identified windows of opportunity where predictions are more likely to be skillful. We calculate accuracy for the 40% and 20% most confident predictions within each testing simulation, and then average across the five testing simulations for each GCM.

We compare the neural network accuracy to a persistence model, which assumes that the future SST anomaly remains unchanged. For example, the SST anomaly prediction for year 1-5 is the same as the SST anomaly for the most recent 5 year period. Because there is no confidence associated with these predictions, we only calculate overall accuracy (not windows of opportunity).

2.4 Evaluating neural network performance on historical observations

We use the NOAA Extended Reconstructed SST Version 5 (ERSSTv5) dataset (Huang et al., 2017) to evaluate how well the trained CNNs can predict historical internal SST variability. The ERSSTv5 dataset includes global coverage at 2°x2° resolution from 1854 to present. We analyze monthly SST averages from January 1854 through October 2022. We perform similar preprocessing steps as for the GCM simulations. We regrid to the same 5°x5° grid and calculate 12-, 36-, and 60-month moving averages. Then, instead of subtracting the GCM ensemble mean, we subtract the third-order polynomial trend from each grid cell to remove any long-term forcing. We then calculate grid-cell means, standard deviations, and tercile thresholds.

In analyzing CNN predictions on the ERSSTv5 data, we focus specifically on windows of opportunity by looking at the accuracy of the top 20% most confident predictions. We also calculate the accuracy of persistence predictions within the ERSSTv5 data as a baseline comparison.

3 Results and Discussion

The CNN accuracy results are shown for one model, *IPSL-CM6-LR*, in Figure 2, with the remaining models shown in Fig. S2-S9 (*Supporting Information*). Because we have removed the forced response from the GCM simulations, these maps show the accuracy of predicting internal SST variability.

