Comparison of climate model large ensembles with observations in the Arctic using simple neural networks

Zachary M. Labe[1,1] and Elizabeth A. Barnes[1,1]

[1]Colorado State University

November 30, 2022

Abstract

Evaluating historical simulations from global climate models (GCMs) remains an important exercise for better understanding future projections of climate change and variability in rapidly warming regions, such as the Arctic. As an alternative approach for comparing climate models and observations, we set up a machine learning classification task using a shallow artificial neural network (ANN). Specifically, we train an ANN on maps of annual mean near-surface temperature in the Arctic from a multi-model large ensemble archive in order to classify which GCM produced each temperature map. After training our ANN on data from the large ensembles, we input annual mean maps of Arctic temperature from observational reanalysis and sort the prediction output according to increasing values of the ANN's confidence for each GCM class. To attempt to understand how the ANN is classifying each temperature map with a GCM, we leverage a feature attribution method from explainable artificial intelligence. By comparing composites from the attribution method for every GCM classification, we find that the ANN is learning regional temperature patterns in the Arctic that are unique to each GCM relative to the multi-model mean ensemble. In agreement with recent studies, we show that ANNs can be useful tools for extracting regional climate signals in GCMs and observations.

Comparison of climate model large ensembles with observations in the Arctic using simple neural networks

Zachary M. Labe[1] **and Elizabeth A. Barnes**[1]

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

Key Points:

- Artificial neural network is trained to identify which climate model produced an annual mean map of near-surface temperature in the Arctic
- The classification network is evaluated using input from atmospheric reanalysis as a method of comparing climate models and observations
- An explainability method reveals regional temperature patterns the network is using to classify observations with different climate models

Corresponding author: Zachary M. Labe, zmlabe@rams.colostate.edu

Abstract

Evaluating historical simulations from global climate models (GCMs) remains an important exercise for better understanding future projections of climate change and variability in rapidly warming regions, such as the Arctic. As an alternative approach for comparing climate models and observations, we set up a machine learning classification task using a shallow artificial neural network (ANN). Specifically, we train an ANN on maps of annual mean near-surface temperature in the Arctic from a multi-model large ensemble archive in order to classify which GCM produced each temperature map. After training our ANN on data from the large ensembles, we input annual mean maps of Arctic temperature from observational reanalysis and sort the prediction output according to increasing values of the ANN's confidence for each GCM class. To attempt to understand how the ANN is classifying each temperature map with a GCM, we leverage a feature attribution method from explainable artificial intelligence. By comparing composites from the attribution method for every GCM classification, we find that the ANN is learning regional temperature patterns in the Arctic that are unique to each GCM relative to the multi-model mean ensemble. In agreement with recent studies, we show that ANNs can be useful tools for extracting regional climate signals in GCMs and observations.

Plain Language Summary

Due to many complex processes in the climate system, the Arctic is warming more rapidly relative to other parts of the globe. To understand the impacts of these changes in the Arctic, it is important to evaluate climate model projections. While there are other existing statistical methods for assessing simulations between different climate models, we introduce a machine learning approach for comparing climate models and observations using a tool called artificial neural networks. We set up our problem by inputting yearly maps of temperature in the Arctic and then task the artificial neural network to classify which climate model produced each map. To understand how the artificial neural network learns where the temperature map is coming from, we utilize a visualization method to peer into the machine learning black box. After training our artificial neural network on data from different climate models, we then input maps of Arctic temperature from observations to evaluate which climate model is classified for every year in the historical record. Using this setup, we find that the artificial neural network is lever-

aging regional patterns of temperatures, and not just overall warm and cold biases, in order to make its climate model and observation predictions.

1 Introduction

The Arctic is warming at a rate of more than three times as fast as the globally averaged mean surface temperature trend (Druckenmiller et al., 2021). This dramatic warming, otherwise known as Arctic amplification, is accompanied by long-term losses of Arctic sea-ice extent and thickness (Schweiger et al., 2019; Parkinson & DiGirolamo, 2021; Kacimi & Kwok, 2022), reductions in the ice mass of glaciers and the Greenland Ice Sheet (Mouginot et al., 2019; Tepes et al., 2021), thawing permafrost and boreal wildfires (McCarty et al., 2021; Miner et al., 2022), changes to deep ocean heat content and biogeochemistry (Timmermans et al., 2018; Solomon et al., 2021), shifts in high latitude phenology (Myers-Smith et al., 2020), and other possible connections to local and remote extreme weather (Graham et al., 2017; Cohen et al., 2020). As summarized in Previdi et al. (2021) and P. C. Taylor et al. (2022), local CO_2 forcing and other positive feedbacks in the Earth system contribute to Arctic amplification, such as from increases in atmosphere-ocean poleward energy transport, changes in clouds and water vapor, the ice-albedo feedback, Planck and lapse rate feedbacks, and other radiative energy imbalances. To further understand the contributions to Arctic amplification and its far-reaching impacts, it is necessary to evaluate climate models of varying orders of complexities (Dutta et al., 2021; Henry et al., 2021; Holland & Landrum, 2021; Hahn et al., 2022). Moreover, fully-coupled atmosphere-ocean global climate models (GCMs) are needed for comparing future assessments of Arctic climate change. However, there are large mean state biases across the Arctic between different GCMs (Davy & Outten, 2020), such as in Coupled Model Intercomparison Project 5 and 6 (CMIP5/6). For example, most CMIP6 models are still too cold over sea ice during the boreal winter (Davy & Outten, 2020). Large internal variability also needs to be accounted for in the high latitudes, especially when considering dynamical changes to the atmospheric circulation (Swart et al., 2015; M. England et al., 2019; Peings et al., 2021). To address some of these issues, one opportunity is to use large ensembles from different GCMs, which includes both internal variability and structural model uncertainties when comparing historical and future Arctic climate change simulations (Deser et al., 2020; Landrum & Holland, 2020).

Improving credibility, understanding, and trust in climate models requires constant evaluation of historical and future projections, especially for considering them in adaptation and mitigation planning in the Arctic. In fact, previous assessment reports from the Intergovernmental Panel on Climate Change have devoted entire chapters to climate model evaluation for summarizing GCM performance and other associated diagnostics (e.g., Randall et al., 2007; Flato et al., 2013). A number of scientific institutions have also developed automated statistical toolboxes, such as the Program for Climate Model Diagnosis and Intercomparison Metrics Package (Gleckler et al., 2016; Lee et al., 2021) and the National Center for Atmospheric Research Climate Variability Diagnostics Package (Phillips et al., 2014, 2020), to assist in methodologically comparing GCMs. These types of software packages usually compare sets of relative skills metrics or rankings for CMIP5/6 models across different mean climate fields, modes of internal variability, trends, extreme events, and teleconnections.

At a basic level, climate model evaluation considers sets of skill metrics, such as measures of bias, variance, pattern correlation, and root-mean-square error (RMSE), for comparing differences between GCMs and observations. The scalar metrics are often then presented in summary displays, such as through Taylor diagrams (K. E. Taylor, 2001) or portrait diagrams of relative error (Gleckler et al., 2008). In recent years, more advanced statistical methods have also been applied to mean climate benchmarks, such as through bias correction, emergent constraints, and model independence and performance-based weighting schemes (Knutti et al., 2017; Eyring et al., 2019; Brunner et al., 2020; Lauer et al., 2020; Merrifield et al., 2020). This includes leveraging output from newly designed GCM large ensembles (Maher et al., 2021). However, these common relative error and emergent constraint measures are not without issues (Chai & Draxler, 2014; Sanderson et al., 2021); in some cases, they may even underestimate the skill of climate models (Willmott et al., 2017). Most of these benchmarks also only consider point-by-point statistics, rather than considering potential (non)linear patterns across space or time. As a result, it is worth exploring new approaches for climate model evaluation, especially considering the growing interest in applying deep learning methods in the geosciences (Reichstein et al., 2019; Nowack et al., 2020; Nichol, Peterson, Fricke, & Peterson, 2021).

Although the use of machine learning methods is still fairly new in climate science applications (Rasu et al., 2019; Boukabara et al., 2021), several studies have already demon-

strated their utility over traditional multiple linear regression for identifying mechanistic processes and extracting patterns of climate change and variability (e.g., Pasini et al., 2017; Barnes et al., 2020; Nichol, Peterson, Peterson, et al., 2021). In this study, we use a form of deep learning called artificial neural networks (ANNs) for classifying Arctic maps of temperature data according to different GCMs. We also leverage an explainable machine learning method to identify regional climate patterns that the ANN is using to make its classification.

Overall, the ANN is quickly able to learn which climate model produces each annual mean map of near-surface temperature by using regional patterns that are unique to each large ensemble simulation, especially relative to the multi-model mean large ensemble. The machine learning explainability method then reveals these relevant regional pattern fingerprints of temperature for each climate model. One motivation for this work is that we are interested in applying inputs from observationally-derived maps to compare with GCMs using the ANN classification scheme and evaluate whether our method produces similar results relative to other climate model evaluation techniques. Here the methodological difference is that by using ANNs we can also consider potential regional nonlinear relationships across the entire Arctic map, rather than only computing point-by-point statistics. Notably, we find that although the ANN is using these regional patterns, the classification results for comparing with observations resemble other simple evaluation methods.

2 Data

2.1 Multi-model large ensemble archive

To train our ANN on climate model data, we use a collection of single model initial-condition large ensemble simulations from the multi-model large ensemble archive (MMLEA) (NCAR, 2020; Deser et al., 2020). The MMLEA consists of seven CMIP5-class GCMs, which range in ensemble size from 16 to 100 members. Specifically, we use the Canadian Earth System Model Large Ensemble (CanESM2; Kirchmeier-Young et al., 2017), Max Planck Institute Grand Ensemble (MPI; Maher et al., 2019), Commonwealth Scientific and Industrial Research Organisation Large Ensemble (CSIRO-MK3.6; Jeffrey et al., 2013), EC-Earth Consortium Large Ensemble (EC-Earth; Hazeleger et al., 2010), Geophysical Fluid Dynamics Laboratory Large Ensemble (GFDL-CM3; Sun et al., 2018),

Geophysical Fluid Dynamics Laboratory Earth System Model Large Ensemble (GFDL-ESM2M; Rodgers et al., 2015), and the Community Earth System Model Large Ensemble Community Project (LENS; Kay et al., 2015).

We include only the first 16 ensemble members from each simulation, since this is the minimum number of ensemble members available to equally weight all seven GCMs (i.e., EC-Earth includes 16 ensemble members) when training and testing our ANN. The GCMs also differ by their initialization protocol (Stainforth et al., 2007; Hawkins et al., 2016) and utilize micro perturbations (i.e., small roundoff error in the atmospheric initial conditions: EC-Earth, GFDL-CM3, LENS), macro perturbations (i.e., different coupled atmosphere-ocean states: MPI, CSIRO-MK3.6, GFDL-ESM2M), or a combination of these two methods (CanESM2). All of the simulations in the MMLEA use historical forcing until 2005 and Representative Concentration Pathway 8.5 (RCP8.5) forcing thereafter (Riahi et al., 2011; K. E. Taylor et al., 2012). Although RCP8.5 is described as an unrealistically high emissions scenario (e.g., Peters & Hausfather, 2020; Hausfather & Peters, 2020), we focus on data from the observational record (1950-2019) and discuss the broader conclusions of using explainable neural networks to compare maps of climate data between different GCMs. Given that the individual RCP scenarios do not substantially diverge until later in the 21st century (Vuuren et al., 2011), the use of this future emissions scenario does not affect the interpretation of our results.

Large ensembles are useful for disentangling the effects of internal variability relative to external climate forcing, especially in regions such as the Arctic. Recently, the MMLEA has been used in studies for evaluating Arctic amplification, (e.g., Landrum & Holland, 2020; M. R. England, 2021; Holland & Landrum, 2021), detection and attribution of extreme events in Siberia and Alaska (e.g., Ciavarella et al., 2021; Weidman et al., 2021), comparing projections of Arctic sea ice (e.g., Topál et al., 2020; Bonan et al., 2021), and identifying extratropical teleconnections (e.g., McKenna & Maycock, 2021; McCrystall & Screen, 2021). The high number of realizations per GCM is also particularly valuable for addressing deep learning and climate science applications, where large sample sizes are required for creating training datasets and improving overall ANN performance. As a recent example, Maher et al. (2022) leveraged the MMLEA and compared different supervised machine learning methods for classifying El Niño-Southern Oscillation events according to their spatial pattern.

In this work, we use monthly near-surface temperature (T2M) data and calculate annual means in each large ensemble simulation. To compare the results of our ANN with observations, we evaluate the 1950 to 2019 temporal period, which overlaps across all of the large ensembles and observations. Since the ANN requires the input maps to be the same size, all climate model data are regridded onto a common spatial grid of 1.9° latitude by 2.5° longitude using a bilinear interpolation scheme. A brief summary of the large ensemble simulations can be found in Table S1.

2.2 Atmospheric reanalysis

We primarily use ERA5 reanalysis to evaluate how the ANN would classify maps of T2M from observations after training the network on only the climate model large ensembles. ERA5 is the fifth generation of atmospheric reanalysis from the European Centre for Medium-Range Weather Forecasts (ECMWF) and provides hourly output on a 31 km horizontal grid with 137 vertical levels (up to 0.01 hPa) (Hersbach et al., 2020). ERA5 is based on the ECWMF's Integrated Forecast System (IFS) release 41r2 and uses four-dimensional variational analysis (4D-Var) as a data assimilation scheme. Output from ERA5 is available from 1979 to near real-time and is constrained by numerous satellite and in situ observations, such as from meteorological stations, ships, buoys, radiosonde profiles, and aircraft. To further extend the available observations back in time, we use the preliminary ERA5 back extension (BE), which is described in Bell et al. (2021).

In addition to being used as one of the primary datasets for monitoring Earth's global mean surface temperature (Dunn et al., 2021), ERA5 has been widely adopted for studies on Arctic climate change and variability (e.g., Davy & Outten, 2020; Cai et al., 2021; Nygård et al., 2021; R. Zhang et al., 2021). Detailed assessments of ERA5's representation of Arctic surface temperature can be found in Graham, Hudson, and Maturilli (2019); Wang et al. (2019); Yu et al. (2021), but in general, ERA5 suffers from a small warm bias over sea ice when compared to buoy observations and other in situ measurements. This bias may result from underestimating surface inversions and the simulation of turbulent and radiative heat flux exchanges, especially during the boreal winter (Graham, Cohen, et al., 2019).

Although ERA5 is a modeled product, its mean long-term trends and interannual variability of T2M compare well with other station-based datasets in the Arctic (Fig-

ure S1). However, in the Supporting Information, we also evaluate the ANN results using a separate observational dataset from the National Oceanic and Atmospheric Administration/Cooperative Institute for Research in Environmental Sciences/Department of Energy Twentieth Century Reanalysis (20CR) version 3 (20CRv3; Slivinski et al., 2019, 2021). The difference between the annual mean Arctic T2M for ERA5-BE and 20CRv3 is within 1°C in most years, and both datasets fall in the warmer envelope of the range in MMLEA mean climate states (Figure S4). However, there are notable regional differences in T2M across the Arctic between ERA5-BE and 20CRv3 (Figure S2-S3), especially in the vicinity of sea ice and Greenland. The implications of these differences for the ANN output will be further discussed in Sections 4-6. Overall, we focus on these atmospheric reanalysis products as they provide both temporarily and spatially-complete gridded data (i.e., no missing data) during our period of interest.

For comparison with the climate model results, we first bilinearly interpolate all reanalysis data onto the slightly coarser 1.9° latitude by 2.5° longitude grid. We then calculate annual mean maps of T2M from monthly output over the period of 1950 to 2019. A summary of the reanalysis data can be found in Table S2.

3 Methods

3.1 Artificial neural network architecture

In this work, we are interested in whether an ANN can correctly identify which climate model simulated an input map of Arctic T2M. As previously discussed, ANNs are useful in the geosciences for approximating nonlinear relationships in data-intensive problems (Boukabara et al., 2021; Irrgang et al., 2021). In climate science, this type of data problem often involves maps of climate variables that are available from large datasets, such as satellite data, gridded observational products, or climate models. If provided enough training data for the ANN to learn, and without it overfitting, the ANN can then make correct predictions on data it has not been seen before. Accompanying ANNs with explainability methods can also provide insights into the prediction by considering the trustworthiness of the ANN through scientific intuition for the specific application. An introduction to ANNs and other deep learning methods can be found in Lecun et al. (2015); Goodfellow et al. (2016); Neapolitan and Jiang (2018).

We use one shallow ANN architecture for the applications in this study, which is described in Figure 1. Our ANN receives flattened maps of Arctic T2M from climate model large ensembles in the MMLEA, where each unit of the input layer corresponds to a grid box on the map. Therefore, the input layer receives a total of 2016 units (i.e., 14 latitudes by 144 longitudes). This input vector is fed into two hidden layers with 10 nodes each. The output layer contains seven nodes, i.e., one for each GCM class. Our ANN is a fully-connected neural network, and the weights and biases are updated iteratively until the loss function is minimized. We use a categorical cross-entropy loss function, which acts to penalize larger model errors due to a logarithmic transformation. The rectified linear unit (ReLU; equation 1; Agarap, 2018) is used in the hidden layers for nonlinear transformation, and a softmax operator is included in the output layer (equation 2). We refer to the output of the ANN after the softmax operator as the ANN's 'confidence.' The softmax function remaps the output values of the ANN so that they sum to one for the class likelihoods of a given prediction. In other words, the GCM class that is ultimately selected receives the highest confidence in the ANN output. Overall, this architecture is very similar to recent studies using ANNs for evaluating patterns in climate models and observations (e.g., Barnes et al., 2019, 2020; Madakumbura et al., 2021), and this simple complexity is well suited for assessing the results of the explainability methods (Toms et al., 2020).

$$f(z_j) = max(0, z_j) \tag{1}$$

$$\tilde{y}_i = \frac{\exp(x_i)}{\sum_{j=1} \exp(x_j)} \tag{2}$$

Before the training process begins, we standardize each map of T2M by subtracting the training data mean across all 7 climate model large ensembles and dividing by the standard deviation of the training data over all 7 climate model large ensembles. This is computed at every grid point across all years (70 years; 1950 to 2019) and ensembles (12 members). We train the ANN using 12 ensemble members (75% of the data), validate on 1 ensemble member, and test on 3 ensemble members. This division of ensemble members is evenly applied across all 7 GCMs. To evaluate the performance of our network, we compute the accuracy classification score on testing data (equation 3).

