# Assessing Decadal Predictability in an Earth-System Model Using Explainable Neural Networks

Benjamin A Toms[1], Elizabeth A. Barnes[1], and James Wilson Hurrell[1]

[1]Colorado State University

November 22, 2022

## Abstract

We show that explainable neural networks can identify regions of oceanic variability that contribute predictability on decadal timescales in a fully coupled Earth system model. The neural networks learn to use sea-surface temperature anomalies to predict future continental surface temperature anomalies. We then use a neural network explainability method called layerwise relevance propagation to infer which oceanic patterns lead to accurate predictions made by the neural networks. In particular, regions within the North Atlantic Ocean and North Pacific Ocean lend the most predictability for surface temperature across continental North America. We apply the proposed methodology to decadal variability, although the concept is generalizable to other timescales of predictability. Furthermore, while our approach focuses on predictable patterns of internal variability within climate models, it should be generalizable to observational data as well. Our study contributes to the growing evidence that interpretable neural networks are important tools for advancing geoscientific knowledge.

# Assessing Decadal Predictability in an Earth-System Model Using Explainable Neural Networks

**Benjamin A. Toms[1], Elizabeth A. Barnes[1], and James W. Hurrell[1]**

[1]Department of Atmospheric Science, Colorado State University

**Key Points:**

- Explainable neural networks can serve as a new tool for identifying patterns of Earth system predictability
- Oceanic patterns that lend predictability in CESM2 occur in similar locations to known oceanic modes
- The proposed method can be used to separate the timing and location of predictable patterns

Corresponding author: Elizabeth A. Barnes, eabarnes@rams.colostate.edu

## Abstract

We show that explainable neural networks can identify regions of oceanic variability that contribute predictability on decadal timescales in a fully coupled Earth system model. The neural networks learn to use sea-surface temperature anomalies to predict future continental surface temperature anomalies. We then use a neural network explainability method called layerwise relevance propagation to infer which oceanic patterns lead to accurate predictions made by the neural networks. In particular, regions within the North Atlantic Ocean and North Pacific Ocean lend the most predictability for surface temperature across continental North America.

We apply the proposed methodology to decadal variability, although the concept is generalizable to other timescales of predictability. Furthermore, while our approach focuses on predictable patterns of internal variability within climate models, it should be generalizable to observational data as well. Our study contributes to the growing evidence that interpretable neural networks are important tools for advancing geoscientific knowledge.

## Plain Language Summary

We use a form of artificial intelligence and machine learning called neural networks to identify patterns within the ocean that can help predict temperature over land. We focus in particular on surface temperatures averaged over multiple years, since a growing body of scientific evidence has suggested that such timescales can be predicted using information about the ocean. We find that several oceanic patterns are associated with surface temperatures across North America in a fully coupled Earth system model. From a broader perspective, this study contributes to the growing body of scientific evidence that artificial intelligence and neural networks can be used to advance geoscientific knowledge.

## 1 Introduction

Explainable neural networks have opened new doorways in Earth science research (Toms, Barnes, & Ebert-Uphoff, 2020), with applications ranging from the identification of climate change indicators (Barnes et al., 2020), hail detection within severe thunderstorms (Gagne II et al., 2019), and the improvement of numerical model parameterizations (Brenowitz et al., 2020), among other applications (Toms, Kashinath, et al., 2020). The specific usage of neural network interpretation techniques ranges substantially across such studies, however, as the interpretations can be used as either direct or indirect tools for scientific discovery. For example, interpretation efforts can be either a secondary objective by ensuring a network's reasoning is consistent with existing physical theory (e.g. Brenowitz et al., 2020; Ebert-Uphoff and Hilburn, 2020; Toms et al., 2020), or the primary objective, with their usage focused on discovering new patterns of Earth system variability (e.g. Toms, Barnes, and Ebert-Uphoff (2020); Barnes et al. (2020)). Here, we focus on the latter application, whereby we use neural networks to identify predictable modes of Earth system variability on decadal timescales in a fully coupled Earth system model.

An extensive body of literature exists on theoretical and observed sources of decadal predictability, and, more recently, on the development of operational decadal prediction systems (Yeager et al., 2018). Modes of regional and global-scale decadal variability within the ocean are well documented (e.g. Barnett et al., 1999; Kirtman and Schopf, 1998; Xie and Tanimoto, 1998), and these patterns have been found to contribute to atmospheric anomalies on decadal timescales via ocean-atmosphere feedbacks (e.g. Newman et al., 2016; Schneider et al., 2002; Wen et al., 2016). The discovery of this coupling has led to the usage of oceanic variability to make decadal predictions of atmospheric anomalies relevant to society. Recently, oceanic observations have been assimilated into Earth system models to generate large ensembles of global decadal predictions (Meehl et al., 2009; van Oldenborgh et al., 2012; Yeager et al., 2018), which have a reasonable amount of prediction skill for variables such as continental temperature and precipitation (Smith et al., 2019) and ocean acidification (Brady et al., 2020). Additional efforts have created statistical decadal prediction models based on knowledge of specific modes of oceanic decadal variability (e.g. Simpson et al., 2019).

There are, however, limitations to decadal predictions that use dynamical Earth system models, including how to initialize the observational fields (He et al., 2017; Kröger et al., 2018) and long-standing model biases in simulating known ocean-atmosphere and land-atmosphere interactions (Black et al., 1999; Chang et al., 1997; Simpson et al., 2019). It is therefore not clear whether regions that lack predictability in decadal prediction ensembles have limited predictability in the observed world, or whether model limitations preclude accurate predictions. This uncertainty also exists for other timescales of Earth system prediction, such as subseasonal-to-seasonal timescales (Jin et al., 2008; Kim et al., 2018, 2019; Koster et al., 2011; Toms, Barnes, Maloney, & van den Heever, 2020). For statistical models, a complete knowledge of which patterns of oceanic variability offer predictability is important for the correct selection of model inputs and thereby a maximization of statistical prediction skill (e.g. DelSole and Banerjee, 2017; Simpson et al., 2019; Wilks, 2008).

Because of these uncertainties, it is useful to identify predictable patterns of Earth system variability within both models and observations. Knowledge of such patterns may, for example, help guide efforts to improve the robustness of observational assimilation within dynamical decadal prediction systems, or inform which variables and regions to include within statistical models. To this end, we use a new method, namely explainable neural networks, to identify sources of decadal predictability within a fully coupled Earth system model. We take a purely methodological approach and test whether the proposed method is viable for identifying such patterns of predictability, which opens opportunities for its application to a broader range of predictability problems in future studies.

## 2 Data and Methods

Our neural network architecture is designed to receive inputs of oceanic fields from an Earth system model and output the predicted sign of a continental temperature anomaly at a given location. Figure 1 describes this neural network design, and the appendix contains additional information about the training procedure. It is important to note that we have opted to keep the neural network as simple as possible to both maximize interpretability and to ensure our approach is valid before venturing into more complex networks in future studies. The neural network has one hidden layer of 32 nodes which is connected to two output nodes, both of which represent a different outcome associated with the input oceanic information. We use the rectified linear unit (ReLu; $max(0, x)$) activation function and apply a softmax operator to the output layer. The softmax operator transforms the neural network outputs into relative likelihoods of the two output climate states.