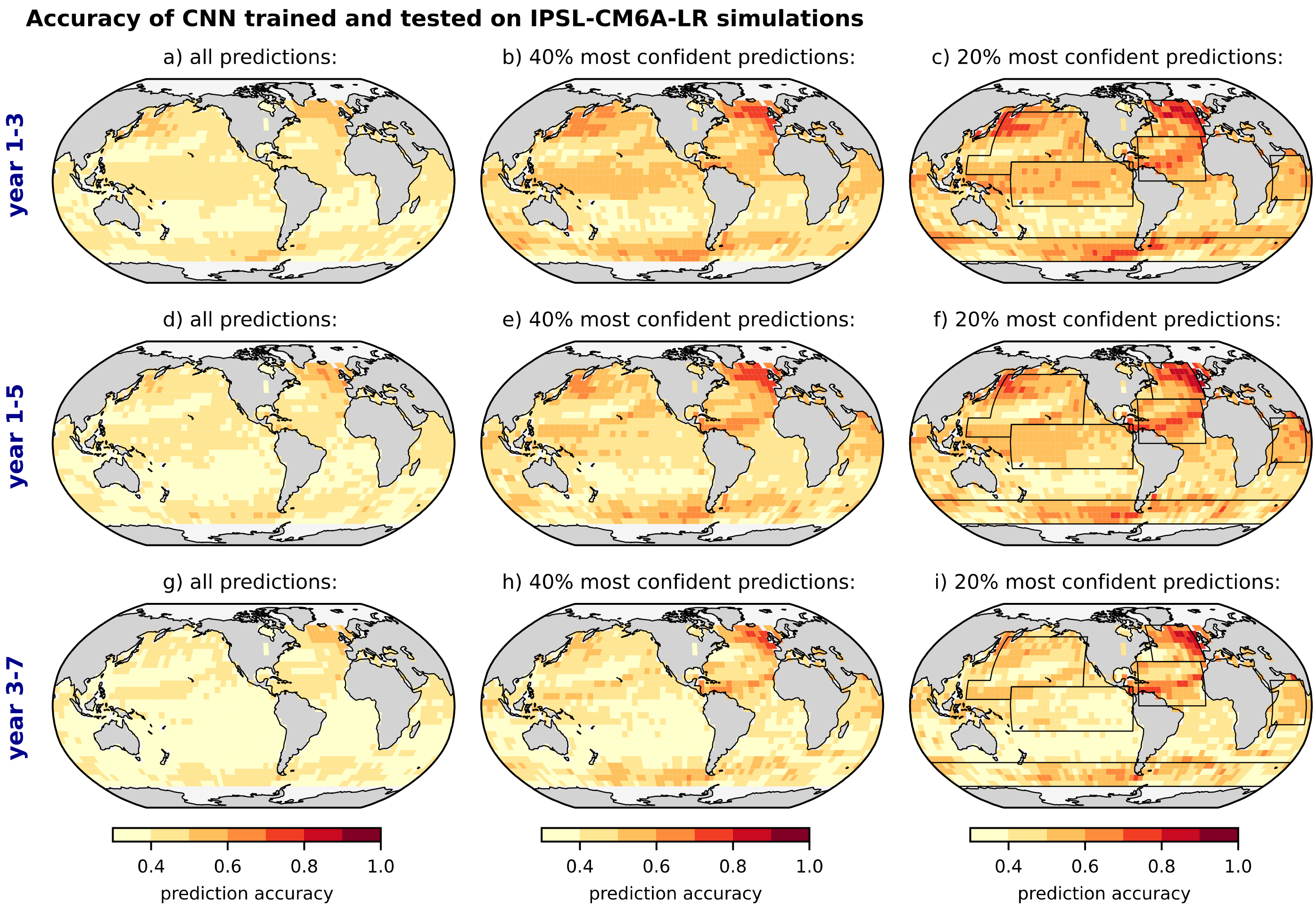

Figure 2. Accuracy of 1-5 year SST predictions using the CNNs trained and tested on *IPSL-CM6A-LR* simulations. **a)** accuracy calculated across all predictions in the test set. **b)** accuracy calculated for the 40% most confident predictions in the test set (see Methods). **c)** same as b) but for the 20% most confident predictions. Black boxes indicate regions shown in Fig 4. Other GCMs are shown in *Supporting Information*, Figs S2-S9.

Overall, we find that the prediction accuracy is higher for years 1-3, decreases for years 1-5, and is lowest for years 3-7. This pattern of higher prediction accuracy at shorter lead times is true across all nine GCMs. When accuracy is calculated across all test samples (e.g. left column of Fig. 2), the CNNs perform slightly better than the persistence model benchmark (*Supporting Information*, Fig. S10-11). However, we find that the CNNs can make much more skillful predictions during windows of opportunity, shown in the middle and right columns of Fig. 2. In some regions, prediction accuracy can approach 80% or higher for these more confident predictions (e.g. Fig. 2c, f). We find that the CNNs are able to identify windows of opportunity with higher prediction accuracy in all of the GCMs analyzed.

Regions where future SSTs are predicted most skillfully include the North Pacific, Tropical Pacific, North Atlantic, Tropical Atlantic and the Southern Ocean (defined here to refer to ocean regions between 45-65S). While many of these regions are similar across the different GCMs, there are also clear inter-model differences. For example, CNNs trained and tested on *CNRM-CM6-1* detect especially strong predictability in the North Atlantic (Fig S3). This is likely due to the stronger persistence of SSTs in North Atlantic in this GCM (*Supporting Information*, Fig. S10). The CNNs trained on *CanESM5* or *NorCPM1* have much higher accuracy in predicting SST anomalies in the Southern Ocean compared to other regions. As a third example, the CNNs trained on *GISS-E2-1-G*, *MIROC-ES2L* and *MIROC6* all show strong 1-3 year SST predictability across the tropics, including parts of the Indian Ocean.

Within each ocean basin, the spatial pattern of predictability varies depending on the GCM. For example, within the North Atlantic, many of the GCMs have the highest predictability in the subpolar North Atlantic (e.g. *ACCESS-ESM1*, *NorCPM1*). For some GCMs, though, the region of high predictability extends to include a band of high predictability in the subtropical North Atlantic (e.g. *CNRM-CM6-1*, *IPSL-CM6A-LR*). Different GCMs also have different spatial patterns of predictability in the North Pacific. Many GCMs show highest predictability in the subpolar (and especially the western subpolar) North Pacific region. Some models, such as *MIROC-ES2L* and *MIROC6*, show higher predictability in the central North Pacific. In the Southern Ocean, the most predictable region depends on both the GCM and the lead time. Many of the GCMs show high predictability across most of the Southern Ocean for year 1-3 predictions. For year 3-7 predictions, the region of high predictability generally narrows to regions of the South Pacific and South Atlantic, especially just west and east of South America (between around 160W to 0W).