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total Number of Predictions}} \tag{3}$$

In addition to the method of processing data using absolute T2M, we also conduct a set of experiments by first removing the mean temperature of each Arctic map before standardizing and allowing the training process to begin. In this set of results, the ANN cannot simply rely on differences in the overall mean state of each climate model large ensemble for making correct predictions. This method has also been successfully utilized in other previous studies for using ANNs to reveal regional indicator patterns of climate change (Barnes et al., 2020; Labe & Barnes, 2021).

As will be discussed later, this overall classification problem is simple for the ANN to learn (100% accuracy), and small changes to the proportions of splitting ensemble members do not affect our results. For training, we use a stochastic gradient descent optimizer (Ruder, 2016) with Nesterov momentum turned on (= 0.9) (Nesterov, 1983), a learning rate of 0.001, a batch size of 32, and we apply early stopping to set the number of epochs. Early stopping is a technique to help prevent overfitting. Here, the ANN is finished training if the validation loss does not decrease for 5 epochs in row. The ANN is then restored to the iteration with the best model weights, which is generally less than 200 epochs for our application. In addition to early stopping, we also apply ridge regularization (L_2; Friedman, 2012) to the first hidden layer in order to reduce overfitting. By limiting the sensitivity of the ANN to outlier weights, L_2 helps to reduce spatial autocorrelation that may exist in fields of climate data, such as T2M, and it is associated with smoother fields for interpreting our explainability maps. Our L_2 is set to 0.1, although we explore the results of testing observational data using different ridge parameters in Figures S14-S15.

3.2 Layer-wise relevance propagation

To evaluate how the ANN is classifying each temperature map with the correct GCM, we use a method of explainable machine learning called layer-wise relevance propagation (LRP; Bach et al., 2015; Montavon et al., 2017, 2018). First introduced by Toms et al. (2020) for applications in the geosciences, LRP has now been used in a wide range of studies across atmospheric and climate sciences for attempting to understand the decision-making process of neural networks (e.g., Hilburn et al., 2020; Gordon et al., 2021; Mayer

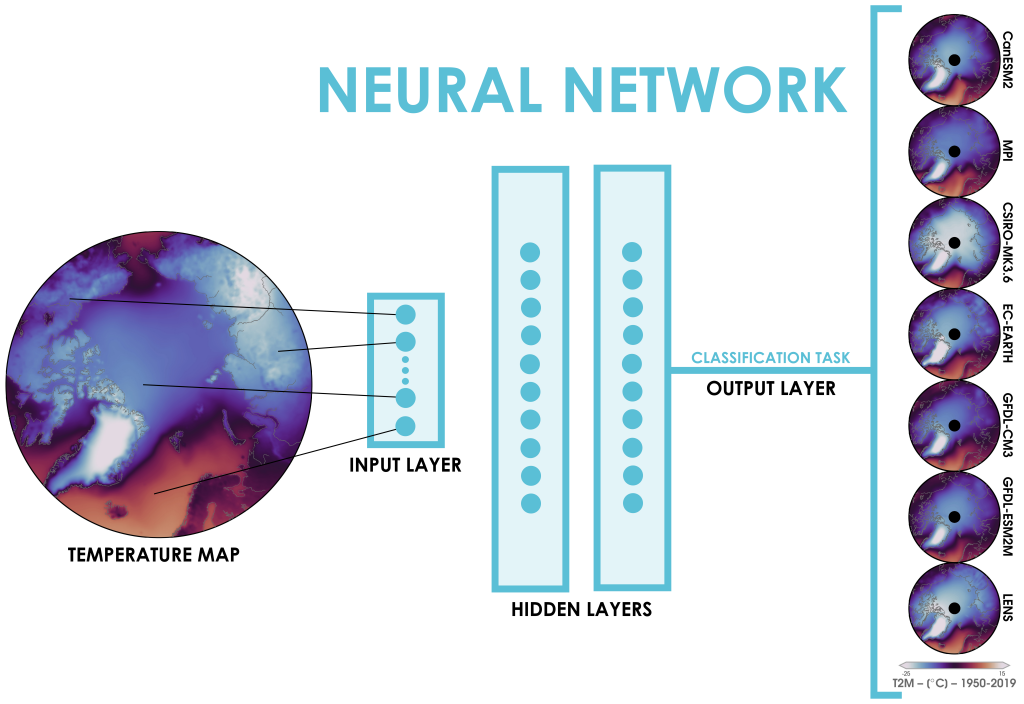

Figure 1. Schematic of the artificial neural network (ANN) used in this study for classifying which climate model large ensemble (output layer) produced a single map of Arctic near-surface temperature averaged over a given year (input layer). The ANN consists of two hidden layers that both contain 10 hidden nodes. The output layer includes a softmax activation function.

& Barnes, 2021; Sonnewald & Lguensat, 2021; Retsch et al., 2022). Importantly for its use in this work, LRP has also been shown to be an effective technique for extracting regional patterns of forced climate change that are collectively found between climate models and observations (e.g., Barnes et al., 2020; Labe & Barnes, 2021; Madakumbura et al., 2021; Rader et al., 2021). Despite a growing number of other machine learning explainability methods (e.g., Hedström et al., 2022), we find that LRP is well suited for the complexity of our simple neural network problem and geospatial input data.

LRP is a form of a posthoc feature attribution, where its output describes the contribution of each input pixel to the overall prediction of the neural network. In other words, LRP returns a heatmap that describes the relevance (unitless) of each input feature with the same dimensions. Specifically, in this study, LRP returns a vectorized heatmap of the relevance value at every latitude and longitude grid point across the Arctic (2016 units per map) for inputs of T2M. Thus, we can make individual composites of LRP heatmaps for every classification output in order to learn the patterns the ANN used to recognize each GCM.

Although overviews of LRP are described in numerous other studies (e.g., Montavon et al., 2019; Toms et al., 2020), we also summarize its implementation here to help improve clarity. After an ANN has been trained, the weights and biases are frozen, and a single input is passed through the network in forward mode to make a prediction. Next, prior to the softmax activation function, the winning output node (i.e., highest likelihood class) is backpropagated through the ANN using a set of decomposition rules. After propagating backward through the ANN to the input layer, we can then obtain relevance values for each input pixel. This entire process is repeated for each prediction, and therefore, we have a relevance heatmap for every annual mean temperature input.

We use the LRP_{z} method, but there are several other forms of LRP following different backpropagation rules (Bach et al., 2015; Samek et al., 2019) and available using the iNNvestigate package (Alber et al., 2019). In a recent comparison of LRP methods for geoscience applications, Mamalakis et al. (2021) demonstrated that LRP_{z} performed well compared to the ground truth using a benchmark dataset with similar characteristics to our climate model large ensemble data. We also compare our results using LRP_{z} with two other explainability methods, LRP_{ϵ} (Bach et al., 2015) and Integrated Gradients (Sundararajan et al., 2017), and find similar relevance spatial patterns (Figure S6).

Finally, while explainability techniques like LRP are useful for assessing whether a neural network is making predictions based on coherent and physically-based processes, we note that it is still subject to user interpretation. The LRP patterns here can only be used to identify the local temperature patterns unique to each GCM which are important for the ANN's decision-making process. However, we cannot directly assess how the ANN may be (non)linearly leveraging and weighting combinations of these regional temperature patterns together.

To improve visual clarity of our LRP output, we normalize each heatmap sample to have a maximum of one and then scale each figure composite by its maximum relevance. We elected to concentrate on positive relevance output for this analysis, which highlights areas that contribute positively to the final ANN classification. This also helps to simplify the interpretation of the explainability results for each of the climate model large ensemble considered here. In summary, locations of higher relevance indicate regions of temperature that are more important for the ANN to make its GCM classification.

4 Classifying climate model large ensembles

To begin exploring the differences between each GCM in the MMLEA, we first analyze their raw composites of annual mean T2M over the historical period in Figure S5. Unsurprisingly, all of the GCMs capture a similar spatial pattern of temperatures between sea-ice covered regions, open water in the North Atlantic and North Pacific, the Greenland Ice Sheet, and across other land areas. However, there are some notable differences in the mean T2M, especially for CSIRO-MK3.6, which is at least 3°C colder across most of the Arctic Ocean (Figure S5c). This is likely in association with an unrealistic sea ice mean state (i.e., higher sea-ice concentration) and slower rate of sea-ice decline over the last one to two decades (Uotila et al., 2013; Topál et al., 2020). It could also be due to biases in albedo, cloud processes, and other atmospheric dynamics, as decomposed for CESM1 by Park et al. (2014). Figure 2 shows that all of the GCMs capture higher interannual variability of T2M across the marginal ice zone in the North Atlantic, such in the Barents Sea region, with respect to ERA5-BE observations (Figure 2a). However, there is greater variability along and north of Siberia for CanESM2 (Figure 2b) and GFDL-CM3 (Figure 2f), which is likely again in response to differences in sea-ice variability. In summary, despite some differences in average T2M and spatial patterns of vari-

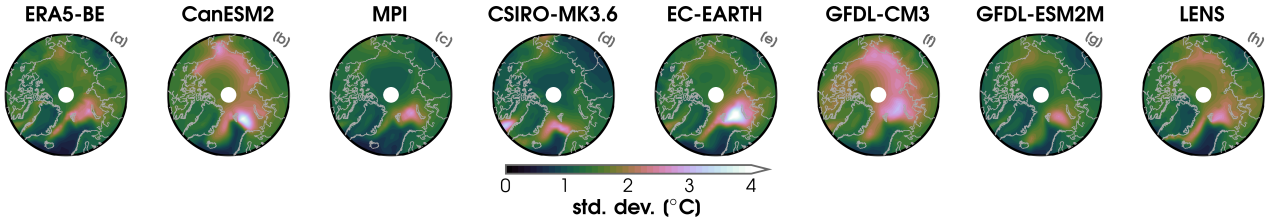

Figure 2. (a) Standard deviation of annual mean T2M (contour interval of 0.1°C) for ERA5-BE calculated over the 1950 to 2019 period. (b-h) Standard deviation of annual mean T2M for the mean of the ensemble members calculated over the 1950 to 2019 period for CanESM2, MPI, CSIRO-MK3.6, EC-EARTH, GFDL-CM3, GFDL-ESM2M, and LENS, respectively.

ability, all of the GCMs capture the general annual mean climatological characteristics of the Arctic.

We now turn to our ANN to see if it can correctly identify which GCM simulates every input map of annual mean T2M from 1950 to 2019. Recall that we train our ANN on 12 ensemble members from each GCM and then test the skill of the ANN using 3 ensemble members. The ANN is quickly able to learn how to identify each T2M map with the correct GCM and achieves a categorical accuracy of 100% on testing data. We hypothesize that this perfect accuracy is due to the easy task for the ANN, since the only differences between training, testing, and validation are due to the selected ensemble members. Thus, the systematic differences among the GCMs may be larger and spatially more persistent in both training and testing data than from that due to internal variability alone (i.e., only considering the differences between ensemble members for each GCM class). To further elucidate this point, Figure 3h-n shows the T2M differences for each GCM relative to the overall multi-model mean ensemble. This more clearly reveals the colder mean state in CSIRO-MK3.6 (Figure 3j), along with other regional differences among the other GCMs, especially across the North Atlantic, Greenland, and Canadian Arctic Archipelago.

We identify the regions that the ANN is leveraging to make its accurate predictions using the LRP explainability method in Figure 3a-g. The LRP heatmaps are composted separately for each GCM class across all testing ensemble members and years (1950-2019). Comparing the areas of higher relevance (i.e., locations that are more important for the ANN to make a prediction) in Figure 3a-g with the differences in T2M for each GCM

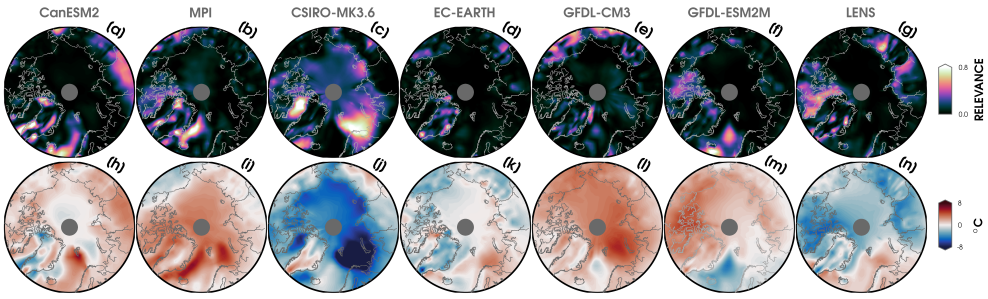

Figure 3. (a-g) Composite heatmap of layer-wise relevance propagation (LRP) for correct testing data predictions averaged over the 1950 to 2019 period for CanESM2, MPI, CSIRO-MK3.6, EC-EARTH, GFDL-CM3, GFDL-ESM2M, and LENS, respectively. (h) Composite of differences in T2M between CanESM2 minus the multi-model mean ensemble averaged over 1950 to 2019. (i-n) As in (h), but for MPI, CSIRO-MK3.6, EC-EARTH, GFDL-CM3, GFDL-ESM2M, and LENS, respectively.

minus the multi-model ensemble mean (Figure 3h-n) reveal clear similarities in the spatial patterns. This suggests that the ANN is learning characteristics of each GCM to make its classification. Importantly though, the relevance patterns indicate that the ANN is not simply using the entire map of T2M differences relative to the multi-model mean. For example, GFDL-CM3 is several degrees warmer than the multi-model ensemble mean in the Barents Sea region (Figure 3l). Yet, the LRP composite in Figure 3e suggests instead that T2M patterns in Alaska and the North Pacific are more relevant for the ANN to make a final prediction. In contrast, sometimes it is the case that the larger T2M differences correspond to areas of higher relevance, such as for LENS when comparing Figure 3g with the colder anomalies in Figure 3n over the Canadian Arctic Archipelago.

Overall, we interpret that the locations of higher relevance show that the ANN is spatially leveraging patterns of T2M that result in a unique set of characteristics or differences between each GCM class. To check that our interpretations of the LRP results are not sensitive to the choice of backpropagation rule, we compare relevance composites using the epsilon-rule (LRP_ϵ) and Integrated Gradients method in Figure S6. The relevance composites are nearly indistinguishable across the three explainability methods for all GCM classes. Given that the ANN is learning distinctive patterns of T2M to characterize each respective GCM, we now turn to observations to consider classifying each year with a GCM as a method of climate model evaluation.

5 Evaluating observations with climate model large ensembles

We first calculate the mean T2M bias for each GCM relative to ERA5-BE (Figure S7). All of the GCMs reveal a cold bias over the sea-ice covered portions of the Arctic Ocean, which has been a persistent issue for several generations of fully-coupled climate models (Chapman & Walsh, 2007; Davy & Outten, 2020). There are also other regional differences in T2M biases between GCMs, especially over Greenland and the Canadian Arctic Archipelago. To test the ANN on inputs from observations (Figure S8a), we first rescale each map by subtracting the training mean (Figure S8b,e) and dividing by the training standard deviation. In other words, the data is processed in the same method as the climate model large ensembles (Section 3.1). Figures S8c shows the difference from ERA5-BE minus the training mean, which again shows the Arctic Ocean cold bias in the climate model data. Although there are some small differences in magnitude, especially over Greenland, we find similar results for rescaling observations using 20CRv3 (Figure S8d-f). Composites of the rescaled T2M observations over three time periods display a persistent spatial pattern of T2M anomalies, except for the long-term background warming associated with Arctic amplification (Figure S9).

Finally, after rescaling the observational maps of annual mean T2M, we input them into the ANN to see which GCM is classified from 1950 to 2019. As discussed in Section 3.1, the ANN outputs the confidence (or likelihood) of a single T2M map belonging to each of the GCMs classes (Figure 4a). After applying the softmax operator, we sort these confidence values from lowest to highest and display these rankings in Figure 5 separately for every map year. Accordingly, the class with the highest confidence value is the GCM ultimately selected for each year and hence given a rank of '1.' If the confidence value is below that of random chance (1/7), the GCM is given a ranking of '7.' For ERA5-BE, we find that GFDL-CM3, EC-EARTH, and MPI are mostly frequently classified with the highest confidence in a single year. Interestingly, we also see a temporal evolution of these three models, with EC-EARTH more frequently classified in earlier years prior to 1979, MPI generally classified between 1979 to 2012, and GFDL-CM3 selected in the last few years. We hypothesize that this temporal evolution may be related to the long-term warming of the Arctic, which closely mirrors the Arctic mean T2M in Figure S1. GFDL-CM3 also observes the largest recent warming trends in the Arctic (not shown).