For our particular application, we input vectorized maps of global sea-surface temperature (SST) and the neural network is trained to output the associated likelihood that future continental surface temperatures across locations of North America will be anomalously warm or cold. The SST and continental surface temperature data are gathered from the Community Earth System Model Version 2 (CESM2; Danabasoglu et al., 2020) pre-industrial control simulation of the Coupled Model Intercomparison Project, Phase 6 (CMIP6; Eyring et al., 2016). We remove the seasonal cycle from both fields and regrid the SST field onto a 4° by 4° grid to reduce the number of inputs into the neural network. This grid spacing still permits the resolution of dominant patterns of oceanic variability, as we will show in Section 3. We also linearly detrend both fields by separately subtracting the linear trend from each grid point to reduce impacts of model drift during the control simulations. The input to the neural networks is a sequence of lagged sea-surface temperature maps that are vectorized and concatenated into a single vector, and includes the most recent SST map along with the 3-month, 6-month, and 9-month time-lagged SST maps. We include the lagged SST information because we find that the neural networks converge on an accurate solution more accurately when we do so.

We also apply a 24-month running average to the SST anomalies and a 60-month running average to the continental surface temperature anomalies, such that for any time the corresponding SST field represents the precedent 24-month mean and the continental surface temperature represents the future 60-month mean. We use these input and output smoothing durations to demonstrate the utility of the proposed methodology, and they can be changed for particular timescales or seasons of interest. The CMIP6 CESM2 pre-industrial control simulation offers 1,200 years of monthly data, the first 900 of which we use to train the neural networks and the last 300 of which we use for validation. We omit the beginning and end of the time-series which are contaminated by the temporal smoothing. We note that because we train the neural networks using a pre-industrial control simulation, all estimates of predictability provided by the neural networks are for internal variability only and do not include information about any predictable response due to anthropogenic forcing.

After training the neural network, we use an interpretation method called layer-wise relevance propagation (LRP; Montavon et al., 2018) to assess what the network has learned. In brief, LRP traces the decision-making process of a neural network for each individual input sample. For each input sample, the network pathways through which information flows to arrive at the associated output is traced backwards and projected back onto the dimensions of the input. Computationally, LRP identifies which patterns within the input lead to increases in value for a particular output node. This projection enables an interpretation of which inputs are most important for making predictions on a case-by-case basis. Our usage of LRP therefore offers insights into which patterns of SST variability lend predictability of decadal surface temperature anomalies over continental North America within CESM2. A more detailed discussion of LRP and its applicability to Earth system research is discussed in Toms, Barnes, and Ebert-Uphoff (2020), and additional applications are available in Barnes et al. (2020), Ebert-Uphoff and Hilburn (2020), and Toms et al. (2020).

## 3 Assessment of Decadal Predictability

We train a separate neural network for each location on a 5° by 5° grid across the globe, and assess the accuracy using the validation data (the last 300 years of the CESM2 pre-industrial control simulation). We choose this resolution due to the computational expense of training a neural network for every location across the globe. Each neural network can then identify patterns of SST that lend predictability unique to each location, which is helpful for understanding if the predictability across different regions of the globe is sourced from different oceanic patterns. Figure 2 shows the resultant accuracy for each of these neural networks in predicting the 1-to-60 month average surface temperature using a global map of the prior 24-month mean SST within the CESM2 pre-industrial control simulation. The accuracy varies across the globe, with southern Africa, southern Australia, the Maritime Continent, and parts of northeastern North America exhibiting the highest accuracy. It is important to note that we choose the neural network parameters to ensure the accuracy on the training and validation datasets are similar, the details of which are provided in the appendix.

We then use LRP to assess which modes of oceanic variability contribute to the predictability within the CESM2 pre-industrial control simulation. The following analysis is applicable to any region of the globe, although we choose North America as an example. We only assess the LRP interpretations for cases when the neural networks make accurate predictions within both the training and validation datasets, although for future use-cases it is likely that assessing the LRP interpretations for inaccurate predictions will also be useful. We further separate the interpretations into accurate predictions of positive and negative temperature anomalies and only show the results for the positive anomalies, although the analysis for the negative anomalies is similar (see supplementary information). Also, while we input a sequence of lagged SST anomalies into the neural networks (as shown in Figure 1), the interpretations for each lag are nearly identical in spatial structure, but with the magnitude of LRP relevance decreasing with increasing lag (see supplementary information).

The composite LRP patterns for four regions across North America suggest that predictability is sourced from different oceanic patterns for different regions (Figure 3). Perhaps surprisingly, continental temperature anomalies within Central America are most associated with SST anomalies off the east coast of Japan (Figure 3a), likely within the Kuroshio Extension (Qiu & Chen, 2005). SST anomalies within the North-Central Pacific Ocean are associated with continental temperature anomalies along the west coast (Figure 3b), while those within the tropical Pacific Ocean contribute to predictability across central North America (Figure 3c). The North Atlantic Ocean contributes predictability to the four locations, although its impacts are particularly prominent across the northeast portions of the continent (Figure 3d). These patterns of predictability occur in similar regions to known modes of oceanic variability, such as the El Niño-Southern Oscillation (Kirtman & Schopf, 1998; Kleeman et al., 1999; Newman et al., 2003), the Pacific Decadal Oscillation (Mantua & Hare, 2002; Newman et al., 2016), and the Atlantic Meridional Overturning Circulation (Knight et al., 2005; Medhaug et al., 2012). A mechanistic study is needed before it can be said whether the identified patterns within CESM2 are associated with any of these three observed modes of oceanic variability, although the regional similarities lend confidence that this may be the case.

A unique aspect of our approach is that LRP highlights which input patterns contribute to predictability on a case-by-case basis. So, we further analyze which patterns of oceanic variability lend continental temperature predictability by using k-means clustering. The composite interpretation in Figure 3 risks averaging together temporally distinct patterns of predictability, and so the clustering approach allows us to analyze these potentially distinct patterns separately. We focus in particular on the west coast of North America in a region that exhibits high continental surface temperature predictability (according to Figure 2). We determine the optimal number of clusters by plotting the number of clusters against the mean Euclidian distance between each cluster, and selecting the number of clusters which falls in the inflection point of this curve (not shown). The inflection point denotes the number of clusters after which the addition of new clusters offers substantially less new information than the previous clusters. This technique is colloquially called the “elbow” technique (e.g. Dimitriadou et al. (2002)).

Using this approach, we find three dominant patterns of oceanic variability within CESM2 that lend predictability at the chosen location along the west coast of North America (Figure 4). These patterns are located in regions also impacted by known modes of oceanic decadal variability. The first mode occurs in a region commonly associated with the Kuroshio Extension (Qiu & Chen, 2005), while the second and third clusters occur in similar regions to the Atlantic Meridional Overturning Circulation (Knight et al., 2005, 2006) and Pacific Decadal Oscillation (Newman et al., 2016), respectively (Figure 4a, b, c). A mechanistic study is needed to tie the patterns identified within CESM2 to the aforementioned known modes of variability, although our analysis at least suggests that decadal predictability within CESM2 can be sourced independently from spatially distinct patterns of oceanic variability. The clustering analysis identifies the most spatially distinct patterns of variability, so it is likely that there are also situations where the identified patterns of variability lend predictability in tandem.

It is worth a quick note that the one-point correlation map of the 24-month smoothed SSTs and the surface temperature at the red dot in Figure 4 highlights most of the globe as correlated with the surface temperature at the west coast location (Supp. Figure 4). The neural network, however, identifies very localized regions as the best predictors, although some of these locations align with hot spots also seen in the one-point correlation map, e.g. the eastern Pacific and the North Atlantic.