After training CNNs on each GCM, we look at how well the CNNs perform when tested on ERSSTv5 observations. These results are shown in Figure 3 for the year 1-5 lead time. Year 1-3 and year 3-7 results are shown in *Supporting Information*, Fig. S12-13. We find that the CNNs are able to make skillful predictions using the ERSSTv5 observations, and that the CNN predictions outperform the historical persistence model (*Supporting Information*, Fig. S14).

The regions with the most accurate predictions in ERSSTv5 are generally the same regions that were most predictable in the GCMs, namely the North Pacific, Tropical Pacific, North Atlantic, Tropical Atlantic, and Southern Ocean. However, there are also differences in the spatial pattern of predictability between ERSSTv5 and the GCMs. As an example, in the North Pacific, the regions of highest predictability in ERSSTv5 appear similar to the PDO horseshoe pattern in the central/eastern North Pacific (e.g. Fig. 3a-e, i). In contrast, when the CNNs are evaluated on the original GCM test simulations (Fig. 2 and *Supporting Information*, Fig. 2-9), most of the GCMs lack the PDO horseshoe pattern and show the highest predictability in the western subpolar North Pacific. There are also some small regions of predictability in the ERSSTv5 observations that did not appear at all in the GCMs, such as along the coast of Chile.

As in the GCM test data, the CNN skill at predicting the ERSSTv5 observations generally decreases at the 3-7 year lead time (Fig. S13). One exception is in the North Pacific for CNNs that were trained on *ACCESS-ESM1-5*, *CNRM-CM6-1*, or *IPSL-CM6A-LR*. We find that these CNNs still make relatively skillful predictions in the North Pacific at 3-7 year lead times