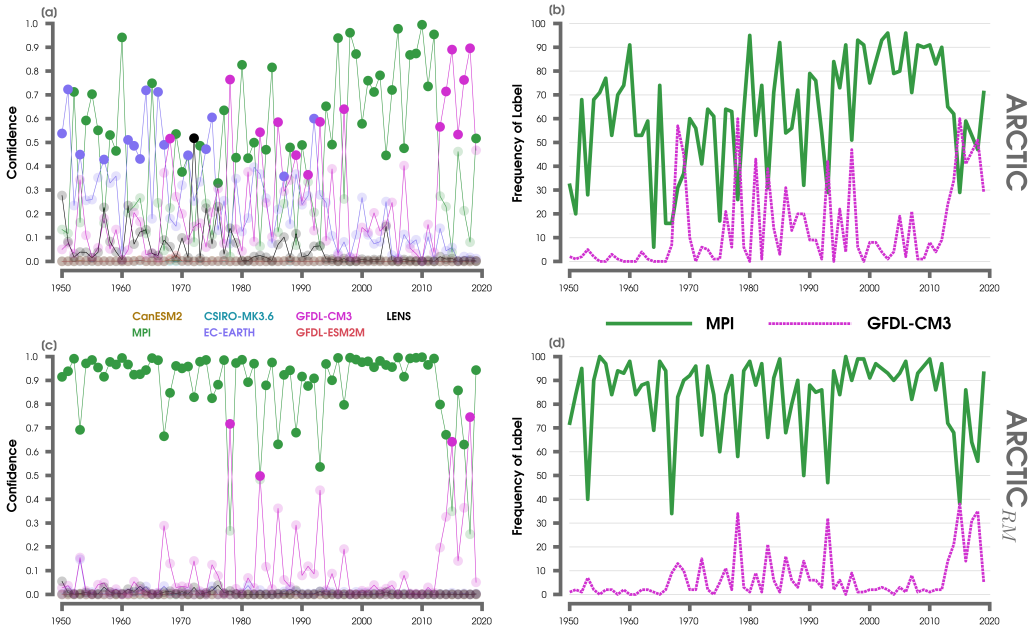

Figure 4. (a) Confidence values (after a softmax operator) from a single seed ANN for each GCM class after inputting an annual mean map of T2M from ERA5-BE over the period of 1950 to 2019. The line color and marker shading is darker for the GCM class with the highest confidence in each year. (b) Frequency of MPI (dark green line) and GFDL-CM3 (pink dashed line) classes for receiving the highest confidence prediction output for each annual mean T2M map from ERA5-BE. The frequency is considered by training 100 ANNs with different combinations of training, testing, and validation data and random initialization seeds. (c-d) As in (a-b) but after removing (RM) the annual mean of each T2M map from every grid point before inputting the observations into the ANN.

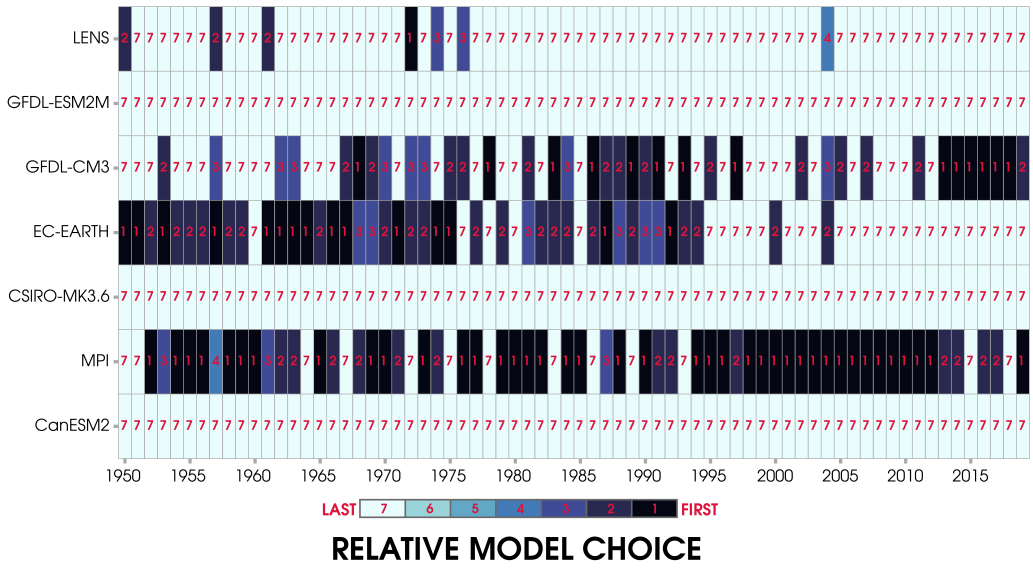

Figure 5. Ranking the order of the ANN confidence values (after a softmax operator) for each GCM class after inputting an annual mean map of T2M from ERA5-BE over the period of 1950 to 2019. A value of 1 indicates that the GCM received the highest confidence (i.e., winning predicted category) for each yearly T2M map. If the confidence value of the ANN output is lower than random chance (≈1/7), the ranking is then set to 7.

To test the robustness of these results, we train 100 separate ANNs using unique random initialization seeds and different combinations of training, testing, and validation data (ensemble members). After training each of these 100 ANNs, we then input the same T2M maps from ERA5-BE and show the frequency of classifying MPI and GFDL-CM3 in Figure 4. Similar to the single seed ANN predictions in Figure 4a, MPI is frequently predicted for the observational maps across the distribution of the 100 ANNs (Figure 4b). However, there are also small differences in the observational predictions, which suggests that there is some uncertainty due to the choice of training ensemble members and ANN initialization states.

As briefly mentioned in Section 3.1, to assess whether the network is simply just using a smaller mean state bias in a GCM for deciding to to make predictions for observations, we try training a new ANN experiment by first processing the climate model large ensembles to remove the annual mean T2M of the entire Arctic map from every grid point and for every year. In this case, by design, the ANN needs to learn regional patterns in order to make its classification. Here, the ANN once again quickly learns unique

spatial characteristics of each GCM and achieves a perfect accuracy for the testing data. We similarly evaluate the ERA5-BE maps by removing the annual mean T2M from each grid point and year (Figure S10). After sorting the confidence values of this new ANN (Figure 4c), we rank the GCMs in Figure S11 for every year of observations. In this case, we find that MPI receives the highest confidence in nearly every year. Again, testing the sensitivity of the observational predictions of this new ANN to the ensemble members selected for training, we compute 100 ANNs and show the frequency of MPI and GFDL-CM receiving the highest confidence in Figure 4d. Notably, processing the data with the annual map mean first removed results in MPI much more frequently labeled than the methodology used in Figure 4b. This suggests that the ANN is instead leveraging regional temperature patterns to more consistently make observational predictions of MPI.

Naturally, a next question is how closely do the ANN results compare with traditional relative error metrics for comparing climate models and observations. As a baseline comparison, we calculate the pattern correlation and RMSE between ERA5-BE and each GCM in Figure 6. The correlations and RMSEs are first computed between the observations and each ensemble member and then averaged together to get an ensemble mean. Most GCMs achieve a high pattern correlation (>0.9), which is unsurprising given the results in Figure S5. The lowest pattern correlation (and highest RMSE) is found for CSIRO-MK3.6, which is related to its cold bias and extensive sea ice mean state. Turning to RMSE, we find that MPI has the lowest error in most years of ERA5-BE. Notably, this is largely consistent with the ANN results in Figure 5. Finally, Figure S12 shows temporal correlations calculated at each grid point between ERA5-BE and the GCMs. Using this metric, GFDL-CM3 has the highest correlation over the Arctic Ocean, but most of the other GCMs have a similar spatial pattern too (>0.5) from the long-term warming trend.

We consider observations from 20CRv3 to assess how sensitive the GCM prediction results are to the choice of observational dataset. Following the same steps, Figure S13 shows the sorted ANN predictions of 20CRv3 maps according to increasing confidence values for each GCM class. MPI is frequently classified for each year of 20CRv3. However, in this exercise, we do not find any years with confidence above random chance for EC-EARTH. Although this differs from the results of ERA5-BE in Figure 5, this is not overly surprising given the mean state differences between the two observational datasets found in Figure S2-S4.

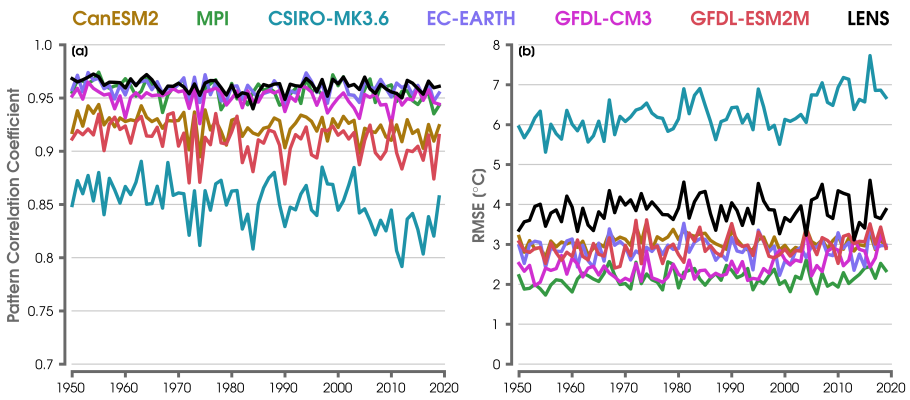

Figure 6. (a) Pattern correlation coefficient of T2M computed for each year between ERA5-BE and the climate model large ensembles from 1950 to 2019. Correlations (area weighted) are first calculated per each ensemble separately and then averaged across ensemble members. (b) Root-mean-square error (RMSE) of T2M for each year between ERA5-BE and the climate model large ensembles from 1950 to 2019. RMSEs (area weighted) are first calculated per each ensemble separately and then averaged across ensemble members.

Subsequently, it is evident that the spatial patterns of T2M are important for the ANN's prediction. This could be related to our choice of L_2 regularization, since a larger L_2 can effectively reduce spatial variability and irregularities in the input data. We test the effect of different L_2 parameters in Figure S14 on the observational predictions for ERA5-BE. Here, we find that a larger L_2 does in fact result in different GCM labels for the T2M maps, which could result from smoothing out the regional patterns that were originally important for the ANN using our L_2 choice of 0.1. Interestingly, we find that repeating this L_2 parameter exercise for ANNs with the annual map mean first removed results in more consistent predictions for observations (Figure S15). In summary, these findings further illustrate that the ANN is learning both information about the mean climate state and regional patterns that are associated with an individual GCM and observations. Moreover, the ANN is particularly sensitive to regional differences in T2M when classifying observations with a GCM.

6 Identifying regional climate patterns

So far, we've shown that an ANN can detect differences in regional T2M patterns that are unique to a particular GCM. We've also shown that observations can be eval-

uated in the ANN for identifying a GCM with each year in the historical record. This tends to result in observational predictions that are still fairly consistent with traditional climate model evaluation metrics like RMSE.

Now we can leverage our LRP explainability method by applying it to the observational data in order to more clearly see where the ANN is looking to make its predictions. Figure 7a-g shows the LRP results composited separately for each GCM that is ultimately classified from 1950 to 2019 (i.e., rankings of 1 in Figure 5). We compare these relevance heatmaps to T2M composites of the rescaled ERA5-BE input data in Figure 7h-n. Although at first glance the patterns of the rescaled T2M composites look fairly similar, it is clear that the ANN is using small regional differences to make its classification, as reflected by the relevance patterns in Figure 7a-g. For example, one year of observations is classified as LENS, which is likely due to the large cold anomaly over the Canadian Arctic Archipelago (Figure 7n) that is similarly reflected as an area of higher relevance in Figure 7g.

Due to some differences in the GCM predictions for 20CRv3 compared to ERA5-BE (Section 5), we show the LRP results for 20CRv3 testing predictions in Figure S16. For the composites of 20CRv3, we see higher relevance predominately over Greenland. Uncertainties in T2M are particularly large over Greenland for many reanalysis and other gridded observational datasets (e.g., Jack et al., 2017; Delhasse et al., 2020; W. Zhang et al., 2021), and this is found to be true for both ERA5-BE and 20CRv3 (Figure S2). Therefore, this may help to explain the differences found for the observational predictions in Section 5.

We can also explore the LRP results of the ANN experiment using T2M data with the annual mean of the Arctic first removed before training and testing. These relevance composites are shown in Figure S17. While MPI is selected for most observational years by this ANN (Figure S17b), we can still see spatial differences in the relevance regions compared with GFDL-CM3 (Figure S17e) particularly over northwestern Canada and eastern Siberia.

Finally, returning to the LRP results of the climate model large ensembles, Figure 8 shows the relevance heatmaps of each GCM from the ANN trained on data with the annual mean of the map first removed (RM). Comparing Figure 8 with the original LRP results of Figure 3a-g shows that the ANN is still using many of the same regional

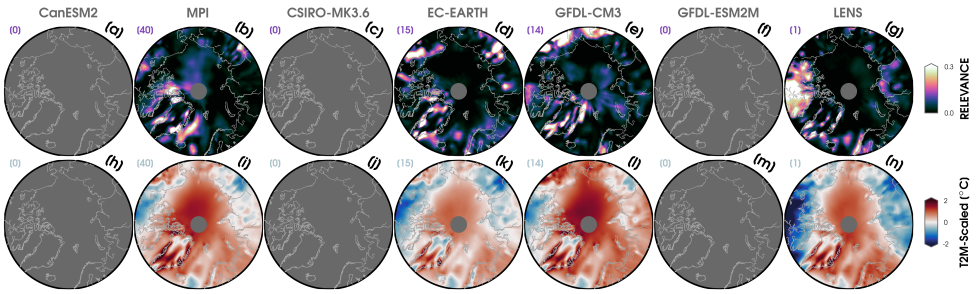

Figure 7. (a-g) Composites of LRP heatmaps for each GCM classification after inputting annual mean maps of T2M from ERA5-BE into the ANN. Higher values indicate greater relevance for the ANN's prediction. (h-n) Composites of T2M from ERA5-BE that are first scaled by the training data mean and training data standard deviation. Maps are then composited according to each predicted GCM class for every year. Maps that are gray indicate that the GCM was never classified, and the number in the upper left-hand corner indicates the number of times the GCM was classified from 1950 to 2019.

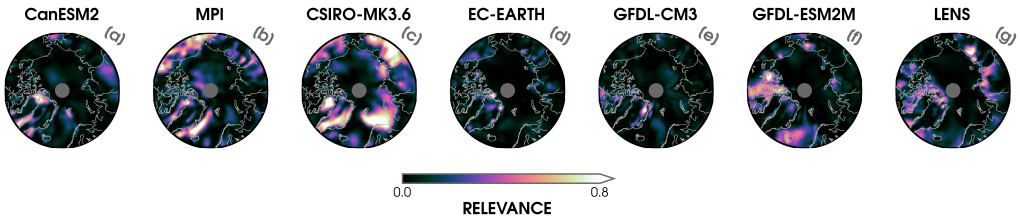

Figure 8. (a) Composite heatmap of LRP averaged over 1950 to 2019 for correct testing data predictions after removing the annual mean of each CanESM2 map before it is fed into the ANN. (b-g) As in (a), but for MPI, CSIRO-MK3.6, EC-EARTH, GFDL-CM3, GFDL-ESM2M, and LENS, respectively.

T2M signals, such as the cold anomaly signatures over the Barents Sea in CSIRO-MK3.6 and near Iceland in GFDL-ESM2M. But there are also some differences in the higher relevance areas, like those found in the heatmap composites for CanESM2 over Siberia and the North Atlantic. These results support our interpretation that the ANN is making predictions by weighting regional patterns of T2M that are unique to each GCM for comparing with observational data.

7 Discussion and Conclusions

There are many existing methods for ranking the skill of climate models against observations (Gleckler et al., 2008; Eyring et al., 2019). This exercise is particularly important for climate sensitive regions, such as the Arctic, which have large spreads and uncertainties in future projections and where guidance for weighting climate model projections is not always necessarily straightforward (Knutti et al., 2017). Some of the advantages for exploring deep learning methods for comparing climate models with observations include their ability to leverage spatial patterns and relationships and approximate any nonlinearities. We attempt to evaluate climate model large ensembles and observational datasets in the Arctic using a simple artificial neural network (ANN) classification framework. That is, we trained ANNs on maps of near-surface temperature (T2M) from the multi-model large ensemble archive and then used the neural network for predicting data from atmospheric reanalysis to see which climate model is classified for each year from 1950 to 2019. To understand the ANN's prediction, we leveraged an explainability method called layer-wise relevance propagation, which revealed that the ANN is using regional temperature patterns, rather than only mean state biases, in order to make each climate model selection.

Although the prediction task itself is quite simple for the ANN to correctly learn which climate model simulated a map of T2M, it is more challenging to interpret the ANN's utility on observations. Here, MPI is most frequently classified by the ANNs for the T2M maps taken from observations, which is likely a result of its mean climate state and patterns of spatial variability that compare closely with ERA5 over both land and ocean areas in the Arctic. Interestingly, we find that this climate model classification for each year of observations produces results rather similar to traditional evaluation metrics, such as comparing with climate models that receive lower root-mean-square-errors. One advantage of our approach is that the ANN can also learn regional relationships across spatial patterns, rather than only computing point-by-point relative error statistics. Further, the relevance maps can be used as tools for highlighting regional pattern fingerprints unique to individual climate models. This is especially true for areas around large temperature gradients. For example, the explainability maps reveal that differences in T2M near Greenland and the marginal ice zone of the North Atlantic are often important for the ANN to correctly identify many of the climate model large ensembles. This is consistent with recent analysis of CMIP6 models (e.g., Cai et al., 2021), which note that cli-

mate model differences may be due to their simulation of Atlantic poleward heat transport.

In future work, it may be interesting to use convolutional neural networks for comparing spatial differences in different climate variables or to try the classification architecture on GCMs prescribed with different future emission scenarios, but that is beyond the scope of this preliminary work. Importantly, we note that the output of this approach is dependent on the selection of preprocessing steps, but these choices can be aligned with the overall scientific question one is interested in addressing. For instance, preliminary work has shown some (albeit lower) skill in classifying maps of temperature anomalies that are calculated with respect to a common baseline or by using data with the ensemble mean first removed. Despite these limitations and future work, this study demonstrates that ANNs have the ability to extract regional patterns that are consistent between climate models and observations, but the overall practicality of translating this approach to existing climate evaluation toolboxes should be further investigated.

Open Research

Climate model large ensemble data used in this study are freely available from the NCAR Climate Data Gateway (`https://www.earthsystemgrid.org/dataset/ucar.cgd.ccsm4.CLIVAR_LE.html`), which is supported by the U.S. National Science Foundation (NSF). Atmospheric reanalysis data are openly available for ERA5 (`https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview`) and the preliminary version of the ERA5 back extension (`https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means-preliminary-back-extension?tab=overview`), which are both supported by the Copernicus Climate Change Service (C3S; Thépaut et al., 2018) Climate Data Store (CDS). Twentieth Century Reanalysis Project version 3 (20CRv3) data are provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA (`https://psl.noaa.gov/data/gridded/data.20thC_ReanV3.html`). References for the datasets are available in Tables S1-S2.