Along with the predictions, the neural networks output likelihoods that the input SST field will lead to positive or negative continental temperature anomalies. We therefore use these likelihoods to assess the oceanic state for highly confident (i.e. high likelihood) accurate predictions, and compare those cases to accurate predictions with lower confidence. In doing so, we find that higher confidence predictions for the west coast of North America are made when non-lagged SST anomalies are of greater magnitude within the northern Atlantic and Pacific oceans (Figure 5). Anomalies within the North Pacific Ocean and North Atlantic Ocean are most magnified in the high confidence predictions. According to LRP, the non-lagged SST anomalies within the North Pacific Ocean are particularly relevant for the high confidence scenarios. The interpretations are spatially similar for the lagged SST fields, but with decreased amplitude of differences in SST and LRP values between the high and low confidence predictions (not shown).

## 4 Discussion

We demonstrate that neural networks can identify patterns of oceanic variability that lend predictability on decadal timescales within Earth system models. In particular, the neural networks identify known patterns of decadal oceanic variability as sources of predictability for continental surface temperature anomalies across North America within the CMIP6 CESM2 pre-industrial control simulation. The identified patterns of oceanic variability each offer distinct sources of predictability, at least across the west coast of North America where the useful oceanic regimes occur in regions also impacted by known modes of decadal oceanic variability such as the Atlantic Meridional Overturning Circulation, Pacific Decadal Oscillation, and Kuroshio Extension. A mechanistic study is needed to assess whether the patterns identified within CESM2 are truly associated with these known modes, or if they simply occur in a similar location.

We propose the methodology in this paper through its application to a single Earth system model (CESM2), although the method can be applied to a collection of climate models to assess the similarities of predictable climate modes across different models. Additionally, while we applied the proposed methods to decadal prediction, the methods are also likely viable for other timescales. Subseasonal-to-seasonal prediction may particularly benefit from such an approach, as these timescales lie at the intersection of predictable processes in the atmosphere, land, and ocean (Koster et al., 2011; Kumar & Hoerling, 1998; Woolnough et al., 2007). Explainable neural networks may therefore be useful in determining coincident patterns of predictability within each domain.

The complexity of the proposed method can be varied as necessary, although we introduce it here with intentional simplicity. For example, the neural networks can be made more nonlinear through the addition of more nodes and hidden layers, temporal information can be included within the inputs and outputs, and numerous Earth-system variables can be input rather than sea-surface temperature alone. The method may also be applicable to observational data, particularly cases for which an extensive observational record exists (e.g. subseasonal-to-seasonal prediction). Our formulation also only tasks the neural network with predicting positive or negative temperature anomalies without regard to magnitude, so the addition of more categories of output temperature anomalies can help separate anomalies of different magnitudes. From a broader perspective, this study contributes to the growing body of evidence that interpretable neural networks can be used to advance geoscientific knowledge.

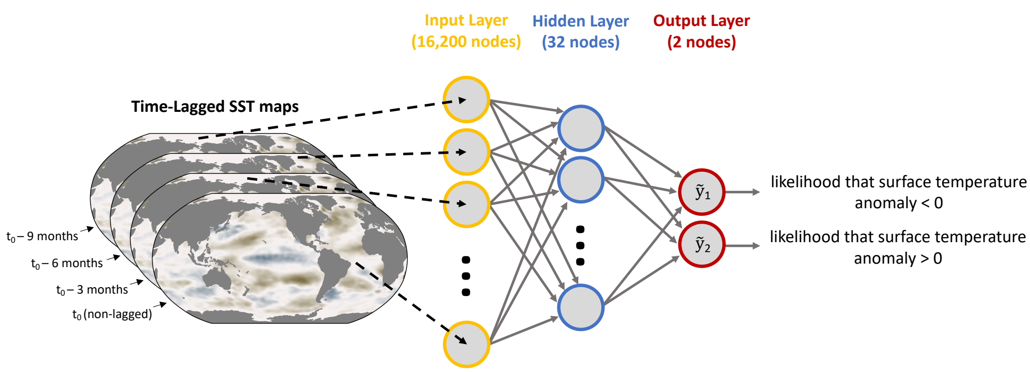



**Figure 1.** Schematic of the neural network design. The neural network receives a concatenated sequence of vectorized sea-surface temperature fields as input, passes the input forward to a single hidden layer of 32 nodes, and finally outputs a likelihood that the input is associated with surface temperature anomalies of a particular sign for a specified location. Note that the input samples include four sea-surface temperature maps that are vectorized and concatenated before being input into the neural network. The input includes the most recent SST map and the time-lagged 3-month, 6-month, and 9-month SST maps.

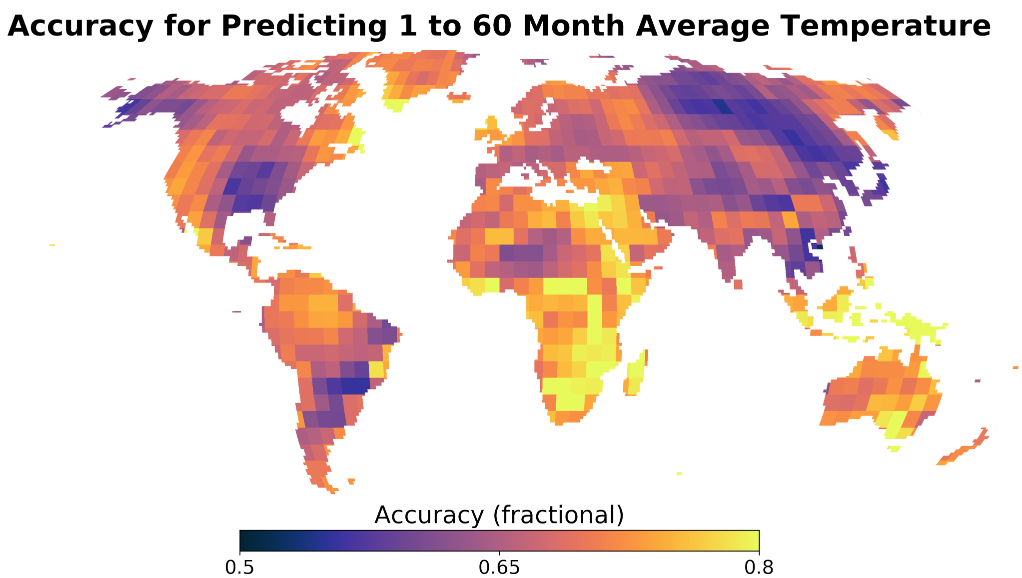



**Figure 2.** Accuracy for the neural network approach using only the validation data (the last 300 years of the CESM2 pre-industrial control simulation). The accuracy is defined in a Boolean sense, and the output node with the highest likelihood is taken as the networks' prediction. The accuracy values therefore represent the fraction of predictions for which the neural networks predict the correct sign of continental surface temperature anomalies. The values shown are the average of five different neural network trained for each location, as discussed within the appendix.