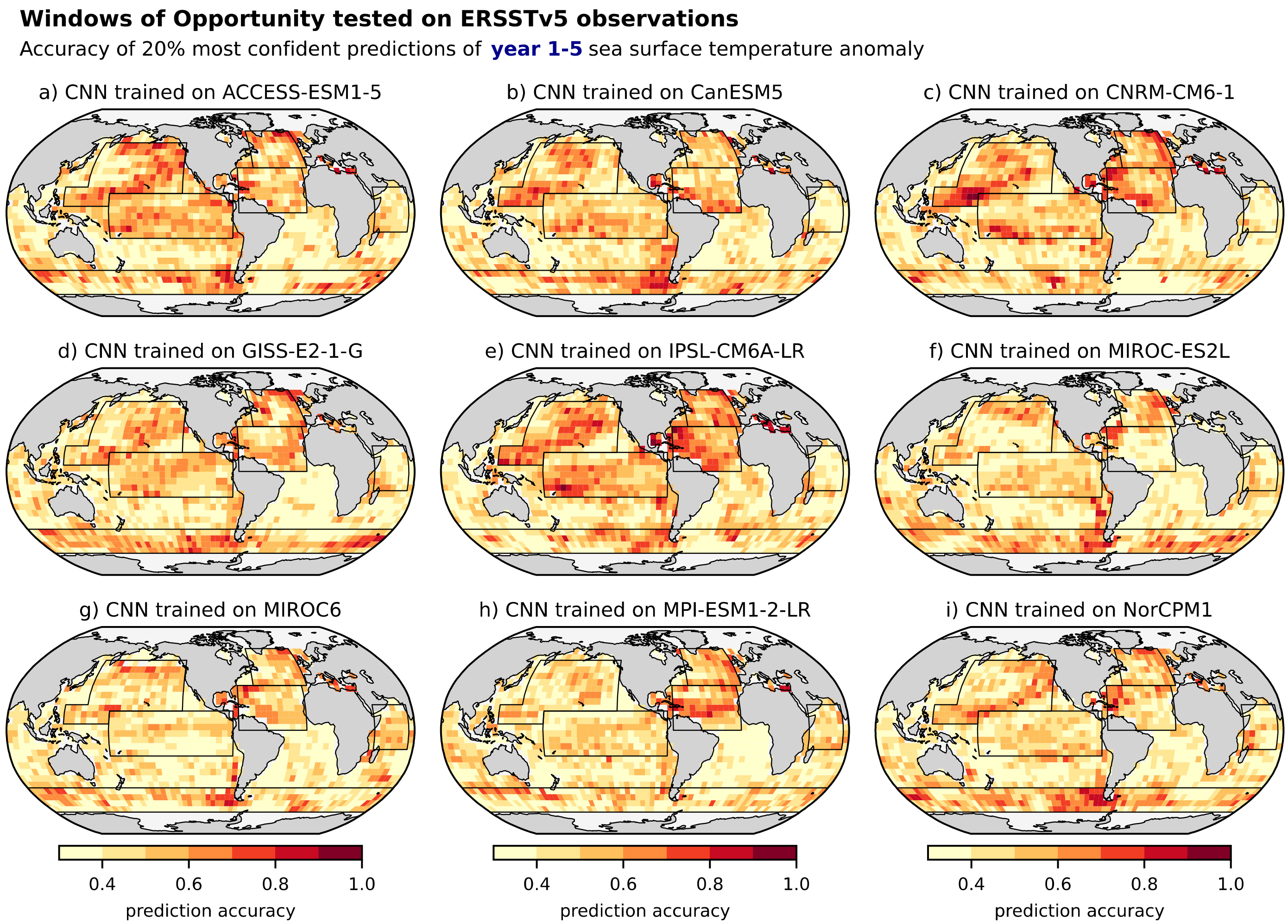

Figure 3. Accuracy of 1-5 year SST predictions for *windows of opportunity* (i.e. 20% most confident predictions) within the ERSSTv5 data. Panels show results for CNNs trained on different GCMs. Other lead times are shown in *Supporting Information*, Fig. S12-13.

when evaluated on the ERSSTv5 observations. In fact, the CNNs trained on *ACCESS-ESM1-5* and *IPSL-CM6A-LR* predict the ERSSTv5 observations in the North Pacific better than they predict their respective GCM testing data at the 3-7 year lead time (Fig. 4f).

Figure 4 summarizes the CNN performance on the GCM testing data versus the ERSSTv5 observations at the global scale (Fig. 4a-c) and for the six regions with the most skillful predictions: North Pacific, Tropical Pacific, Southern Ocean, North Atlantic, Tropical Atlantic, and West Indian Ocean. There are a few interesting patterns that emerge. We find that higher predictability in a GCM does not necessarily lead to higher predictability in the ERSSTv5 observations. For example, in the North Pacific for years 1-3 and in the Tropical Pacific for years 1-3 and 1-5, the GCMs that correspond to the highest prediction accuracy have lower accuracy when the CNNs are tested on ERSSTv5 (shown by negative correlations in Fig. 4). However, in other locations, such as the Tropical Atlantic for years 1-5 and years 3-7, higher predictability in the GCM does correspond to higher predictability in ERSSTv5. For the most part, prediction