Preprocessing steps were completed using NCO v4.9.3 (Zender, 2008), CDO v1.9.8 (Schulzweida, 2019), and NCL v6.2.2 (NCAR, 2019). Computer code for the ANN architecture, figures, and other exploratory data analysis is available at `https://zenodo`

.org/record/6564106. Python v3.7.6 (Van Rossum & Drake, 2009) packages used for this analysis include Numpy v1.19 (Harris et al., 2020), SciPy v1.4.1 (Virtanen et al., 2020), and Scikit-learn v0.24.2 (Pedregosa et al., 2011). Additional open source software used for development of the ANN and LRP heatmaps include TensorFlow v1.15.0 (Abadi et al., 2016) and iNNvestigate v1.0.8 (Alber et al., 2019). Matplotlib v3.2.2 (Hunter, 2007) was used for plotting figures, and colormaps were provided by cmocean v2.0 (Thyng et al., 2016), Palettable's cubehelix v3.3.0 (Green, 2011), and Scientific v7.0.0 (Crameri, 2018; Crameri et al., 2020).

Conflict of Interest

The Authors declare no conflicts of interest in regard to this study.

Acknowledgments

We thank two anonymous reviewers and the editor for their constructive comments and suggestions, which helped us to improve this manuscript. This study was supported by NOAA MAPP grant NA19OAR4310289 and by the Regional and Global Model Analysis program area of the U.S. Department of Energy's (DOE) Office of Biological and Environmental Research (BER) as part of the Program for Climate Model Diagnosis and Intercomparison project. We would like to acknowledge the US CLIVAR Working Group on Large Ensembles and high-performance computing support from NCAR's Computational and Information Systems Laboratory's (CISL) Cheyenne (doi:10.5065/D6RX99HX) for the development of the Multi-Model Large Ensemble Archive (https://www.cesm.ucar.edu/projects/community-projects/MMLEA/). Lastly, we would like to acknowledge support for the Twentieth Century Reanalysis Project version 3 (20CRv3) dataset provided by the U.S. DOE Office of Science BER, by the NOAA Climate Program Office, and by the NOAA Physical Sciences Laboratory.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th usenix symposium on operating systems design and implementation, osdi 2016.*

Agarap, A. F. (2018, mar). Deep Learning using Rectified Linear Units (ReLU).

arXiv. Retrieved from http://arxiv.org/abs/1803.08375

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., ... Kindermans, P. J. (2019). INNvestigate neural networks! *Journal of Machine Learning Research*, *20*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015, jul). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, *10*(7), e0130140. Retrieved from http://www.hfsp.org/, doi: 10.1371/journal.pone.0130140

Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019, nov). Viewing Forced Climate Patterns Through an AI Lens. *Geophysical Research Letters*, *46*(22), 13389–13398. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2019GL084944 doi: 10.1029/2019GL084944

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020, sep). Indicator Patterns of Forced Change Learned by an Artificial Neural Network. *Journal of Advances in Modeling Earth Systems*, *12*(9). Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2020MS002195 doi: 10.1029/2020MS002195

Bell, B., Hersbach, H., Simmons, A., Berrisford, P., Dahlgren, P., Horányi, A., ... Thépaut, J.-N. (2021). The ERA5 Global Reanalysis: Preliminary Extension to 1950. *Quarterly Journal of the Royal Meteorological Society*. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4174 doi: 10.1002/QJ.4174

Bonan, D. B., Lehner, F., & Holland, M. M. (2021, jan). Partitioning uncertainty in projections of Arctic sea ice. *Environmental Research Letters*. Retrieved from https://iopscience.iop.org/article/10.1088/1748-9326/abe0ec doi: 10.1088/1748-9326/abe0ec

Boukabara, S.-A., Krasnopolsky, V., Penny, S. G., Stewart, J. Q., McGovern, A., Hall, D., ... Hoffman, R. N. (2021, may). Outlook for Exploiting Artificial Intelligence in the Earth and Environmental Sciences. *Bulletin of the American Meteorological Society*, *102*(5), E1016–E1032. Retrieved from https://journals.ametsoc.org/view/journals/bams/102/5/BAMS-D-20-0031.1.xml doi: 10.1175/BAMS-D-20-0031.1

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti,

R. (2020, nov). Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dynamics*, *11*(4), 995–1012. doi: 10.5194/ESD-11-995-2020

Cai, Z., You, Q., Wu, F., Chen, H. W., Chen, D., & Cohen, J. (2021). Arctic warming revealed by multiple CMIP6 models: Evaluation of historical simulations and quantification of future projection uncertainties. *Journal of Climate*, *34*(12). doi: 10.1175/JCLI-D-20-0791.1

Chai, T., & Draxler, R. R. (2014, jun). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. doi: 10.5194/GMD-7-1247-2014

Chapman, W. L., & Walsh, J. E. (2007, feb). Simulations of Arctic Temperature and Pressure by Global Coupled Models. *Journal of Climate*, *20*(4), 609–632. Retrieved from `https://journals.ametsoc.org/view/journals/clim/20/4/jcli4026.1.xml` doi: 10.1175/JCLI4026.1

Ciavarella, A., Cotterill, D., Stott, P., Kew, S., Philip, S., van Oldenborgh, G. J., ... Zolina, O. (2021, may). Prolonged Siberian heat of 2020 almost impossible without human influence. *Climatic Change*, *166*(1-2), 9. Retrieved from `https://link.springer.com/10.1007/s10584-021-03052-w` doi: 10.1007/s10584-021-03052-w

Cohen, J., Zhang, X., Francis, J., Jung, T., Kwok, R., Overland, J., ... Yoon, J. (2020, dec). Divergent consensuses on Arctic amplification influence on midlatitude severe winter weather. *Nature Climate Change*, 1–10. doi: 10.1038/s41558-019-0662-y

Crameri, F. (2018, jan). Scientific colour maps. *Zenodo*. Retrieved from `https://zenodo.org/record/4153113` doi: 10.5281/ZENODO.4153113

Crameri, F., Shephard, G. E., & Heron, P. J. (2020, dec). The misuse of colour in science communication. *Nature Communications*, *11*(1), 1–10. Retrieved from `https://doi.org/10.1038/s41467-020-19160-7` doi: 10.1038/s41467-020-19160-7

Davy, R., & Outten, S. (2020). The Arctic surface climate in CMIP6: Status and developments since CMIP5. *Journal of Climate*, *33*(18). doi: 10.1175/JCLI-D-19-0990.1

Delhasse, A., Kittel, C., Amory, C., Hofer, S., Van As, D., Fausto, R. S., & Fettweis, X. (2020, mar). Brief communication: Evaluation of the near-surface climate in ERA5 over the Greenland Ice Sheet. *Cryosphere*, *14*(3), 957–965. doi: 10.5194/TC-14-957-2020

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., ... Ting, M. (2020, mar). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, 1–10. Retrieved from `http://www.nature.com/articles/s41558-020-0731-2` doi: 10.1038/s41558-020-0731-2

Druckenmiller, M. L., Moon, T. A., Thoman, R. L., Ballinger, T. J., Berner, L. T., Bernhard, G. H., ... Ziel, R. (2021, aug). The Arctic [in "State of the Climate in 2020"]. *Bulletin of the American Meteorological Society*, *102*(8), S263–S316. Retrieved from `https://journals.ametsoc.org/view/journals/bams/102/8/BAMS-D-21-0086.1.xml` doi: 10.1175/BAMS-D-21-0086.1

Dunn, R. J. H., Aldred, F., Gobron, N., Miller, J. B., Willett, K. M., Ades, M., ... Zotta, R. M. (2021, aug). Global Climate [in "State of the Climate in 2020"]. *Bulletin of the American Meteorological Society*, *102*(8), S11–S142. Retrieved from `https://journals.ametsoc.org/view/journals/bams/102/8/BAMS-D-21-0098.1.xml` doi: 10.1175/BAMS-D-21-0098.1

Dutta, D., Sherwood, S. C., Meissner, K. J., Gupta, A. S., Lunt, D. J., Tourte, G. J., ... Brown, J. R. (2021, jul). A Multimodel Investigation of Atmospheric Mechanisms for Driving Arctic Amplification in Warmer Climates. *Journal of Climate*, *34*(14), 5723–5740. Retrieved from `https://journals.ametsoc.org/view/journals/clim/34/14/JCLI-D-20-0354.1.xml` doi: 10.1175/JCLI-D-20-0354.1

England, M., Jahn, A., & Polvani, L. (2019, jul). Nonuniform Contribution of Internal Variability to Recent Arctic Sea Ice Loss. *Journal of Climate*, *32*(13), 4039–4053. Retrieved from `http://journals.ametsoc.org/doi/10.1175/JCLI-D-18-0864.1` doi: 10.1175/JCLI-D-18-0864.1

England, M. R. (2021). Are Multi-Decadal Fluctuations in Arctic and Antarctic Surface Temperatures a Forced Response to Anthropogenic Emissions or Part of Internal Climate Variability? *Geophysical Research Letters*, *48*(6). doi: 10.1029/2020GL090631

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., ... Williamson, M. S. (2019, jan). Taking climate model evaluation to the next level. *Nature Climate Change 2019 9:2*, *9*(2), 102–110. Retrieved from https://www.nature.com/articles/s41558-018-0355-y doi: 10.1038/s41558-018-0355-y

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., ... Rummukainen, M. (2013). IPCC AR5 - Chapter 9: Evaluation of Climate Models. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 741–866.

Friedman, J. H. (2012, jul). Fast sparse regression and classification. *International Journal of Forecasting*, *28*(3), 722–738. doi: 10.1016/j.ijforecast.2012.05.001

Gleckler, P. J., Doutriaux, C., Durack, P. J., Taylor, K. E., Zhang, Y., Williams, D. N., ... Servonnat, J. (2016). A more powerful reality test for climate models. *Eos (United States)*, *97*(12). doi: 10.1029/2016eo051663

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008, mar). Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, *113*(D6), 6104. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1029/2007JD008972 doi: 10.1029/2007JD008972

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*.

Gordon, E. M., Barnes, E. A., & Hurrell, J. W. (2021, nov). Oceanic Harbingers of Pacific Decadal Oscillation Predictability in CESM2 Detected by Neural Networks. *Geophysical Research Letters*, *48*(21), e2021GL095392. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021GL095392 doi: 10.1029/2021GL095392

Graham, R. M., Cohen, L., Ritzhaupt, N., Segger, B., Graversen, R. G., Rinke, A., ... Hudson, S. R. (2019, jul). Evaluation of Six Atmospheric Reanalyses over Arctic Sea Ice from Winter to Early Summer. *Journal of Climate*, *32*(14), 4121–4143. Retrieved from https://journals.ametsoc.org/view/journals/clim/32/14/jcli-d-18-0643.1.xml doi: 10.1175/JCLI-D-18-0643.1

Graham, R. M., Hudson, S. R., & Maturilli, M. (2019, jun). Improved Performance of ERA5 in Arctic Gateway Relative to Four Global Atmospheric Reanalyses. *Geophysical Research Letters*. doi: 10.1029/2019GL082781

Graham, R. M., Rinke, A., Cohen, L., Hudson, S. R., Walden, V. P., Granskog, M. A., ... Maturilli, M. (2017, jun). A comparison of the two Arctic atmospheric winter states observed during N-ICE2015 and SHEBA. *Journal of Geophysical Research: Atmospheres*, *122*(11), 5716–5737. Retrieved from `http://doi.wiley.com/10.1002/2016JD025475` doi: 10.1002/2016JD025475

Green, D. A. (2011). A colour scheme for the display of astronomical intensity images. *Bulletin of the Astronomical Society of India*, *39*(2).

Hahn, L. C., Armour, K. C., Battisti, D. S., Eisenman, I., & Bitz, C. M. (2022, mar). Seasonality in Arctic Warming Driven by Sea Ice Effective Heat Capacity. *Journal of Climate*, *35*(5), 1629–1642. Retrieved from `https://journals-ametsoc-org.ezproxy2.library.colostate.edu/view/journals/clim/35/5/JCLI-D-21-0626.1.xml` doi: 10.1175/JCLI-D-21-0626.1

Harris, C. R., Jarrod Millman, K., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, sep). Array programming with NumPy. *Nature*, *585*(7825), 357. Retrieved from `https://doi.org/10.1038/s41586-020-2649-2` doi: 10.1038/s41586-020-2649-2

Hausfather, Z., & Peters, G. P. (2020). *RCP8.5 is a problematic scenario for near-term emissions* (Vol. 117) (No. 45). doi: 10.1073/pnas.2017124117

Hawkins, E., Smith, R. S., Gregory, J. M., & Stainforth, D. A. (2016, jun). Irreducible uncertainty in near-term climate projections. *Climate Dynamics*, *46*(11-12), 3807–3819. Retrieved from `https://link.springer.com/article/10.1007/s00382-015-2806-8` doi: 10.1007/S00382-015-2806-8

Hazeleger, W., Severijns, C., Semmler, T., Ştefănescu, S., Yang, S., Wang, X., ... Willén, U. (2010, oct). EC-Earth: A Seamless Earth-System Prediction Approach in Action. *Bulletin of the American Meteorological Society*, *91*(10), 1357–1364. Retrieved from `https://journals.ametsoc.org/view/journals/bams/91/10/2010bams2877{\_}1.xml` doi: 10.1175/2010BAMS2877.1

Hedström, A., Weber, L., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., & Höhne, M. M. C. (2022, feb). Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations. *arXiv*. Retrieved from `https://arxiv.org/abs/2202.06861v1`

Henry, M., Merlis, T. M., Lutsko, N. J., & Rose, B. E. (2021, mar). Decomposing the Drivers of Polar Amplification with a Single-Column Model. *Journal of*

Climate, *34*(6), 2355–2365. Retrieved from https://journals.ametsoc.org/view/journals/clim/34/6/JCLI-D-20-0178.1.xml doi: 10.1175/JCLI-D-20-0178.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020, may). The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803 doi: 10.1002/qj.3803

Hilburn, K. A., Ebert-Uphoff, I., & Miller, S. D. (2020). Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using goes-r satellite observations. *Journal of Applied Meteorology and Climatology*, *60*(1). doi: 10.1175/JAMC-D-20-0084.1

Holland, M. M., & Landrum, L. (2021). The Emergence and Transient Nature of Arctic Amplification in Coupled Climate Models. *Frontiers in Earth Science*, *9*. doi: 10.3389/feart.2021.719024

Hunter, J. D. (2007, may). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 99–104. doi: 10.1109/MCSE.2007.55

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021, aug). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, *3*(8), 667–674. Retrieved from https://www.nature.com/articles/s42256-021-00374-3 doi: 10.1038/s42256-021-00374-3

Jack, J. E., Eyre, R., & Zeng, X. (2017). Evaluation of Greenland near surface air temperature datasets. *The Cryosphere*, *11*, 1591–1605. Retrieved from https://doi.org/10.5194/tc-11-1591-2017 doi: 10.5194/tc-11-1591-2017

Jeffrey, S., Rotstayn, L., Collier, M., Dravitzki, S., Hamalainen, C., Moeseneder, C., ... Syktus, J. (2013). Australia's CMIP5 submission using the CSIRO-Mk3.6 model. *Australian Meteorological and Oceanographic Journal*, *63*(1). doi: 10.22499/2.6301.001

Kacimi, S., & Kwok, R. (2022, mar). Arctic snow depth, ice thickness and volume from ICESat-2 and CryoSat-2: 2018-2021. *Geophysical Research Letters*, e2021GL097448. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1029/2021GL097448 doi: 10.1029/2021GL097448

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., ... Vertenstein, M. (2015, aug). 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. Retrieved from http://journals.ametsoc.org/doi/10.1175/BAMS-D-13-00255.1 doi: 10.1175/BAMS-D-13-00255.1

Kirchmeier-Young, M. C., Zwiers, F. W., & Gillett, N. P. (2017, jan). Attribution of Extreme Events in Arctic Sea Ice Extent. *Journal of Climate*, *30*(2), 553–571. Retrieved from https://journals.ametsoc.org/view/journals/clim/30/2/jcli-d-16-0412.1.xml doi: 10.1175/JCLI-D-16-0412.1

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., Eyring, V., ... Eyring, V. (2017, feb). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*(4), 1909–1918. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2016GL072012 doi: 10.1002/2016GL072012

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. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021MS002464 doi: 10.1029/2021MS002464

Landrum, L., & Holland, M. M. (2020). Extremes become routine in an emerging new Arctic. *Nature Climate Change*, *10*(12). doi: 10.1038/s41558-020-0892-z

Lauer, A., Eyring, V., Bellprat, O., Bock, L., Gier, B. K., Hunter, A., ... Zechlau, S. (2020, sep). Earth System Model Evaluation Tool (ESMValTool) v2.0 - Diagnostics for emergent constraints and future projections from Earth system models in CMIP. *Geoscientific Model Development*, *13*(9), 4205–4228. doi: 10.5194/GMD-13-4205-2020

Lecun, Y., Bengio, Y., & Hinton, G. (2015, may). *Deep learning* (Vol. 521) (No. 7553). Nature Publishing Group. Retrieved from https://www.nature.com/articles/nature14539 doi: 10.1038/nature14539

Lee, J., Gleckler, P., Ordonez, A., Ahn, M.-S., Ullrich, P., Vo, T., & Boutte, J. (2021, nov). PCMDI/pcmdi_metrics: PMP Version 2.2.1. *Zenodo*.