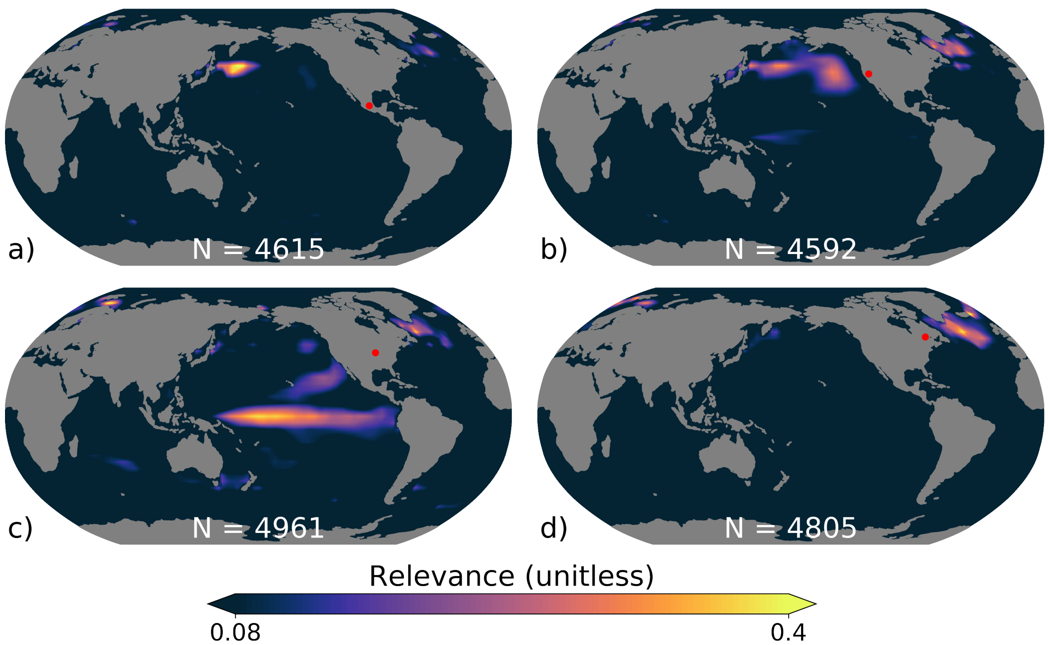



**Figure 3.** Composite (i.e. simple average) of layerwise relevance propagation interpretations for the non-lagged SST field for accurate predictions of positive surface temperature anomalies at four locations across North America. The continental locations associated with the composites are denoted by the red dots in each panel. The LRP interpretation for each sample is normalized between a value of 0 and 1 before compositing to ensure each prediction carries the same weight in the composite. The number of samples used in each composite ($N$) is shown within each sub-figure. Relevance values below the 95th percentile confidence bounds (0.08) are not shown. Confidence bounds were determined using a null hypothesis of no predictability by randomly shuffling the order of the input sea-surface temperature maps, and calculating the 95th percentile values of the associated LRP composites. An example of LRP heatmaps for the lagged SST fields is provided in the supplementary information.

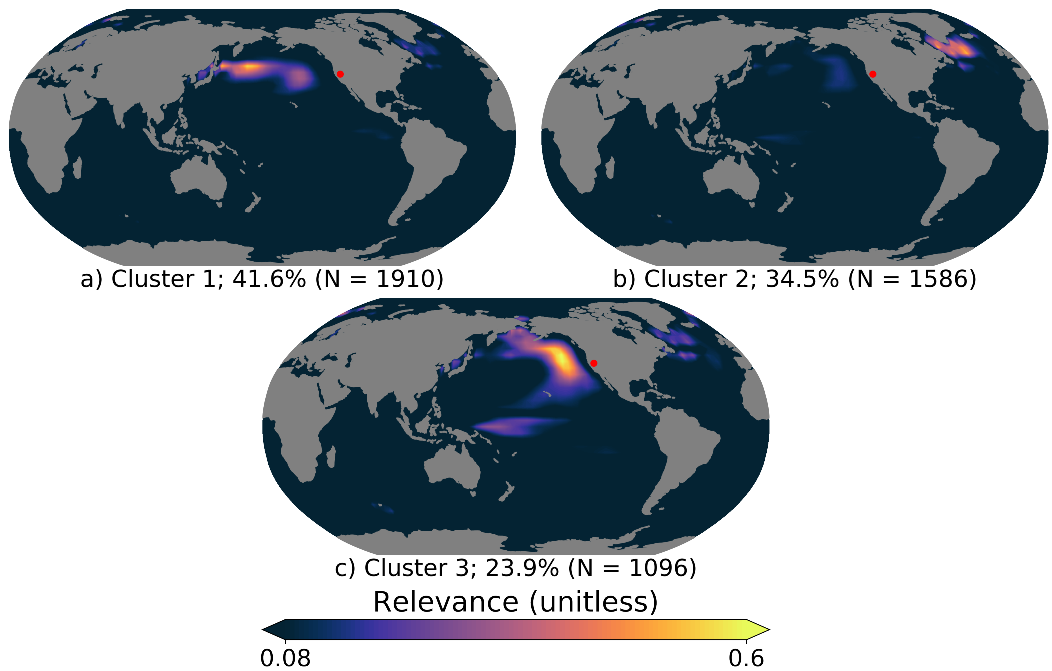



**Figure 4.** K-means clusters of the layerwise relevance propagation interpretations for the non-lagged SST field for accurate predictions of positive surface temperature anomalies at the red dot. The percentage of cases corresponding to each cluster is listed in the bottom left of each sub-panel and sum to 100%. The LRP values for each sample are normalized between a value of 0 and 1 before compositing to ensure each prediction carries the same weight in the composite. The number of samples used in each composite ($N$) is also shown. Relevance values below the 95th percentile confidence bounds (0.08) are not shown. Confidence bounds were determined using a null hypothesis of no predictability by randomly shuffling the order of the input sea-surface temperature maps, and calculating the 95th percentile values of the associated LRP composites.