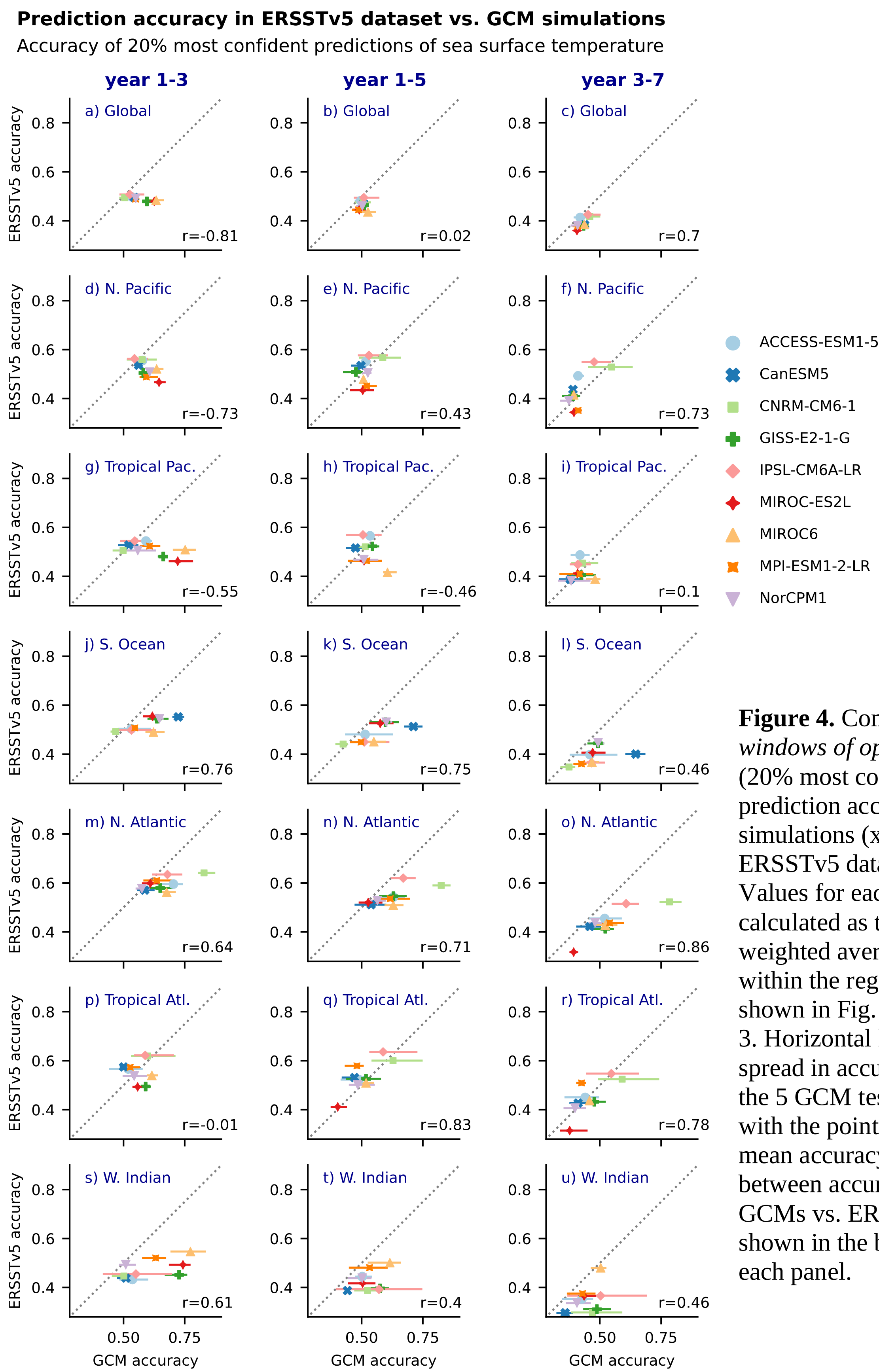

Figure 4. Comparison of *windows of opportunity* (20% most confident) prediction accuracy in GCM simulations (x-axis) vs. the ERSSTv5 data (y-axis). Values for each region are calculated as the area-weighted average accuracy within the region boundaries shown in Fig. 2c,f,i and Fig. 3. Horizontal lines show spread in accuracy across the 5 GCM test simulations, with the points showing the mean accuracy. Correlation between accuracy in the GCMs vs. ERSSTv5 is shown in the bottom right of each panel.

accuracy is higher in the original GCM test data than in the ERSSTv5 observations (shown by most points falling below the one-to-one lines). However, in addition to the example given above for the North Pacific, some CNNs can make more skillful predictions in the Tropical Pacific and Tropical Atlantic in ERSSTv5 observations than in the original GCM test data (Fig. 4h, i, p-r).

The spread in prediction accuracy across the five ensemble members in each GCM test set is shown by horizontal bars in Fig. 4. In general, the differences in predictability between different GCMs are larger than the differences in predictability between individual simulations. However, we do find that there can be substantial spread in prediction accuracy depending on both the region and the GCM. The West Indian Ocean and Tropical Atlantic have the highest spread in predictability across different simulations (although not in all GCMs). Overall, this indicates that a ~150 year record (the length of our training and testing simulations) may not be sufficient to characterize multiyear predictability at a given location, which should be taken into account when comparing predictability across individual simulations or in the historical record.

Overall, many of these results are consistent with prior studies on multidecadal climate prediction. One difference is that we measure prediction skill with classification accuracy and using the window of opportunity framework rather than metrics like the anomaly correlation coefficients. Further, many prior studies on multiyear prediction, including those that use initialized hindcast experiments, evaluate skill in predicting the combined forced response and internal variability. Still, the regions that we find have the most predictability across the GCMs and ERSSTv5 observations include many regions that have been identified in prior work, such as the North Atlantic (Borchert et al., 2021; Yeager et al., 2018; Yeager & Robson, 2017), Southern Ocean (Zhang et al., 2023), and North Pacific (Choi & Son, 2022; Gordon et al., 2021; Qin et al., 2022).