Retrieved from https://zenodo.org/record/5784459 doi: 10.5281/ZENODO.5784459

Madakumbura, G. D., Thackeray, C. W., Norris, J., Goldenson, N., & Hall, A. (2021, jul). Anthropogenic influence on extreme precipitation over global land areas seen in multiple observational datasets. *Nature Communications 2021 12:1*, *12*(1), 1–9. Retrieved from https://www.nature.com/articles/s41467-021-24262-x doi: 10.1038/s41467-021-24262-x

Maher, N., Milinski, S., & Ludwig, R. (2021, apr). Large ensemble climate model simulations: introduction, overview, and future prospects for utilising multiple types of large ensemble. *Earth System Dynamics*, *12*(2), 401–418. Retrieved from https://esd.copernicus.org/articles/12/401/2021/ doi: 10.5194/esd-12-401-2021

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., ... Marotzke, J. (2019, jul). The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2050–2069. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019MS001639 doi: 10.1029/2019MS001639

Maher, N., Tabarin, T. P., & Milinski, S. (2022). Combining machine learning and SMILEs to classify, better understand, and project changes in ENSO events. *Earth System Dynamics Discussions*, 1–24. Retrieved from https://doi.org/10.5194/esd-2021-105 doi: 10.5194/esd-2021-105

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021, mar). Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset. *arXiv*. Retrieved from http://arxiv.org/abs/2103.10005

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

McCarty, J. L., Aalto, J., Paunu, V.-V., Arnold, S. R., Eckhardt, S., Klimont, Z., ... Wilson, S. (2021, sep). Reviews and syntheses: Arctic fire regimes and emissions in the 21st century. *Biogeosciences*, *18*(18), 5053–5083. Retrieved from https://bg.copernicus.org/articles/18/5053/2021/ doi:

10.5194/BG-18-5053-2021

McCrystall, M. R., & Screen, J. A. (2021). Arctic Winter Temperature Variations Correlated With ENSO Are Dependent on Coincidental Sea Ice Changes. *Geophysical Research Letters*, *48*(8). doi: 10.1029/2020GL091519

McKenna, C. M., & Maycock, A. C. (2021). Sources of Uncertainty in Multimodel Large Ensemble Projections of the Winter North Atlantic Oscillation. *Geophysical Research Letters*, *48*(14). doi: 10.1029/2021GL093258

Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., & Knutti, R. (2020, sep). An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. *Earth System Dynamics*, *11*(3), 807–834. doi: 10.5194/ESD-11-807-2020

Miner, K. R., Turetsky, M. R., Malina, E., Bartsch, A., Tamminen, J., McGuire, A. D., ... Miller, C. E. (2022, jan). Permafrost carbon emissions in a changing Arctic. *Nature Reviews Earth & Environment 2022 3:1*, *3*(1), 55–67. Retrieved from `https://www.nature.com/articles/s43017-021-00230-3` doi: 10.1038/s43017-021-00230-3

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-Wise Relevance Propagation: An Overview. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 11700 LNCS). doi: 10.1007/978-3-030-28954-6_10

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017, may). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*, 211–222. doi: 10.1016/j.patcog.2016.11.008

Montavon, G., Samek, W., & Müller, K. R. (2018, feb). *Methods for interpreting and understanding deep neural networks* (Vol. 73). Elsevier Inc. doi: 10.1016/j.dsp.2017.10.011

Mouginot, J., Rignot, E., Bjørk, A. A., van den Broeke, M., Millan, R., Morlighem, M., ... Wood, M. (2019, may). Forty-six years of Greenland Ice Sheet mass balance from 1972 to 2018. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(19), 9239–9244. doi: 10.1073/PNAS.1904242116/SUPPL_FILE/PNAS.1904242116.SD02.XLSX

Myers-Smith, I. H., Kerby, J. T., Phoenix, G. K., Bjerke, J. W., Epstein, H. E.,

Assmann, J. J., ... Wipf, S. (2020, jan). Complexity revealed in the greening of the Arctic. *Nature Climate Change 2020 10:2*, *10*(2), 106–117. Retrieved from `https://www.nature.com/articles/s41558-019-0688-1` doi: 10.1038/s41558-019-0688-1

NCAR. (2019). *The NCAR Command Language (Version 6.6.2).* Boulder, Colorado. Retrieved from `http://dx.doi.org/10.5065/D6WD3XH5` doi: http://dx.doi.org/10.5065/D6WD3XH5

NCAR. (2020). *US CLIVAR Multi-Model LE Archive.* Retrieved from `https://www.cesm.ucar.edu/projects/community-projects/MMLEA/`

Neapolitan, R. E., & Jiang, X. (2018, nov). Neural Networks and Deep Learning. In *Artificial intelligence* (pp. 389–411). Chapman and Hall/CRC. doi: 10.1201/b22400-15

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence o(1/kˆ2). *Doklady AN USSR*, *269*.

Nichol, J. J., Peterson, M. G., Fricke, G. M., & Peterson, K. J. (2021). Learning Why: Data-Driven Causal Evaluations of Climate Models. *Tackling Climate Change with Machine Learning Workshop at ICML 2021*, 1–5.

Nichol, J. J., Peterson, M. G., Peterson, K. J., Fricke, G. M., & Moses, M. E. (2021, oct). Machine learning feature analysis illuminates disparity between E3SM climate models and observed climate change. *Journal of Computational and Applied Mathematics*, *395*, 113451. doi: 10.1016/J.CAM.2021.113451

Nowack, P., Runge, J., Eyring, V., & Haigh, J. D. (2020, mar). Causal networks for climate model evaluation and constrained projections. *Nature Communications 2020 11:1*, *11*(1), 1–11. Retrieved from `https://www.nature.com/articles/s41467-020-15195-y` doi: 10.1038/s41467-020-15195-y

Nygård, T., Tjernström, M., & Naakka, T. (2021, dec). Winter thermodynamic vertical structure in the Arctic atmosphere linked to large-scale circulation. *Weather and Climate Dynamics*, *2*(4), 1263–1282. doi: 10.5194/WCD-2-1263-2021

Park, T. W., Deng, Y., Cai, M., Jeong, J. H., & Zhou, R. (2014). A dissection of the surface temperature biases in the Community Earth System Model. *Climate Dynamics*, *43*(7-8). doi: 10.1007/s00382-013-2029-9

Parkinson, C. L., & DiGirolamo, N. E. (2021, dec). Sea ice extents continue to set

new records: Arctic, Antarctic, and global results. *Remote Sensing of Environment*, *267*, 112753. doi: 10.1016/J.RSE.2021.112753

Pasini, A., Racca, P., Amendola, S., Cartocci, G., & Cassardo, C. (2017, dec). Attribution of recent temperature behaviour reassessed by a neural-network method. *Scientific Reports 2017 7:1*, *7*(1), 1–10. Retrieved from https://www.nature.com/articles/s41598-017-18011-8 doi: 10.1038/s41598-017-18011-8

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*.

Peings, Y., Labe, Z. M., & Magnusdottir, G. (2021). Are 100 ensemble members enough to capture the remote atmospheric response to + 2C Arctic sea ice loss ? *Journal of Climate*, *34*(10), 3751–3769. Retrieved from https://journals.ametsoc.org/view/journals/clim/aop/JCLI-D-20-0613.1/JCLI-D-20-0613.1.xml doi: 10.1175/JCLI-D-20-0613.1.

Peters, G. P., & Hausfather, Z. (2020). Emissions - the 'business as usual' story is misleading. *Nature*, *577*.

Phillips, A. S., Deser, C., & Fasullo, J. (2014, dec). Evaluating Modes of Variability in Climate Models. *Eos, Transactions American Geophysical Union*, *95*(49), 453–455. Retrieved from http://doi.wiley.com/10.1002/2014EO490002 doi: 10.1002/2014EO490002

Phillips, A. S., Deser, C., Fasullo, J., Schneider, D. P., & Simpson, I. R. (2020). *Assessing Climate Variability and Change in Model Large Ensembles: A User's Guide to the Climate Variability Diagnostics Package for Large Ensembles Version 1.0* (Tech. Rep.). Boulder, Colorado: NCAR. Retrieved from https://opensky.ucar.edu/islandora/object/manuscripts:1001 doi: 10.5065/h7c7-f961

Previdi, M., Smith, K. L., & Polvani, L. M. (2021). *Arctic amplification of climate change: A review of underlying mechanisms* (Vol. 16) (No. 9). doi: 10.1088/1748-9326/ac1c29

Rader, J. K., Barnes, E. A., Ebert-Uphoff, I., & Anderson, C. (2021). Detection of forced change within combined climate fields using explainable neural networks. *ESSOAr*. Retrieved from http://www.essoar.org/doi/10.1002/

983 essoar.10509261.1 doi: 10.1002/ESSOAR.10509261.1

984 Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., ...
985 Taylor, K. E. (2007). *Cilmate Models and Their Evaluation. In: Climate*
986 *Change 2007: The Physical Science Basis.* (Tech. Rep.). Retrieved from
987 https://www.ipcc.ch/report/ar4/wg1/

988 Rasu, E., Bernstein, R., Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen,
989 H. M., ... Yang, H. (2019). Machine learning and artificial intelligence
990 to aid climate change research and preparedness. *Environ. Res. Lett*, *14*,
991 124007. Retrieved from https://doi.org/10.1088/1748-9326/ab4e55 doi:
992 10.1088/1748-9326/ab4e55

993 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais,
994 N., & Prabhat. (2019, feb). Deep learning and process understanding
995 for data-driven Earth system science. *Nature*, *566*(7743), 195–204. Re-
996 trieved from https://www.nature.com/articles/s41586-019-0912-1 doi:
997 10.1038/s41586-019-0912-1

998 Retsch, M. H., Jakob, C., & Singh, M. S. (2022, feb). Identifying Relations Be-
999 tween Deep Convection and the Large-Scale Atmosphere Using Explainable
1000 Artificial Intelligence. *Journal of Geophysical Research: Atmospheres*, *127*(3),
1001 e2021JD035388. Retrieved from https://onlinelibrary.wiley.com/doi/
1002 full/10.1029/2021JD035388 doi: 10.1029/2021JD035388

1003 Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., ... Rafaj, P. (2011,
1004 aug). RCP 8.5—A scenario of comparatively high greenhouse gas emissions.
1005 *Climatic Change*, *109*(1-2), 33–57. Retrieved from http://link.springer
1006 .com/10.1007/s10584-011-0149-y doi: 10.1007/s10584-011-0149-y

1007 Rodgers, K. B., Lin, J., & Frölicher, T. L. (2015, jun). Emergence of multiple ocean
1008 ecosystem drivers in a large ensemble suite with an Earth system model. *Bio-*
1009 *geosciences*, *12*(11), 3301–3320. doi: 10.5194/BG-12-3301-2015

1010 Ruder, S. (2016, sep). An overview of gradient descent optimization algorithms.
1011 *arXiv*. Retrieved from http://arxiv.org/abs/1609.04747

1012 Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Muller, K.-R. (2019). Ex-
1013 plainable AI: Interpreting, Explaining and Visualizing Deep Learning. *Lecture*
1014 *Notes in Computer Science (LNCS)*, *11700*.

1015 Sanderson, B. M., Pendergrass, A. G., Koven, C. D., Brient, F., Booth, B. B.,

Fisher, R. A., & Knutti, R. (2021, aug). The potential for structural errors in emergent constraints. *Earth System Dynamics*, *12*(3), 899–918. doi: 10.5194/ESD-12-899-2021

Schulzweida, U. (2019, feb). CDO User Guide. *Zenodo*. Retrieved from `https://zenodo.org/record/2558193` doi: 10.5281/ZENODO.2558193

Schweiger, A. J., Wood, K. R., & Zhang, J. (2019, aug). Arctic Sea Ice Volume Variability over 1901–2010: A Model-Based Reconstruction. *Journal of Climate*, *32*(15), 4731–4752. Retrieved from `http://journals.ametsoc.org/doi/10.1175/JCLI-D-19-0008.1` doi: 10.1175/JCLI-D-19-0008.1

Slivinski, L. C., Compo, G. P., Sardeshmukh, P. D., Whitaker, J. S., McColl, C., Allan, R. J., ... Wyszynski, P. (2021, feb). An Evaluation of the Performance of the Twentieth Century Reanalysis Version 3. *Journal of Climate*, *34*(4), 1417–1438. Retrieved from `https://journals.ametsoc.org/view/journals/clim/34/4/JCLI-D-20-0505.1.xml` doi: 10.1175/JCLI-D-20-0505.1

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., ... Wyszyński, P. (2019, oct). Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 2876–2908. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3598` doi: 10.1002/qj.3598

Solomon, A., Heuzé, C., Rabe, B., Bacon, S., Bertino, L., Heimbach, P., ... Tang, H. (2021, aug). Freshwater in the Arctic Ocean 2010-2019. *Ocean Science*, *17*(4), 1081–1102. doi: 10.5194/OS-17-1081-2021

Sonnewald, M., & Lguensat, R. (2021, aug). Revealing the Impact of Global Heating on North Atlantic Circulation Using Transparent Machine Learning. *Journal of Advances in Modeling Earth Systems*, *13*(8), e2021MS002496. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021MS002496` doi: 10.1029/2021MS002496

Stainforth, D. A., Allen, M. R., Tredger, E. R., & Smith, L. A. (2007, aug). Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2145–2161. Retrieved from `https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2007.2074`

doi: 10.1098/RSTA.2007.2074

Sun, L., Alexander, M., & Deser, C. (2018, oct). Evolution of the Global Coupled Climate Response to Arctic Sea Ice Loss during 1990–2090 and Its Contribution to Climate Change. *Journal of Climate*, *31*(19), 7823–7843. Retrieved from `https://journals.ametsoc.org/view/journals/clim/31/19/jcli-d-18-0134.1.xml` doi: 10.1175/JCLI-D-18-0134.1

Sundararajan, M., Taly, A., & Yan, Q. (2017, mar). Axiomatic Attribution for Deep Networks. *34th International Conference on Machine Learning, ICML 2017*, *7*, 5109–5118. Retrieved from `https://arxiv.org/abs/1703.01365v2` doi: 10.48550/arxiv.1703.01365

Swart, N. C., Fyfe, J. C., Hawkins, E., Kay, J. E., & Jahn, A. (2015, jan). Influence of internal variability on Arctic sea-ice trends. *Nature Climate Change*, *5*(2), 86–89. Retrieved from `http://www.nature.com/doifinder/10.1038/nclimate2483` doi: 10.1038/nclimate2483

Taylor, K. E. (2001, apr). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183–7192. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2000JD900719` doi: 10.1029/2000JD900719

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012, apr). *An overview of CMIP5 and the experiment design* (Vol. 93) (No. 4). American Meteorological Society. doi: 10.1175/BAMS-D-11-00094.1

Taylor, P. C., Boeke, R. C., Boisvert, L. N., Feldl, N., Henry, M., Huang, Y., ... Tan, I. (2022, feb). Process Drivers, Inter-Model Spread, and the Path Forward: A Review of Amplified Arctic Warming. *Frontiers in Earth Science*, *9*, 1391. doi: 10.3389/FEART.2021.758361/BIBTEX

Tepes, P., Gourmelen, N., Nienow, P., Tsamados, M., Shepherd, A., & Weissgerber, F. (2021, aug). Changes in elevation and mass of Arctic glaciers and ice caps, 2010–2017. *Remote Sensing of Environment*, *261*, 112481. doi: 10.1016/J.RSE.2021.112481

Thépaut, J. N., Pinty, B., Dee, D., & Engelen, R. (2018). The Copernicus programme and its climate change service. In *International geoscience and remote sensing symposium (igarss)* (Vol. 2018-July). doi: 10.1109/IGARSS.2018.8518067

Thyng, K., Greene, C., Hetland, R., Zimmerle, H., & DiMarco, S. (2016, sep). True Colors of Oceanography: Guidelines for Effective and Accurate Colormap Selection. *Oceanography*, *29*(3), 9–13. Retrieved from `https://tos.org/oceanography/article/true-colors-of-oceanography-guidelines-for-effective-and-accurate-colormap` doi: 10.5670/oceanog.2016.66

Timmermans, M. L., Toole, J., & Krishfield, R. (2018, aug). Warming of the interior Arctic Ocean linked to sea ice losses at the basin margins. *Science Advances*, *4*(8), eaat6773. doi: 10.1126/sciadv.aat6773

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020, sep). Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, *12*(9). Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2019MS002002` doi: 10.1029/2019MS002002

Topál, D., Ding, Q., Mitchell, J., Baxter, I., Herein, M., Haszpra, T., ... Li, Q. (2020). An internal atmospheric process determining summertime Arctic sea ice melting in the next three decades: Lessons learned from five large ensembles and multiple CMIP5 climate simulations. *Journal of Climate*, *33*(17). doi: 10.1175/JCLI-D-19-0803.1

Uotila, P., O'Farrell, S., Marsland, S. J., & Bi, D. (2013). The sea-ice performance of the Australian climate models participating in the CMIP5. *Australian Meteorological and Oceanographic Journal*, *63*(1). doi: 10.22499/2.6301.008

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3). doi: 10.1038/s41592-019-0686-2

Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., ... Rose, S. K. (2011, aug). The representative concentration pathways: an overview. *Climatic Change*, *109*(1-2), 5–31. Retrieved from `http://link.springer.com/10.1007/s10584-011-0148-z` doi: 10.1007/s10584-011-0148-z

Wang, C., Graham, R. M., Wang, K., Gerland, S., & Granskog, M. A. (2019, jun).