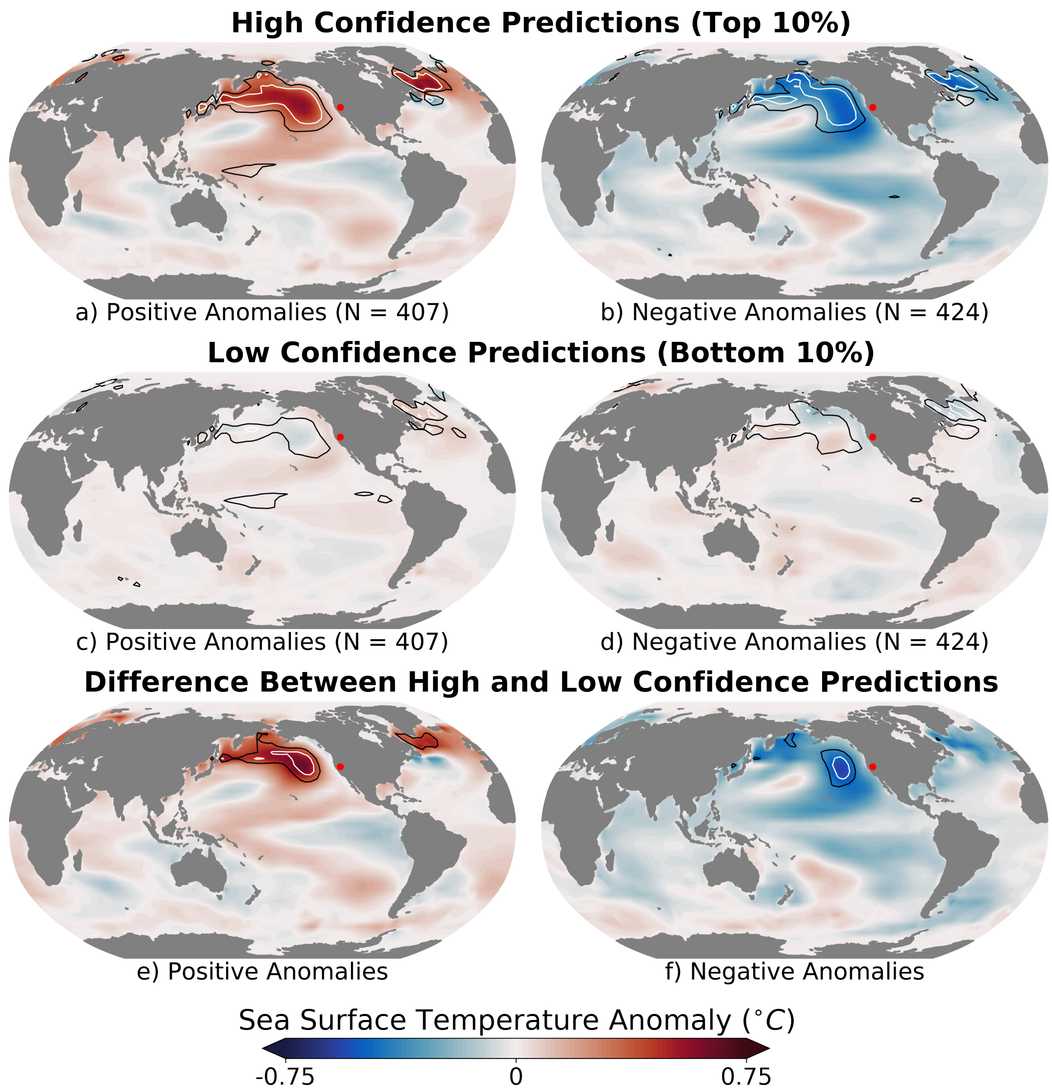



**Figure 5.** Differences in sea-surface temperature anomalies and LRP relevance for the 10% highest and 10% lowest confidence correct predictions for (a, c, e) positive surface temperature anomalies and (b, d, f) negative surface temperature anomalies at the red dot. The non-lagged sea-surface temperature anomalies are shown in fill, and LRP is shown in open contours. For subpanels a, b, c, and d, the black (white) contour denotes an LRP value of 0.3 (0.6). For subpanels e and f, the black (white) contour denotes an LRP difference of +0.1 (+0.2). Negative LRP relevance differences are also allowed to be shown, although none exist with magnitudes of -0.1 or greater.

# Appendix A Neural Network Details

This section includes details of how the neural networks were trained. Each neural network was trained using the Adam optimizer, with an initial learning rate of 1E-4. We do not change the learning rate throughout training. The single hidden layer of neurons is regularized with an L2 (ridge) regularization coefficient of 10, which ensures the neural network uses information from broader spatial regions and can not overfit to individual locations. This regularization parameter also ensures the accuracy for the training and validation datasets are similar. The networks were allowed to train for 100 epochs, which was sufficient for convergence in all cases. The model iteration that resulted in the highest accuracy on the validation data was selected and used for analysis. We train five neural networks for each location because it is possible that each network will find a different optimal solution, and so training numerous networks increases the likelihood that we capture the full range of optimal solutions. The accuracy values presented in Figure 2 represent the mean accuracy from the five networks. The interpretations presented in Figures 3, 4, and 5 are similar across each of the five network iterations, and so we randomly select one of the five neural networks and use this network for these analyses. We find that the networks converge on similar optimal solutions based on the LRP interpretations, and so training five models is sufficient for our purposes.

## Acknowledgments

Data from the CMIP6 CESM2 pre-industrial control simulation can be found on various CMIP6 archives, one of which is the Lawrence Livermore National Laboratory node of the Earth System Grid Federation domain: https://esgf-node.llnl.gov/projects/cmip6/. Benjamin A. Toms was supported by the Department of Energy Computational Science Graduate Fellowship via grant DE-FG02-97ER25308. Elizabeth A. Barnes was supported, in part, by NSF CAREER AGS-1749261 under the Climate and Large-scale Dynamics program.

# References

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, e2020MS002195. doi: 10.1029/2020MS002195

Barnett, T. P., Pierce, D. W., Saravanan, R., Schneider, N., Dommenget, D., & Latif, M. (1999). Origins of the midlatitude pacific decadal variability. *Geophysical Research Letters*, *26*(10), 1453–1456.

Black, D. E., Peterson, L. C., Overpeck, J. T., Kaplan, A., Evans, M. N., & Kashgarian, M. (1999). Eight centuries of north atlantic ocean atmosphere variability. *Science*, *286*(5445), 1709–1713.

Brady, R. X., Lovenduski, N. S., Yeager, S. G., Long, M. C., & Lindsay, K. (2020). Skillful multiyear predictions of ocean acidification in the california current system. *Nature Communications*, *11*(1), 1–9.

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *arXiv preprint arXiv:2003.06549*.

Chang, P., Ji, L., & Li, H. (1997). A decadal climate variation in the tropical atlantic ocean from thermodynamic air-sea interactions. *Nature*, *385*(6616), 516–518.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D., DuVivier, A., Edwards, J., … others (2020). The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001916.

DelSole, T., & Banerjee, A. (2017). Statistical seasonal prediction based on regularized regression. *Journal of Climate*, *30*(4), 1345–1361.

Dimitriadou, E., Dolničar, S., & Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, *67*(1), 137–159.

Ebert-Uphoff, I., & Hilburn, K. A. (2020). Evaluation, tuning and interpretation of neural networks for meteorological applications. *arXiv preprint arXiv:2005.03126*.

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.

Gagne II, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, *147*(8), 2827–2845.

He, Y., Wang, B., Liu, M., Liu, L., Yu, Y., Liu, J., … others (2017). Reduction of initial shock in decadal predictions using a new initialization strategy. *Geophysical Research Letters*, *44*(16), 8538–8547.

Jin, E. K., Kinter, J. L., Wang, B., Park, C.-K., Kang, I.-S., Kirtman, B., … others (2008). Current status of enso prediction skill in coupled ocean–atmosphere models. *Climate Dynamics*, *31*(6), 647–664.

Kim, H., Janiga, M. A., & Pegion, K. (2019). Mjo propagation processes and mean biases in the subx and s2s reforecasts. *Journal of Geophysical Research: Atmospheres*, *124*(16), 9314–9331.

Kim, H., Vitart, F., & Waliser, D. E. (2018). Prediction of the madden–julian oscillation: A review. *Journal of Climate*, *31*(23), 9425–9443.

Kirtman, B. P., & Schopf, P. S. (1998). Decadal variability in enso predictability and prediction. *Journal of Climate*, *11*(11), 2804–2822.

Kleeman, R., McCreary Jr, J. P., & Klinger, B. A. (1999). A mechanism for generating enso decadal variability. *Geophysical Research Letters*, *26*(12), 1743–1746.

Knight, J. R., Allan, R. J., Folland, C. K., Vellinga, M., & Mann, M. E. (2005). A signature of persistent natural thermohaline circulation cycles in observed climate. *Geophysical Research Letters*, *32*(20).

Knight, J. R., Folland, C. K., & Scaife, A. A. (2006). Climate impacts of the atlantic multidecadal oscillation. *Geophysical Research Letters*, *33*(17).

Koster, R., Mahanama, S., Yamada, T., Balsamo, G., Berg, A., Boisserie, M., … others (2011). The second phase of the global land–atmosphere coupling experiment: soil moisture contributions to subseasonal forecast skill. *Journal of Hydrometeorology*, *12*(5), 805–822.

Kröger, J., Pohlmann, H., Sienz, F., Marotzke, J., Baehr, J., Köhl, A., … others (2018). Full-field initialized decadal predictions with the mpi earth system model: An initial shock in the north atlantic. *Climate Dynamics*, *51*(7-8), 2593–2608.