Our results also emphasize the importance of considering prediction uncertainty or confidence using the window of opportunity framework. We find many windows of opportunity for multiyear SST predictability, including for most regions, across all GCMs studied, and at all three lead times studied. These findings are aligned with other recent work demonstrating the occurrence of windows of opportunity within the climate system across multiple timescales (Gordon & Barnes, 2022; Mayer & Barnes, 2021).

One recurring question within multidecadal prediction is the occurrence of the signal-to-noise paradox, in which a climate model ensemble predicts observed variability better than it predicts individual ensemble members (Eade et al., 2014; Scaife & Smith, 2018). Here, we also find examples where the patterns learned from GCMs lead to more predictable behaviour in the observations compared to the climate models. While we do not attribute our results to the signal-to-noise paradox, it highlights additional differences in predictability between climate models and observations that could be studied in future work.

4 Conclusions

We show that machine learning, specifically convolutional neural networks, can learn patterns of global, multiyear SST predictability from existing, unitialized climate model simulations. Because our approach does not require new GCM simulations, we can efficiently analyze and compare predictability across many different GCMs. We find that the regions with the highest predictability on interannual and decadal lead times include the North Pacific, North Atlantic, Tropical Pacific, Tropical Atlantic and the Southern Ocean. However, when comparing predictability across nine GCMs, we find notable differences in the spatial patterns and magnitude of SST prediction skill. The patterns learned by the CNNs also lead to skillful predictions when tested on historical SST observations, but the amount of prediction skill in each region varies based on the GCM used for training. We also find different spatial patterns of SST predictability in the ERSSTv5 observations compared to the GCMs, although the most predictable regions are generally similar.

These results could lead to multiple future research directions. It is beyond the scope of the current study to explore why differences in SST predictability exist across GCMs and the observations. However, recent related work has shown that "explainable ML" methods can be used to understand why CNNs make certain predictions (Davenport & Diffenbaugh, 2021; Gordon et al., 2021; Labe & Barnes, 2021; Toms et al., 2020). These same methods could be applied to the CNNs used here to understand the sources of SST predictability and how they differ across GCMs and observations, providing insight into both the mechanisms involved in multiyear variability and into GCM biases in how these mechanisms are represented. Further, while the focus of this study was to explore differences in predictability across GCMs, future efforts could focus on training CNNs to produce the best predictions in the observed climate. This might be accomplished by selecting certain GCMs to use as training data for different regions, or using a combination of GCM and observational data for training through approaches like transfer learning (e.g. Ham et al., 2019). Overall, this research supports a growing body of literature that shows ML is a valuable tool for advancing the field of skillful multiyear climate prediction.

Acknowledgments

This work was funded, in part, by grant AGS-2210068 from the National Science Foundation and with special thanks to David Wallerstein.

Data Availability

We use historical simulations from the CMIP6 archive available through the Earth System Grid (https://esgf-node.llnl.gov/projects/cmip6/). We use historical sea surface temperature data from the ERSSTv5 dataset available from the National Oceanic and Atmospheric Administration (https://psl.noaa.gov/data/gridded/data.noaa.ersst.v5.html).

Code Availability

The analysis code used to train the convolutional neural networks and generate figures in the paper will be made available on github and archived using Zenodo (DOI will be created and provided here before publication).

References

Bethke, I., Wang, Y., Counillon, F., Keenlyside, N., Kimmritz, M., Fransner, F., et al. (2021). NorCPM1 and its contribution to CMIP6 DCPP. *Geoscientific Model Development*, *14*(11), 7073–7116. https://doi.org/10.5194/gmd-14-7073-2021

Borchert, L. F., Menary, M. B., Swingedouw, D., Sgubin, G., Hermanson, L., & Mignot, J. (2021). Improved Decadal Predictions of North Atlantic Subpolar Gyre SST in CMIP6. *Geophysical Research Letters*, *48*(3), e2020GL091307. https://doi.org/10.1029/2020GL091307

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., et al. (2020). Presentation and Evaluation of the IPSL-CM6A-LR Climate Model. *Journal of Advances in Modeling Earth Systems*, *12*(7). https://doi.org/10.1029/2019MS002010

Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I.-S., & Caltabiano, N. (2018). Decadal Climate Variability and Predictability: Challenges and Opportunities. *Bulletin of the American Meteorological Society*, *99*(3), 479–490. https://doi.org/10.1175/BAMS-D-16-0286.1