Comparison of ERA5 and ERA-Interim near-surface air temperature, snowfall and precipitation over Arctic sea ice: effects on sea ice thermodynamics and evolution. *Cryosphere*, *13*(6), 1661–1679. doi: 10.5194/tc-13-1661-2019

Weidman, S. K., Delworth, T. L., Kapnick, S. B., & Cooke, W. F. (2021, aug). The Alaskan Summer 2019 Extreme Heat Event: The Role of Anthropogenic Forcing, and Projections of the Increasing Risk of Occurrence. *Earth's Future*, *9*(8), e2021EF002163. Retrieved from `https://onlinelibrary.wiley.com/doi/full/10.1029/2021EF002163` doi: 10.1029/2021EF002163

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2017). *Climate and other models may be more accurate than reported* (Vol. 98) (No. 9). doi: 10.1029/2017eo074939

Yu, Y., Xiao, W., Zhang, Z., Cheng, X., Hui, F., & Zhao, J. (2021). Evaluation of 2-m air temperature and surface temperature from ERA5 and ERA-I using buoy observations in the arctic during 2010–2020. *Remote Sensing*, *13*(14). doi: 10.3390/rs13142813

Zender, C. S. (2008). Analysis of self-describing gridded geoscience data with netCDF Operators (NCO). *Environmental Modelling and Software*, *23*(10-11). doi: 10.1016/j.envsoft.2008.03.004

Zhang, R., Wang, H., Fu, Q., Rasch, P. J., Wu, M., & Maslowski, W. (2021). Understanding the Cold Season Arctic Surface Warming Trend in Recent Decades. *Geophysical Research Letters*, *48*(19). doi: 10.1029/2021GL094878

Zhang, W., Wang, Y., Smeets, P. C., Reijmer, C. H., Huai, B., Wang, J., & Sun, W. (2021, sep). Estimating near-surface climatology of multi-reanalyses over the Greenland Ice Sheet. *Atmospheric Research*, *259*, 105676. doi: 10.1016/J.ATMOSRES.2021.105676

References From the Supporting Information

Bell, B., Hersbach, H., Simmons, A., Berrisford, P., Dahlgren, P., Horányi, A., ... Thépaut, J.-N. (2021). The ERA5 Global Reanalysis: Preliminary Extension to 1950. *Quarterly Journal of the Royal Meteorological Society*. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4174` doi: 10.1002/QJ.4174

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N.,

... Ting, M. (2020, mar). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, 1–10. Retrieved from http://www.nature.com/articles/s41558-020-0731-2 doi: 10.1038/s41558-020-0731-2

Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010, dec). GLOBAL SURFACE TEMPERATURE CHANGE. *Reviews of Geophysics*, *48*(4), 4004. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2010RG000345 doi: 10.1029/2010RG000345

Hazeleger, W., Severijns, C., Semmler, T., Ştefănescu, S., Yang, S., Wang, X., ... Willén, U. (2010, oct). EC-Earth: A Seamless Earth-System Prediction Approach in Action. *Bulletin of the American Meteorological Society*, *91*(10), 1357–1364. Retrieved from https://journals.ametsoc.org/view/journals/bams/91/10/2010bams2877{_}1.xml doi: 10.1175/2010BAMS2877.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020, may). The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803 doi: 10.1002/qj.3803

Jeffrey, S., Rotstayn, L., Collier, M., Dravitzki, S., Hamalainen, C., Moeseneder, C., ... Syktus, J. (2013). Australia's CMIP5 submission using the CSIRO-Mk3.6 model. *Australian Meteorological and Oceanographic Journal*, *63*(1). doi: 10.22499/2.6301.001

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., ... Vertenstein, M. (2015, aug). 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. Retrieved from http://journals.ametsoc.org/doi/10.1175/BAMS-D-13-00255.1 doi: 10.1175/BAMS-D-13-00255.1

Kirchmeier-Young, M. C., Zwiers, F. W., & Gillett, N. P. (2017, jan). Attribution of Extreme Events in Arctic Sea Ice Extent. *Journal of Climate*, *30*(2), 553–571. Retrieved from https://journals.ametsoc.org/view/journals/clim/30/2/jcli-d-16-0412.1.xml doi: 10.1175/JCLI-D-16-0412.1

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., & Zyss, D. (2019, jun). Improvements in the GISTEMP Uncertainty Model. *Journal of Geophysical Research: Atmospheres*, *124*(12), 6307–6326. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2018JD029522 doi: 10.1029/2018JD029522

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., ... Marotzke, J. (2019, jul). The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2050–2069. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019MS001639 doi: 10.1029/2019MS001639

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., ... Simpson, I. R. (2021, feb). An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set. *Journal of Geophysical Research: Atmospheres*, *126*(3), e2019JD032361. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019JD032361 doi: 10.1029/2019JD032361

NCAR. (2020). *US CLIVAR Multi-Model LE Archive.* Retrieved from https://www.cesm.ucar.edu/projects/community-projects/MMLEA/

Rodgers, K. B., Lin, J., & Frölicher, T. L. (2015, jun). Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences*, *12*(11), 3301–3320. doi: 10.5194/BG-12-3301-2015

Rohde, R. A., & Hausfather, Z. (2020, dec). The Berkeley Earth Land/Ocean Temperature Record. *Earth System Science Data*, *12*(4), 3469–3479. Retrieved from https://essd.copernicus.org/articles/12/3469/2020/ doi: 10.5194/essd-12-3469-2020

Slivinski, L. C., Compo, G. P., Sardeshmukh, P. D., Whitaker, J. S., McColl, C., Allan, R. J., ... Wyszynski, P. (2021, feb). An Evaluation of the Performance of the Twentieth Century Reanalysis Version 3. *Journal of Climate*, *34*(4), 1417–1438. Retrieved from https://journals.ametsoc.org/view/journals/clim/34/4/JCLI-D-20-0505.1.xml doi: 10.1175/JCLI-D-20-0505.1

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., ... Wyszyński, P. (2019, oct). Towards a more reliable historical

reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 2876–2908. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3598` doi: 10.1002/qj.3598

Sun, L., Alexander, M., & Deser, C. (2018, oct). Evolution of the Global Coupled Climate Response to Arctic Sea Ice Loss during 1990–2090 and Its Contribution to Climate Change. *Journal of Climate*, *31*(19), 7823–7843. Retrieved from `https://journals.ametsoc.org/view/journals/clim/31/19/jcli-d-18-0134.1.xml` doi: 10.1175/JCLI-D-18-0134.1

Supporting Information for "Comparison of climate model large ensembles with observations in the Arctic using simple neural networks"

Zachary M. Labe[1]and Elizabeth A. Barnes[1]

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

Contents of this file

Table S.1. Summary of the climate model simulation used from the multi-model large ensemble archive (MMLEA) in this study (see NCAR, 2020; Deser et al., 2020).

Name	CMIP5 Forcing	Years	# Members	Horizontal Resolution (Atmosphere / Ocean)	Reference
CanESM2	Historical to 2005, RCP 8.5	1950-2019	16	$\sim 2.8^\circ \times 2.8^\circ / \sim 1.4^\circ \times 0.9^\circ$	Kirchmeier-Young, Zwiers, and Gillett (2017)
MPI	Historical to 2005, RCP 8.5	1950-2019	16	$\sim 1.9^\circ \times 1.9^\circ$/nominal 1.5°	Maher et al. (2019)
CSIRO-MK3.6	Historical to 2005, RCP 8.5	1950-2019	16	$\sim 1.9^\circ \times 1.9^\circ / \sim 1.9^\circ \times 1.0^\circ$	Jeffrey et al. (2013)
EC-EARTH	Historical to 2005, RCP 8.5	1950-2019	16	$\sim 1.1^\circ \times 1.1^\circ$/nominal 1.0°	Hazeleger et al. (2010)
GFDL-CM3	Historical to 2005, RCP 8.5	1950-2019	16	$\sim 2.0^\circ \times 2.5^\circ / \sim 1.0^\circ \times 0.9^\circ$	Sun, Alexander, and Deser (2018)
GFDL-ESM2M	Historical to 2005, RCP 8.5	1950-2019	16	$\sim 2.0^\circ \times 2.5^\circ / \sim 1.0^\circ \times 0.9^\circ$	Rodgers, Lin, and Frölicher (2015)
LENS	Historical to 2005, RCP 8.5	1950-2019	16	$\sim 1.3^\circ \times 0.9^\circ$/nominal 1.0°	Kay et al. (2015)

Table S.2. Summary of atmospheric reanalysis data used in this study.

Name	Data Set	Years	Reference
ERA5	ECMWF Reanalysis v5	1979–2019	Hersbach et al. (2020)
ERA5-BE	ECMWF Reanalysis v5 Back Extension	1950-1978	Bell et al. (2021)
20CRv3	NOAA-CIRES-DOE Twentieth Century Reanalysis v3	1950-2015	Slivinski et al. (2019, 2021)

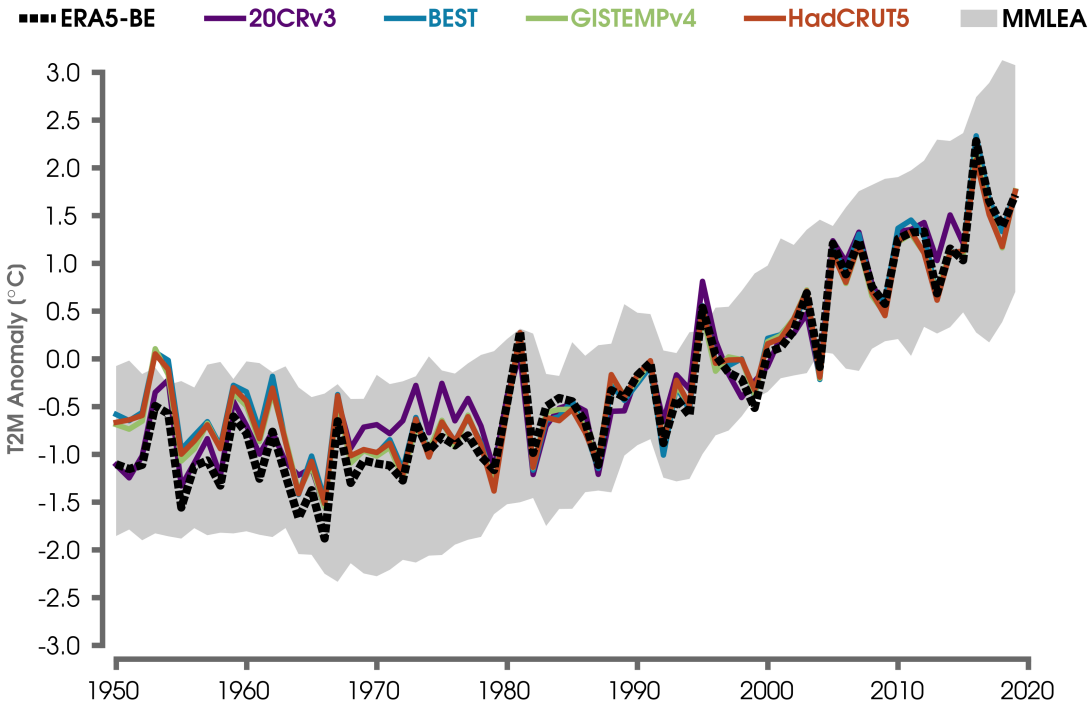

Figure S1. Time series showing annual mean near-surface temperature (T2M; °C) anomalies averaged over the Arctic (60°N-87°N) from 1950 to 2019 using European Centre for Medium-Range Weather Forecasts ERA5 preliminary back extension (ERA5-BE; dashed black line) (Hersbach et al., 2020; Bell et al., 2021), National Oceanic and Atmospheric Administration/Cooperative Institute for Research in Environmental Sciences/Department of Energy Twentieth Century Reanalysis (20CR) version 3 (20CRv3; solid purple line) (Slivinski et al., 2019), Berkeley Earth Land/Ocean Temperature Record (BEST; solid blue line) (Rohde & Hausfather, 2020), Goddard Institute for Space Studies Surface Temperature product version 4 (GISTEMPv4; solid green line) (Hansen et al., 2010; Lenssen et al., 2019), and Hadley Centre/Climatic Research Unit Temperature version 5.0.1.0 (HadCRUT5; solid orange line) (Morice et al., 2021). Note that 20CRv3 is only available from 1950 to 2015. Gray shading shows the 5th-95th percentiles of T2M anomalies in the Arctic across all 7 global climate models with 16 ensemble members each (as used in the main analysis) from the multi-model large ensemble archive (MMLEA) (Deser et al., 2020). All anomalies are computed in respect to their 1981 to 2010 climatology.

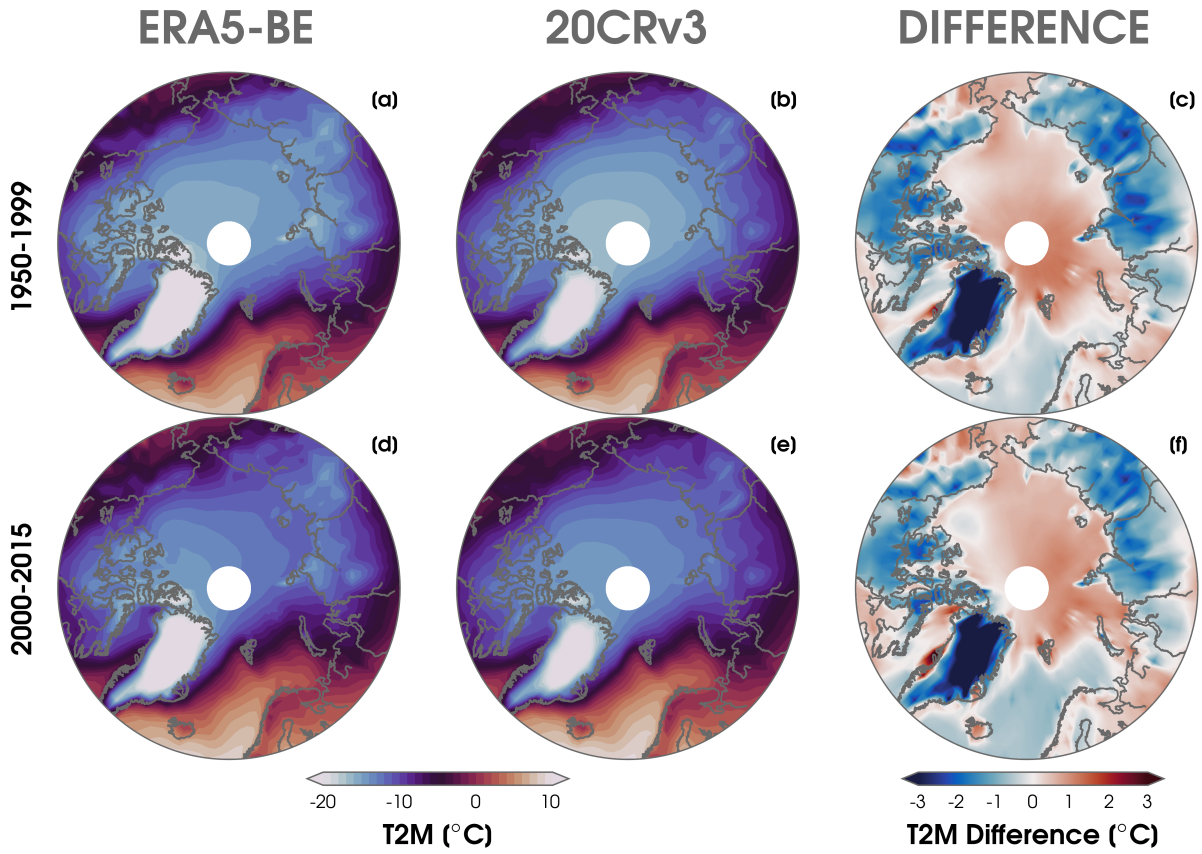

Figure S2. (a) Composite of T2M (contour interval of 1°C) for ERA5-BE reanalysis averaged over the 1950 to 1999 period. (b), As in (a), but for 20CRv3. (c) Difference in annual mean T2M for ERA5-BE minus 20CRv3 over the 1950 to 1999 period. (d-f) As in (a-c), but for the 2000 to 2015 period.

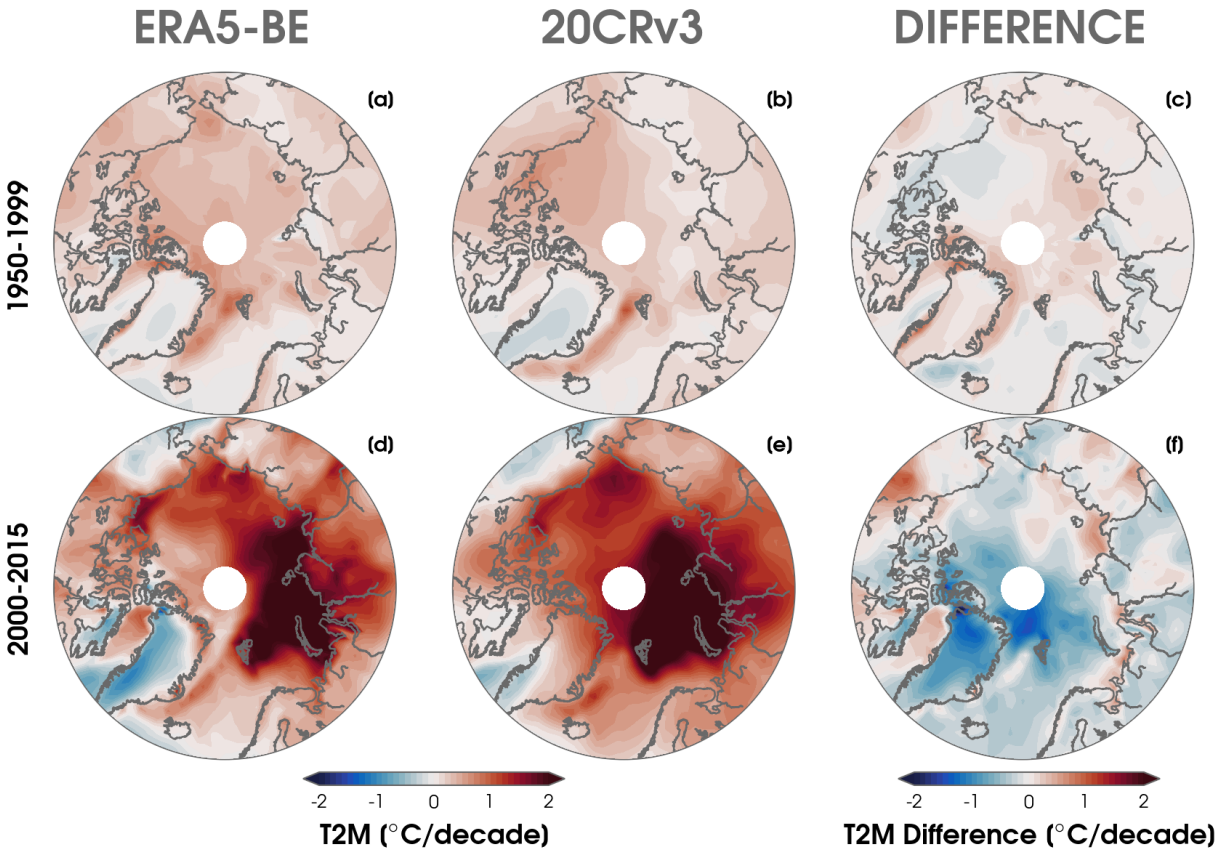

Figure S3. (a) Annual linear least squares trends of T2M (°C per decade) for ERA5-BE reanalysis over the 1950 to 1999 period. (b), As in (a), but for 20CRv3. (c) Difference in decadal trends of T2M for ERA5-BE minus 20CRv3 over the 1950 to 1999 period. (d-f) As in (a-c), but for the 2000 to 2015 period.