Kumar, A., & Hoerling, M. P. (1998, 12). Annual Cycle of PacificNorth American Seasonal Predictability Associated with Different Phases of ENSO. *Journal of Climate*, *11*(12), 3295-3308. doi: 10.1175/1520-0442(1998)011⟨3295:ACOPNA⟩2.0.CO;2

Mantua, N. J., & Hare, S. R. (2002). The pacific decadal oscillation. *Journal of oceanography*, *58*(1), 35–44.

Medhaug, I., Langehaug, H. R., Eldevik, T., Furevik, T., & Bentsen, M. (2012). Mechanisms for decadal scale variability in a simulated atlantic meridional overturning circulation. *Climate dynamics*, *39*(1-2), 77–93.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., … others (2009). Decadal prediction: Can it be skillful? *Bulletin of the American Meteorological Society*, *90*(10), 1467–1486.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

Newman, M., Alexander, M. A., Ault, T. R., Cobb, K. M., Deser, C., Di Lorenzo, E., ... others (2016). The pacific decadal oscillation, revisited. *Journal of Climate*, *29*(12), 4399–4427.

Newman, M., Compo, G. P., & Alexander, M. A. (2003). Enso-forced variability of the pacific decadal oscillation. *Journal of Climate*, *16*(23), 3853–3857.

Qiu, B., & Chen, S. (2005). Variability of the kuroshio extension jet, recirculation gyre, and mesoscale eddies on decadal time scales. *Journal of Physical Oceanography*, *35*(11), 2090–2103.

Schneider, N., Miller, A. J., & Pierce, D. W. (2002). Anatomy of north pacific decadal variability. *Journal of climate*, *15*(6), 586–605.

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.

Smith, D., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T., ... others (2019). Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science*, *2*(1), 1–10.

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. doi: 10.1029/2019MS002002

Toms, B. A., Barnes, E. A., Maloney, E. D., & van den Heever, S. C. (2020). The global teleconnection signature of the madden-julian oscillation and its modulation by the quasi-biennial oscillation. *Journal of Geophysical Research: Atmospheres*, *125*(7), e2020JD032653.

Toms, B. A., Kashinath, K., Prabhat, & Yang, D. (2020). Testing the reliability of interpretable neural networks in geoscience using the madden-julian oscillation. *Geoscientific Model Development Discussions*, *2020*, 1–22. doi: 10.5194/gmd-2020-152

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.

Wen, N., Frankignoul, C., & Gastineau, G. (2016). Active amoc–nao coupling in the ipsl-cm5a-mr climate model. *Climate Dynamics*, *47*(7-8), 2105–2119.

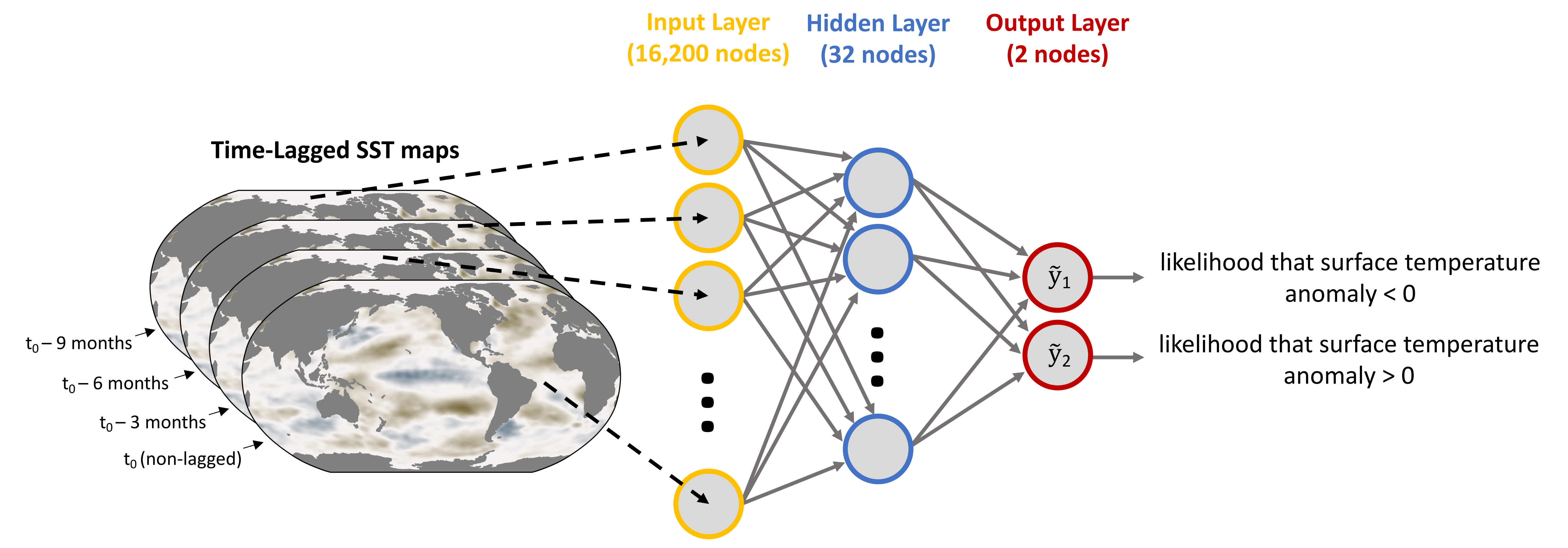
Wilks, D. S. (2008). Improved statistical seasonal forecasts using extended training data. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *28*(12), 1589–1598.

Woolnough, S. J., Vitart, F., & Balmaseda, M. A. (2007). The role of the ocean in the maddenjulian oscillation: Implications for mjo prediction. *Quarterly Journal of the Royal Meteorological Society*, *133*(622), 117-128. doi: 10.1002/qj.4

Xie, S.-P., & Tanimoto, Y. (1998). A pan-atlantic decadal climate oscillation. *Geophysical Research Letters*, *25*(12), 2185–2188.

Yeager, S., Danabasoglu, G., Rosenbloom, N., Strand, W., Bates, S., Meehl, G., ... others (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.

**Figure1.**



Input Layer
(16,200 nodes)
Hidden Layer
(32 nodes)
Output Layer
(2 nodes)
Time-Lagged SST maps
$t_0$ – 9 months
$t_0$ – 6 months
$t_0$ – 3 months
$t_0$ (non-lagged)
$\tilde{y}_1$
$\tilde{y}_2$
likelihood that surface temperature anomaly < 0
likelihood that surface temperature anomaly > 0