Choi, J., & Son, S.-W. (2022). Seasonal-to-decadal prediction of El Niño–Southern Oscillation and Pacific Decadal Oscillation. *Npj Climate and Atmospheric Science*, *5*(1), 29. https://doi.org/10.1038/s41612-022-00251-9

Davenport, F. V., & Diffenbaugh, N. S. (2021). Using Machine Learning to Analyze Physical Causes of Climate Change: A Case Study of U.S. Midwest Extreme Precipitation. *Geophysical Research Letters*, *48*(15), e2021GL093787. https://doi.org/10.1029/2021GL093787

Delgado-Torres, C., Donat, M. G., Gonzalez-Reviriego, N., Caron, L.-P., Athanasiadis, P. J., Bretonnière, P.-A., et al. (2022). Multi-Model Forecast Quality Assessment of CMIP6 Decadal Predictions. *JOURNAL OF CLIMATE*, *35*.

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*(15), 5620–5628. https://doi.org/10.1002/2014GL061146

Enfield, D. B., Mestas-Nuñez, A. M., & Trimble, P. J. (2001). The Atlantic Multidecadal Oscillation and its relation to rainfall and river flows in the continental U.S. *Geophysical Research Letters*, *28*(10), 2077–2080. https://doi.org/10.1029/2000GL012745

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. https://doi.org/10.5194/gmd-9-1937-2016

Findell, K. L., Sutton, R., Caltabiano, N., Brookshaw, A., Heimbach, P., Kimoto, M., et al. (2023). Explaining and Predicting Earth System Change: A World Climate Research Programme Call to Action. *Bulletin of the American Meteorological Society*, *104*(1), E325–E339. https://doi.org/10.1175/BAMS-D-21-0280.1

Gordon, E. M., & Barnes, E. A. (2022). Incorporating Uncertainty into a Regression Neural Network Enables Identification of Decadal State-Dependent Predictability. *Geophysical Research Letters*, *49*(e2022GL098635). https://doi.org/10.1029/2022GL098635

Gordon, E. M., Barnes, E. A., & Hurrell, J. W. (2021). Oceanic Harbingers of Pacific Decadal Oscillation Predictability in CESM2 Detected by Neural Networks. *Geophysical Research Letters*, *48*, e2021GL095392. https://doi.org/10.1029/2021GL095392

Gordon, E. M., Barnes, E. A., & Davenport, F. V. (2023). Separating internal and forced contributions to near term SST predictability in the CESM2-LE. *Environmental Research Letters*, *18*(10), 104047. https://doi.org/10.1088/1748-9326/acfdbc

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., et al. (2020). Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks. *Geoscientific Model Development*, *13*(5), 2197–2244. https://doi.org/10.5194/gmd-13-2197-2020

Ham, Y. G., Kim, J. H., & Luo, J. J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572. https://doi.org/10.1038/s41586-019-1559-7

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., et al. (2017). Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. *Journal of Climate*, *30*(20), 8179–8205. https://doi.org/10.1175/JCLI-D-16-0836.1

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., et al. (2020). GISS-E2.1: Configurations and Climatology. *Journal of Advances in Modeling Earth Systems*, *12*(8). https://doi.org/10.1029/2019MS002025

Kushnir, Y., Scaife, A. A., Arritt, R., Balsamo, G., Boer, G., Doblas-Reyes, F., et al. (2019). Towards operational predictions of the near-term climate. *Nature Climate Change*, *9*(2), 94–101. https://doi.org/10.1038/s41558-018-0359-7

Labe, Z. M., & Barnes, E. A. (2021). Detecting Climate Signals Using Explainable AI With Single-Forcing Large Ensembles. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2021MS002464. https://doi.org/10.1029/2021MS002464

Labe, Z. M., & Barnes, E. A. (2022). Predicting Slowdowns in Decadal Climate Warming Trends With Explainable Neural Networks. *Geophysical Research Letters*, *49*(9). https://doi.org/10.1029/2022GL098173

Martin, E. R., & Thorncroft, C. (2014). Sahel rainfall in multimodel CMIP5 decadal hindcasts. *Geophysical Research Letters*, *41*(6), 2169–2175. https://doi.org/10.1002/2014GL059338

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019). Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO_2. *Journal of Advances in Modeling Earth Systems*, *11*(4), 998–1038. https://doi.org/10.1029/2018MS001400