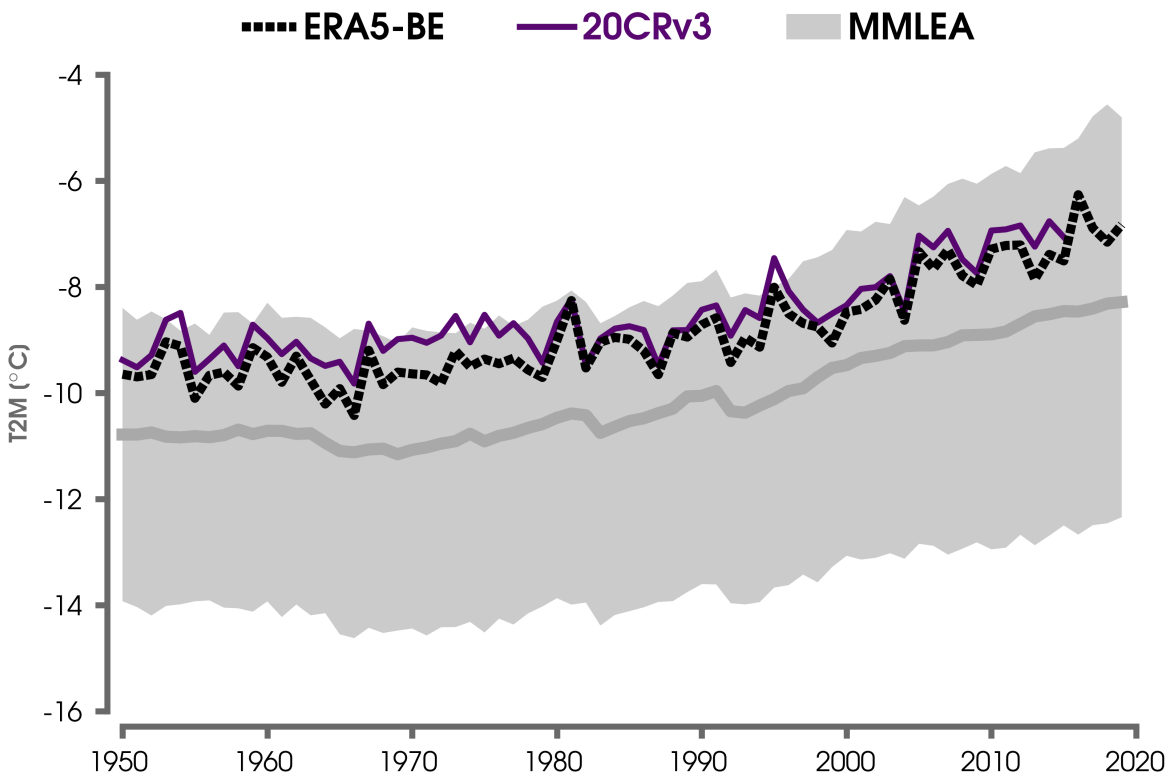

Figure S4. Time series showing annual mean T2M (°C) averaged over the Arctic (60°N-87°N) from 1950 to 2019 using ERA5-BE (dashed black line) and 20CRv3 (solid purple line). Note that 20CRv3 is only available from 1950 to 2015. Gray shading shows the 5th-95th percentiles of T2M in the Arctic across all 7 global climate models with 16 ensemble members each (as used in the main analysis) from the MMLEA. The solid gray line shows the mean by considering all 7 global climate models and their ensemble members.

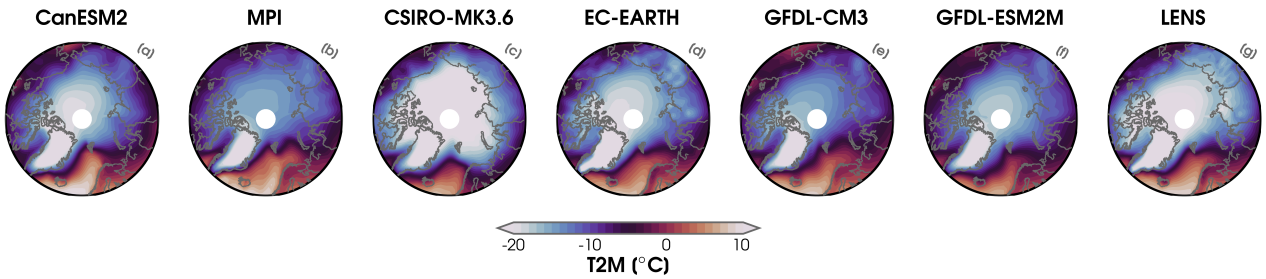

Figure S5. (a-g) Composite of T2M (contour interval of 1°C) for the ensemble mean averaged over the 1950 to 2019 period for CanESM2, MPI, CSIRO-MK3.6, EC-EARTH, GFDL-CM3, GFDL-ESM2M, and LENS, respectively.

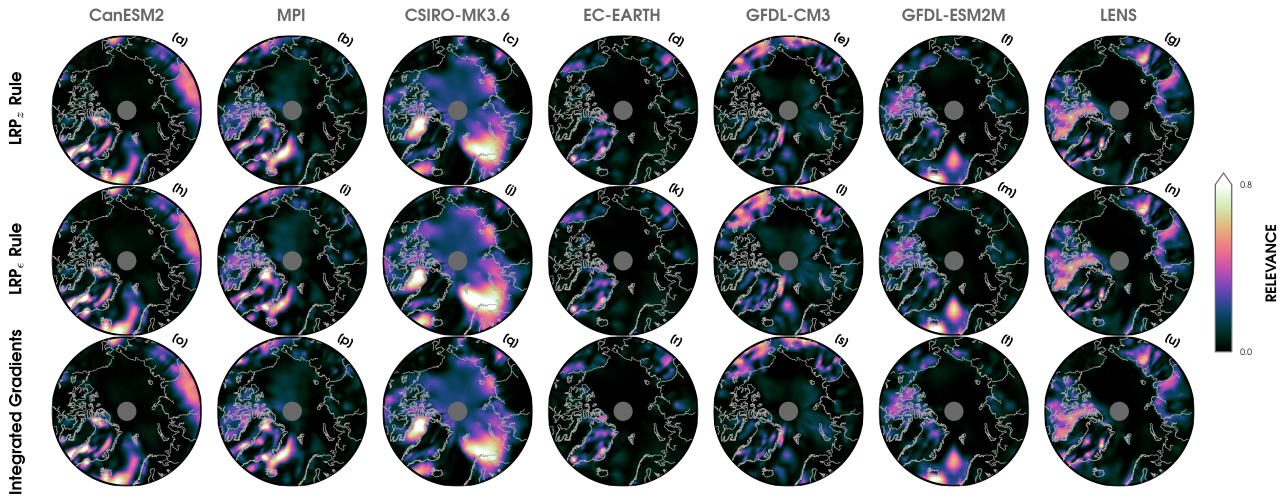

Figure S6. (a-g) Composite heatmap using the layer-wise relevance propagation z-rule (LRP_z) for correct testing data predictions averaged over the 1950 to 2019 period for CanESM2, MPI, CSIRO-MK3.6, EC-EARTH, GFDL-CM3, GFDL-ESM2M, and LENS, respectively. (h-n) As in (a-g), but calculated using the LRP epsilon-rule (LRP_ϵ). (o-u) As in (a-g), but calculated using the Integrated Gradients method. Higher values indicate greater relevance for the ANN's prediction.

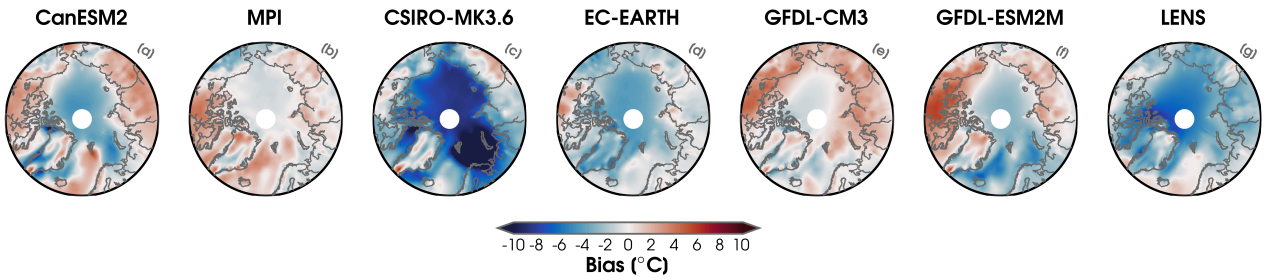

Figure S7. (a-g) Composite of the mean T2M bias (contour interval of 0.1°C) for the ensemble mean averaged over the 1950 to 2019 period for CanESM2, MPI, CSIRO-MK3.6, EC-EARTH, GFDL-CM3, GFDL-ESM2M, and LENS, respectively. The T2M bias is calculated as the difference from each climate model ensemble member minus ERA5-BE.

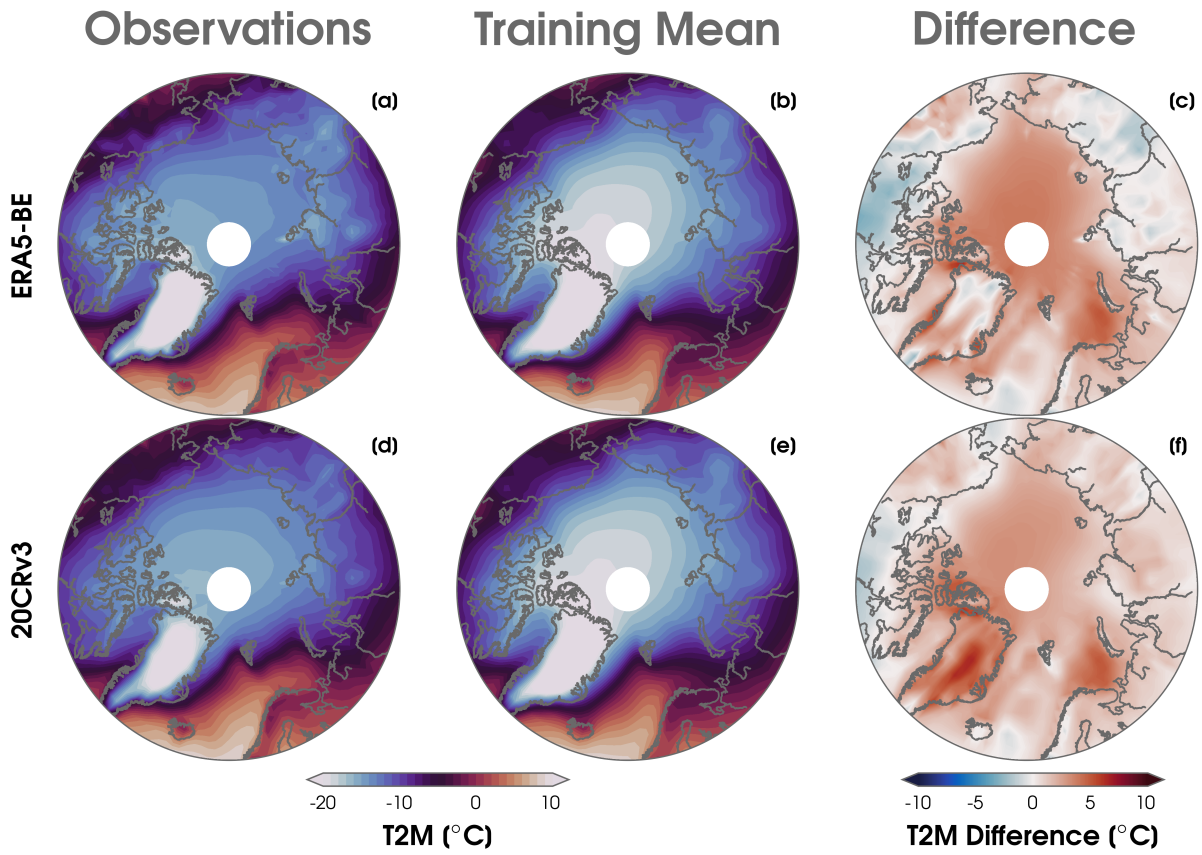

Figure S8. (a) Composite of T2M (contour interval of 1°C) for ERA5-BE reanalysis averaged over the 1950 to 2019 period. (b) Composite of T2M calculated from the mean of the training data (see text for details) using 12 ensemble members from the 7 climate models over 1950 to 2019. (c) Difference in annual mean T2M for ERA5-BE minus the training mean. (d-f) As in (a-c), but for using 20CRv3 from 1950 to 2015.

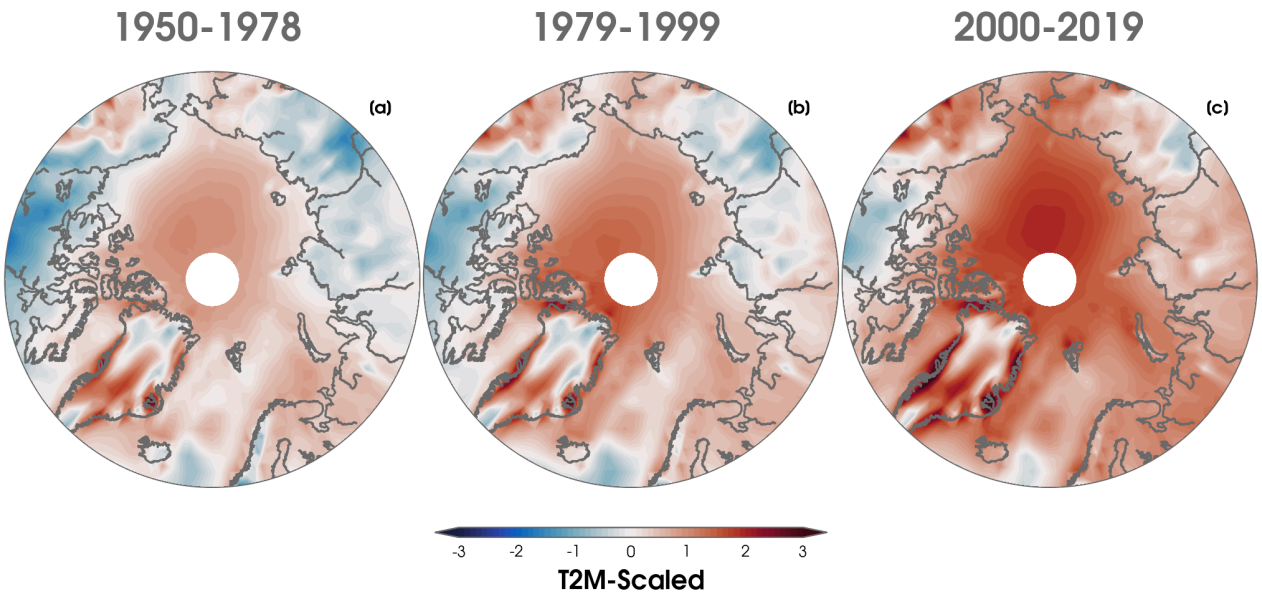

Figure S9. (a) Composite of ERA5-BE T2M over the 1950 to 1978 period, which is rescaled by subtracting the mean of the training data and dividing by the standard deviation of the training data. The training mean and standard deviation are calculated at every grid point by considering all years from 1950 to 2019. (b) As in (a), but for a composite over 1979 to 1999. (c) As in (a), but for a composite over 2000 to 2019.

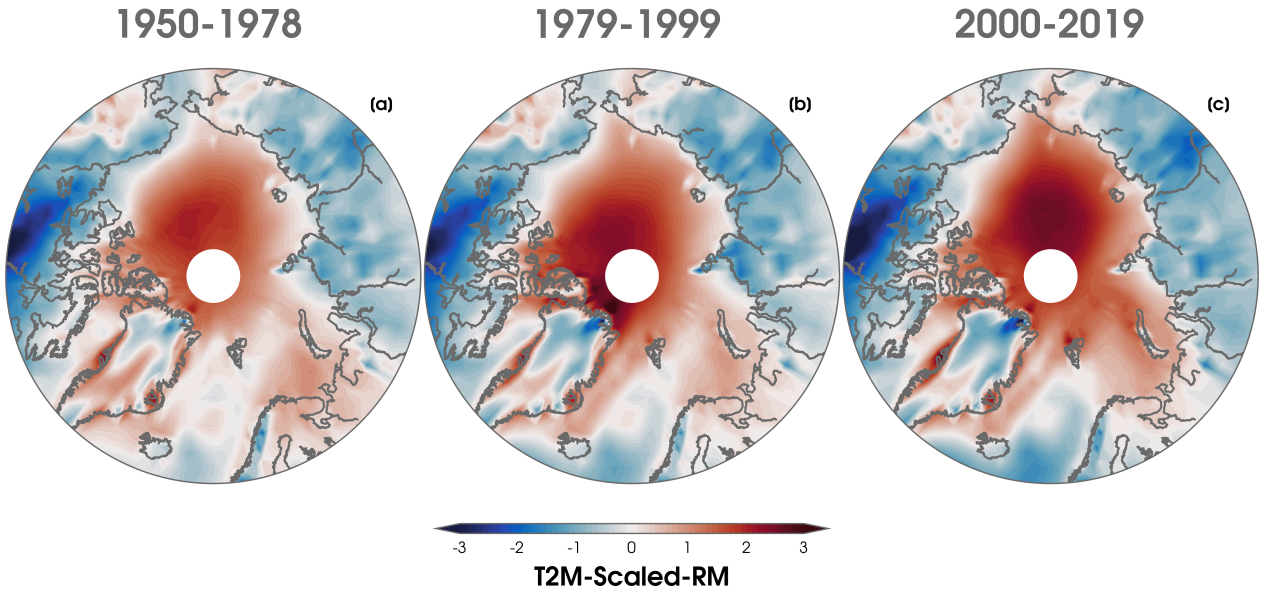

Figure S10. (a) Composite of ERA5-BE T2M with the annual mean of the Arctic first removed separately in each year from 1950 to 1978 from every grid point. The composite is rescaled by subtracting the mean of the training data and dividing by the standard deviation of the training data (with their annual mean also first removed; RM). The training mean and standard deviation are calculated at every grid point by considering all years from 1950 to 2019. (b) As in (a), but for a composite over 1979 to 1999. (c) As in (a), but for a composite over 2000 to 2019.