**Figure2.**

# Accuracy for Predicting 1 to 60 Month Average Temperature

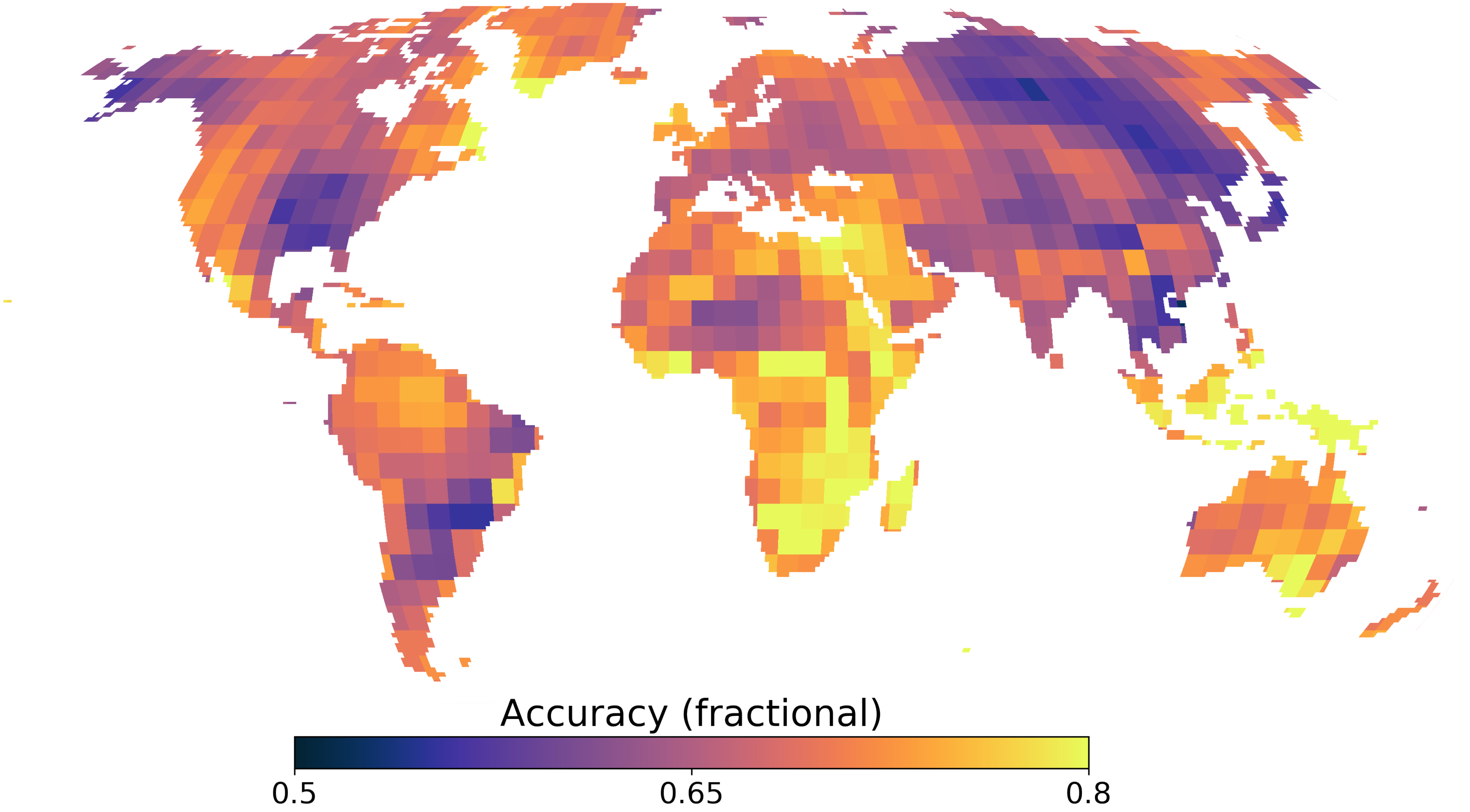
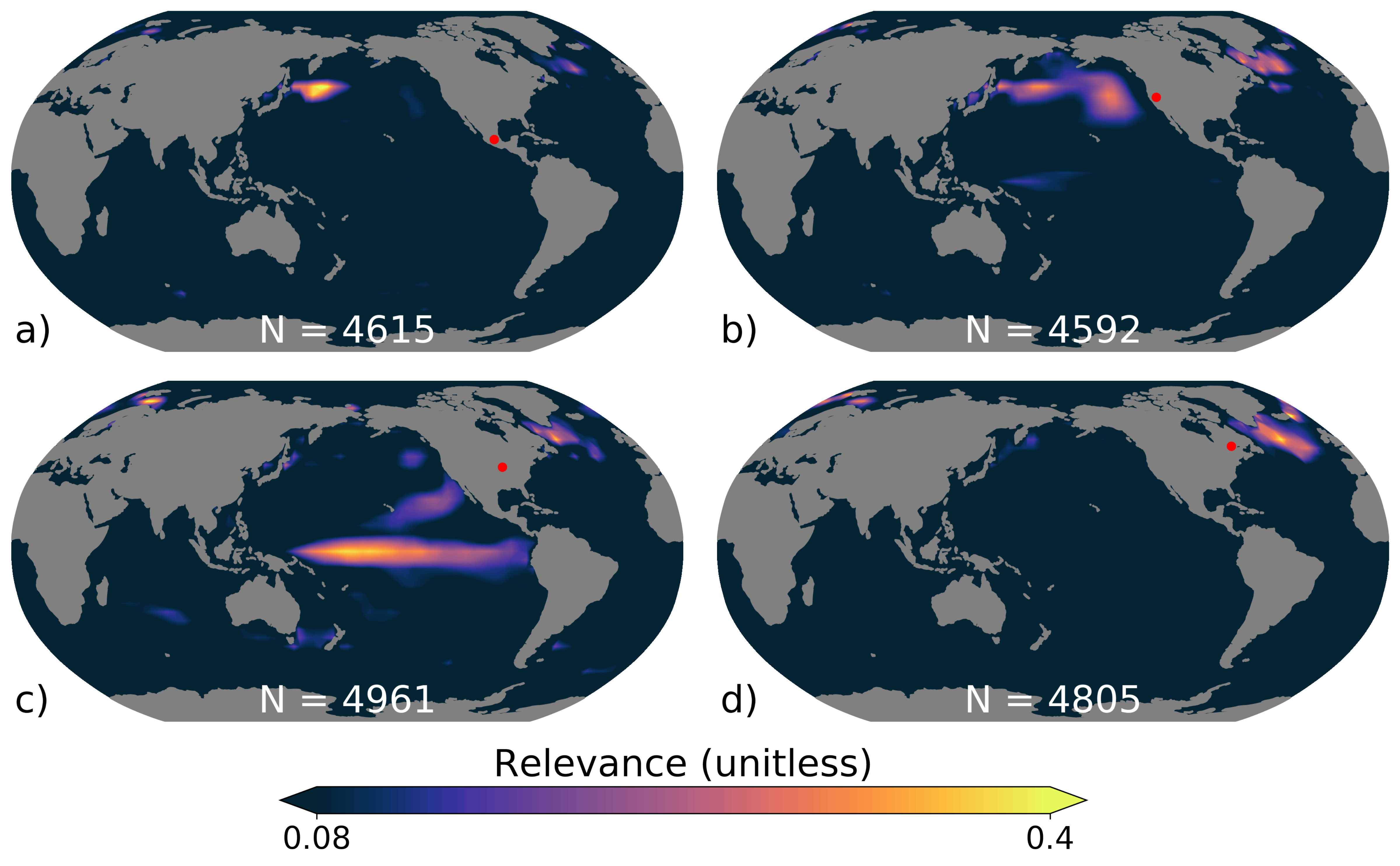