Mayer, K. J., & Barnes, E. A. (2021). Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network. *Geophysical Research Letters*, *48*(10), e2020GL092092. https://doi.org/10.1029/2020GL092092

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., et al. (2009). Decadal Prediction: Can It Be Skillful? *Bulletin of the American Meteorological Society*, *90*(10), 1467–1486. https://doi.org/10.1175/2009BAMS2778.1

Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F., et al. (2021). Initialized Earth System prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment*, *2*(5), 340–357. https://doi.org/10.1038/s43017-021-00155-x

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. https://doi.org/10.1007/s00382-022-06272-7

Qin, M., Du, Z., Hu, L., Cao, W., Fu, Z., Qin, L., et al. (2022). Deep Learning for Multi-Timescales Pacific Decadal Oscillation Forecasting. *Geophysical Research Letters*, *49*(6). https://doi.org/10.1029/2021GL096479

Risbey, J. S., Squire, D. T., Black, A. S., DelSole, T., Lepore, C., Matear, R. J., et al. (2021). Standard assessments of climate forecast skill can be misleading. *Nature Communications*, *12*(1), 4346. https://doi.org/10.1038/s41467-021-23771-z

Scaife, A. A., & Smith, D. (2018). A signal-to-noise paradox in climate science. *Npj Climate and Atmospheric Science*, *1*(1), 28. https://doi.org/10.1038/s41612-018-0038-4

Simpson, I. R., Yeager, S. G., McKinnon, K. A., & Deser, C. (2019). Decadal predictability of late winter precipitation in western Europe through an ocean–jet stream connection. *Nature Geoscience*, *12*(8), 613–619. https://doi.org/10.1038/s41561-019-0391-x

Smith, D. M., Eade, R., Dunstone, N. J., Fereday, D., Murphy, J. M., Pohlmann, H., & Scaife, A. A. (2010). Skilful multi-year predictions of Atlantic hurricane frequency. *Nature Geoscience*, *3*(12), 846–849. https://doi.org/10.1038/ngeo1004

Sutton, R. T., & Hodson, D. L. R. (2005). Atlantic Ocean Forcing of North American and European Summer Climate. *Science*, *309*(5731), 115–118. https://doi.org/10.1126/science.1109496

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019). The Canadian Earth System Model version 5 (CanESM5.0.3). *Geoscientific Model Development*, *12*(11), 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., et al. (2019). Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, *12*(7), 2727–2765. https://doi.org/10.5194/gmd-12-2727-2019

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. https://doi.org/10.1029/2019MS002002

Toms, B. A., Barnes, E. A., & Hurrell, J. W. (2021). Assessing Decadal Predictability in an Earth-System Model Using Explainable Neural Networks. *Geophysical Research Letters*, *48*(12), e2021GL093842. https://doi.org/10.1029/2021GL093842

Van Oldenborgh, G. J., Doblas-Reyes, F. J., Wouters, B., & Hazeleger, W. (2012). Decadal prediction skill in a multi-model ensemble. *Climate Dynamics*, *38*(7–8), 1263–1280. https://doi.org/10.1007/s00382-012-1313-4

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., et al. (2019). Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2177–2213. https://doi.org/10.1029/2019MS001683

Yeager, S. G., & Robson, J. I. (2017). Recent Progress in Understanding and Predicting Atlantic Decadal Climate Variability. *Current Climate Change Reports*, *3*(2), 112–127. https://doi.org/10.1007/s40641-017-0064-z

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., et al. (2018). Predicting Near-Term Changes in the Earth System: A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model. *Bulletin of the American Meteorological Society*, *99*(9), 1867–1886. https://doi.org/10.1175/BAMS-D-17-0098.1

Zhang, L., Delworth, T. L., Yang, X., Morioka, Y., Zeng, F., & Lu, F. (2023). Skillful decadal prediction skill over the Southern Ocean based on GFDL SPEAR Model-Analogs. *Environmental Research Communications*, *5*(2), 021002. https://doi.org/10.1088/2515-7620/acb90e

Ziehn, T., Chamberlain, M. A., Law, R. M., Lenton, A., Bodman, R. W., Dix, M., et al. (2020). The Australian Earth System Model: ACCESS-ESM1.5. *Journal of Southern Hemisphere Earth Systems Science*, *70*(1), 193. https://doi.org/10.1071/ES19035