**Figure3.**



a)
N = 4615
b)
N = 4592
c)
N = 4961
d)
N = 4805
Relevance (unitless)
0.08
0.4

**Figure4.**

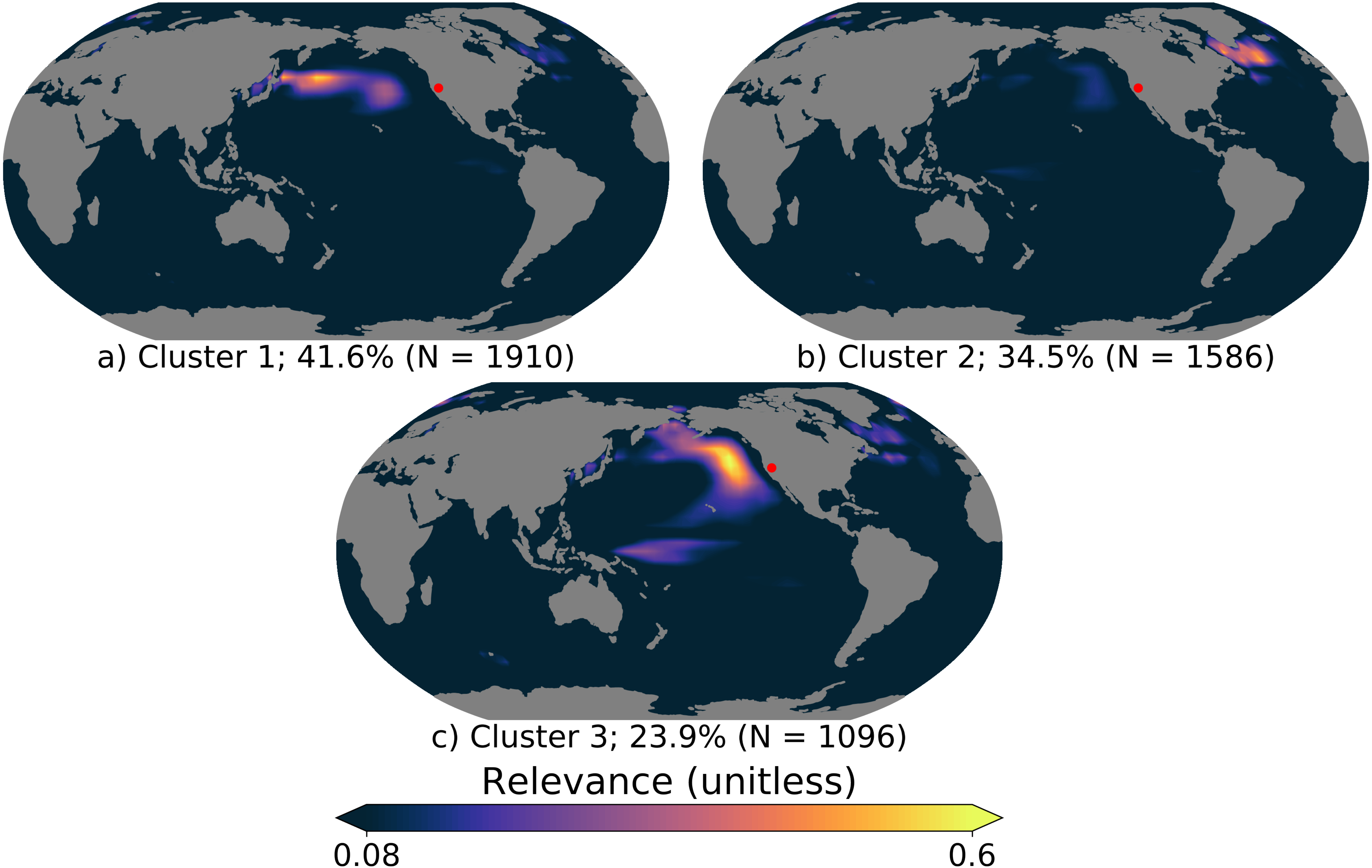
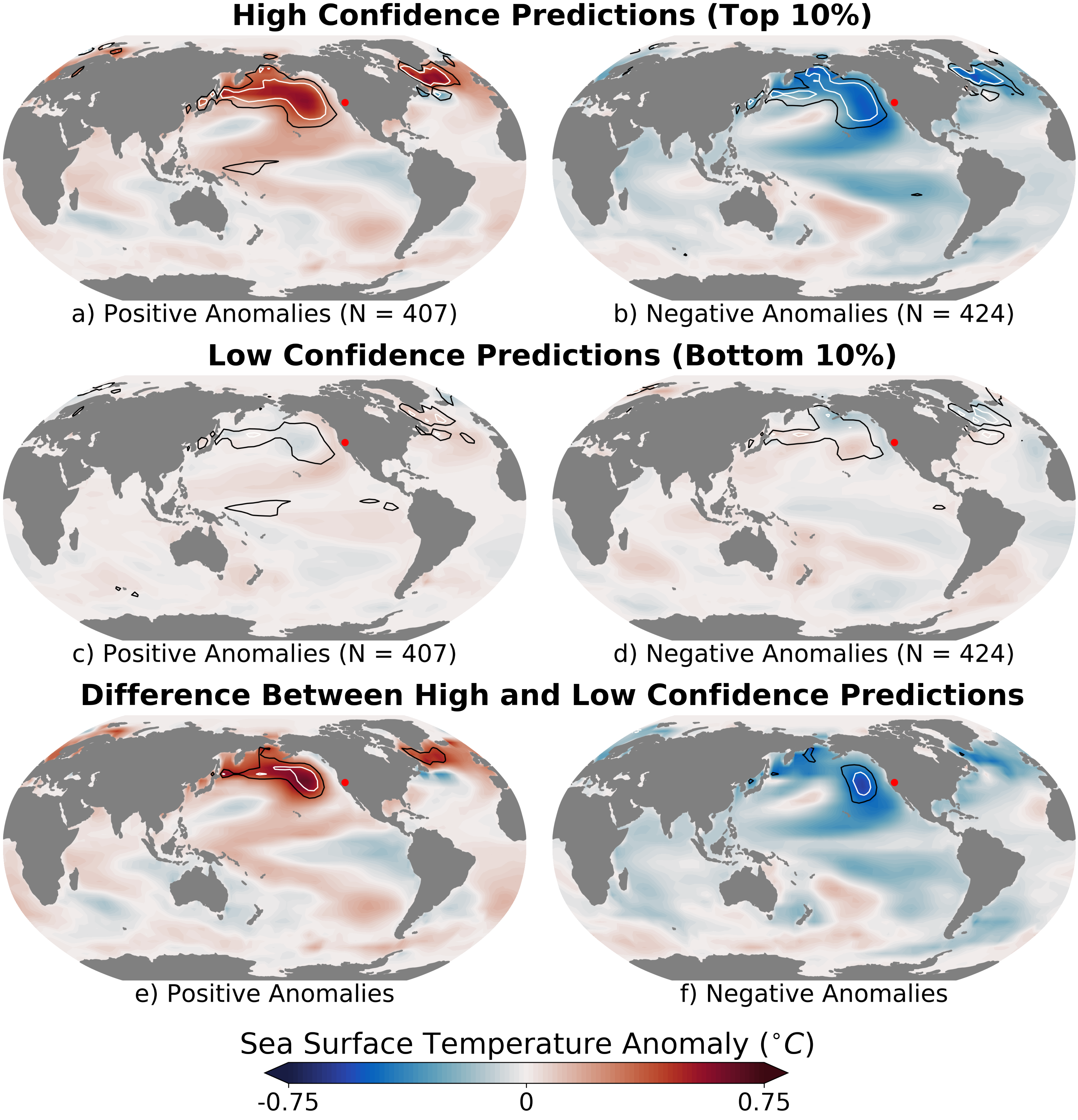

a) Cluster 1; 41.6% (N = 1910)
b) Cluster 2; 34.5% (N = 1586)
c) Cluster 3; 23.9% (N = 1096)
Relevance (unitless)
0.08
0.6

**Figure5.**


High Confidence Predictions (Top 10%)
a) Positive Anomalies (N = 407)
b) Negative Anomalies (N = 424)
Low Confidence Predictions (Bottom 10%)
c) Positive Anomalies (N = 407)
d) Negative Anomalies (N = 424)
Difference Between High and Low Confidence Predictions
e) Positive Anomalies
f) Negative Anomalies
Sea Surface Temperature Anomaly (°C)
-0.75
0
0.75