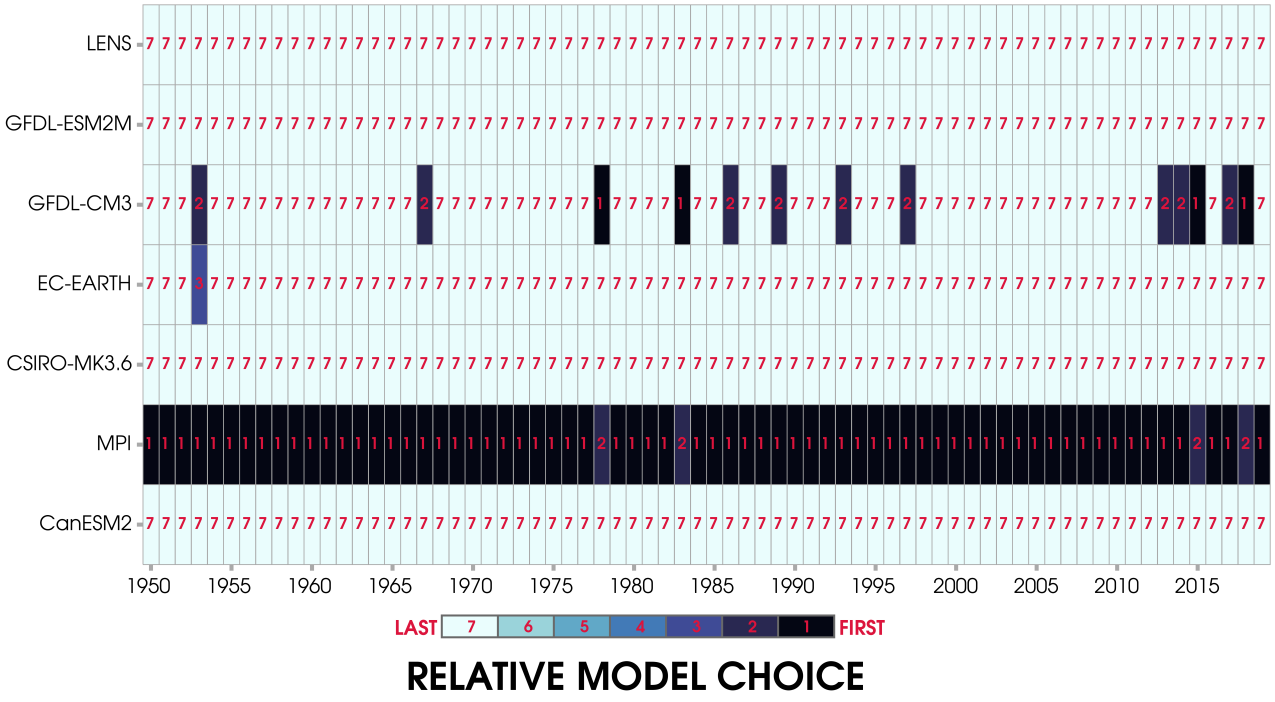

Figure S11. Ranking the order of the ANN confidence values (after a softmax operator) for each GCM class after inputting an annual mean map of T2M from ERA5-BE over the period of 1950 to 2019. For every T2M map, the annual mean of the Arctic is first removed separately in each year from every grid point. A value of 1 indicates that the GCM received the highest confidence (i.e., winning predicted category) for each yearly T2M map. If the confidence value of the ANN output is lower than random chance (≈1/7), the ranking is then set to 7.

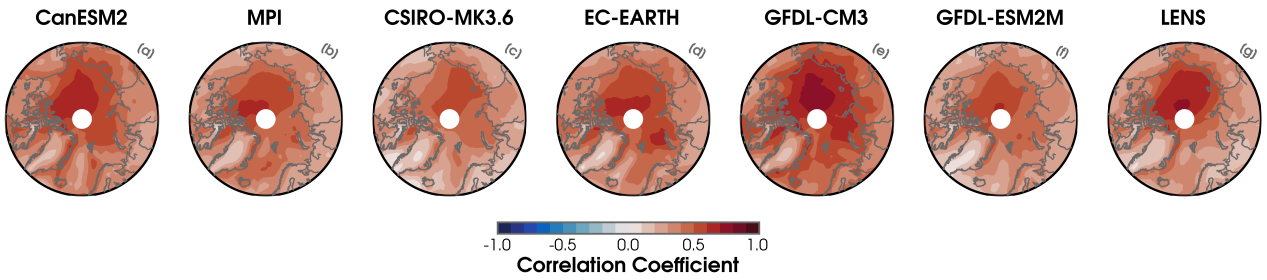

Figure S12. (a) Spatial pattern of temporal correlation coefficients (interval of 0.1) of annual mean T2M between ERA5-BE and CanESM2 from 1950 to 2019. Correlations are first calculated per each ensemble separately and then averaged across ensemble members. (b-g), As in (a), but for MPI, CSIRO-MK3.6, EC-EARTH, GFDL-CM3, GFDL-ESM2M, and LENS, respectively.

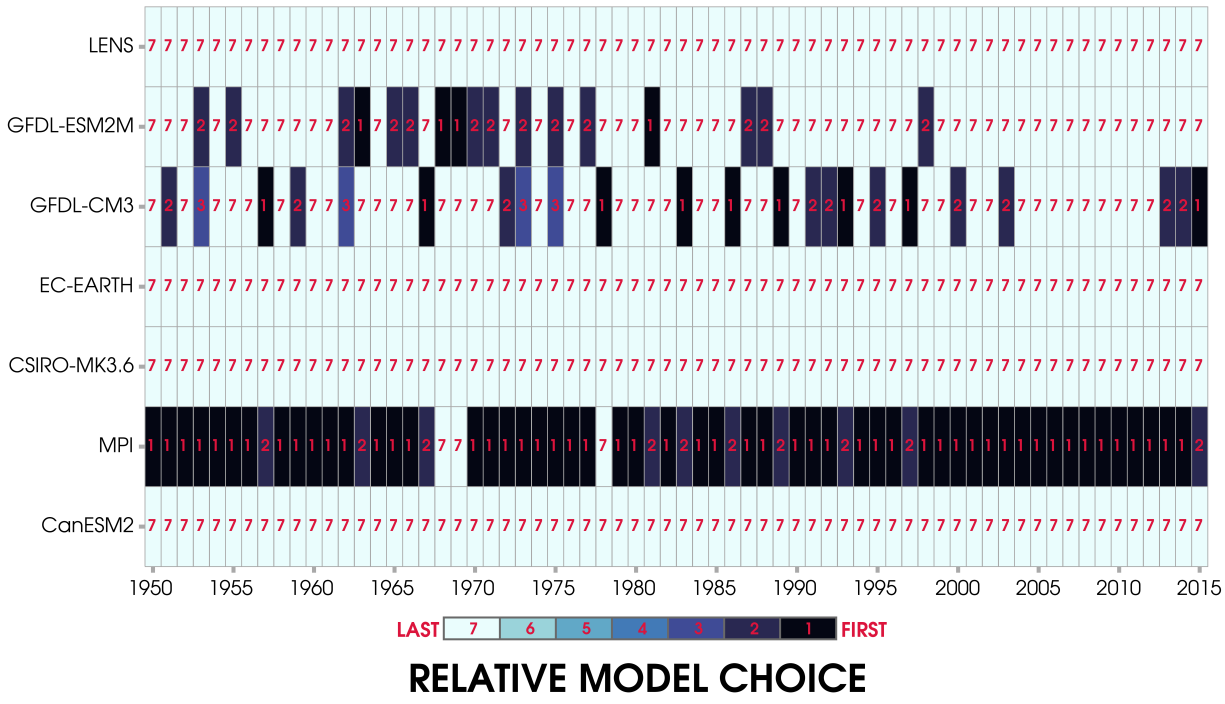

Figure S13. Ranking the order of the ANN confidence values (after a softmax operator) for each GCM class after inputting an annual mean map of T2M from 20CRv3 over the period of 1950 to 2015. A value of 1 indicates that the GCM received the highest probability (i.e., winning predicted category) for each yearly T2M map. If the confidence value of the ANN output is lower than random chance (≈1/7), the ranking is then set to 7.

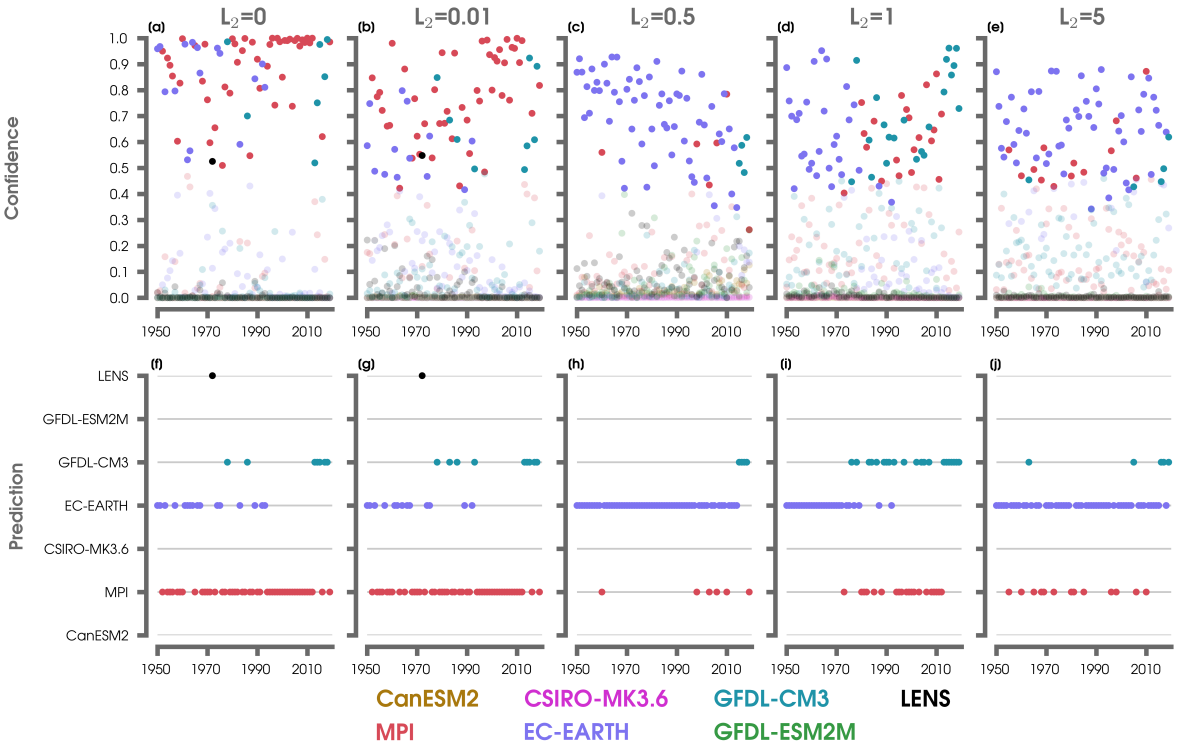

Figure S14. (a) Confidence values for all GCM class outputs (after a softmax operator) for inputs of ERA5-BE from 1950 to 2019 after training an ANN with the same architecture as the main analysis, but using a L_2 regularization value of 0. The marker color is darker for the GCM class with the highest confidence in each year. (b) Same as (a), but using a L_2 regularization value of 0.01. (c) Same as (a), but using a L_2 regularization value of 0.5. (d) Same as (a), but using a L_2 regularization value of 1, (e) Same as (a), but using a L_2 regularization value of 5. (f) As in (a-e), but only showing the winning GCM class label (i.e., highest confidence value) for each T2M map input from ERA5-BE over the period of 1950 to 2019.

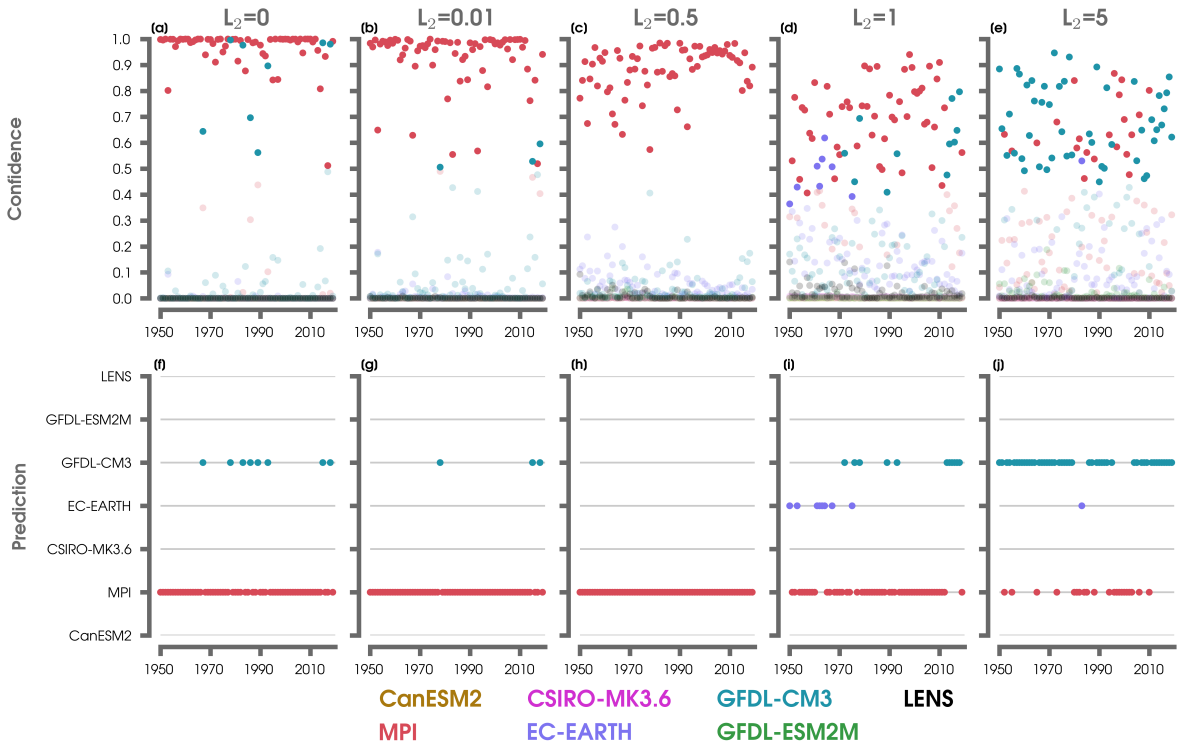

Figure S15. As in Figure S14, but with the annual mean of each T2M map from ERA5-BE first removed at every grid point before inputting into the ANN.

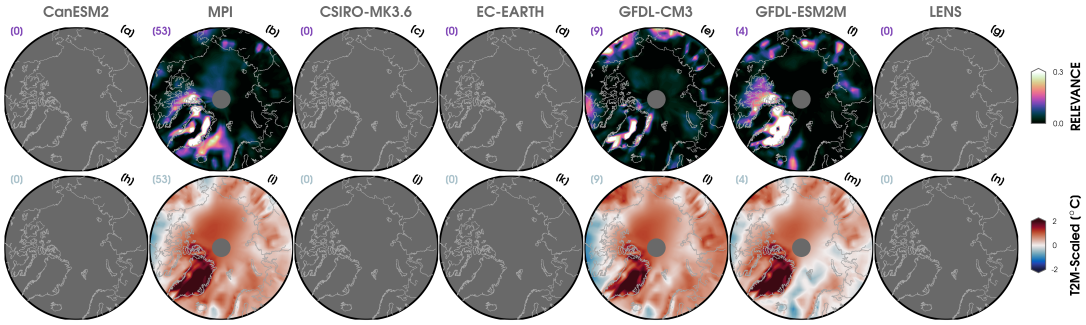

Figure S16. (a-g) Composites of LRP heatmaps for each GCM classification after inputting annual mean maps of T2M from 20CRv3 into the ANN. Higher values indicate greater relevance for the ANN's prediction. (h-n) Composites of T2M from 20CRv3 that are first scaled by the training data mean and training data standard deviation. Maps are then composited according to each predicted GCM class for every year. Maps that are gray indicate that the GCM was never classified, and the number in the upper left-hand corner indicates the number of times the GCM was classified from 1950 to 2015.

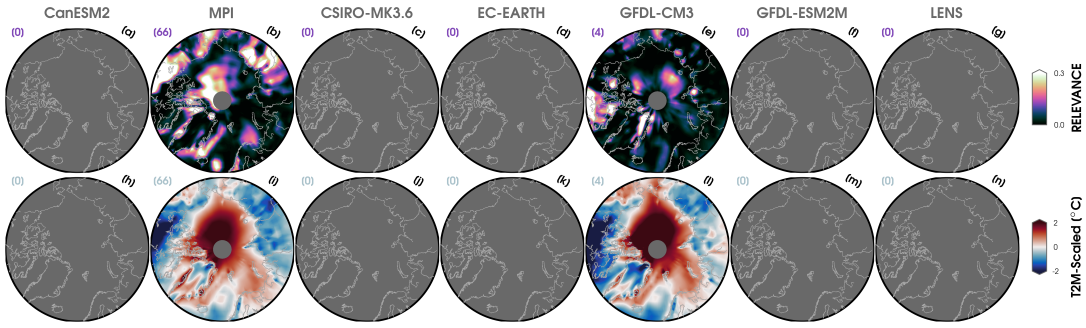

Figure S17. (a-g) Composites of LRP heatmaps for each GCM classification after inputting annual mean maps of T2M from ERA5-BE into the ANN. The annual mean of each map is first removed at every grid point before inputting into the ANN. Higher values indicate greater relevance for the ANN's prediction. (h-n) Composites of T2M (annual mean removed) from ERA5-BE that are first scaled by the training data mean and training data standard deviation. Maps are then composited each year that the GCM was classified, respectively. Maps that are gray indicate that the GCM was never classified, and the number in the upper left-hand corner indicates the number of times the GCM was classified from 1950 to 2019.

References

Bell, B., Hersbach, H., Simmons, A., Berrisford, P., Dahlgren, P., Horányi, A., ... Thépaut, J.-N. (2021). The ERA5 Global Reanalysis: Preliminary Extension to 1950. *Quarterly Journal of the Royal Meteorological Society*. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4174 doi: 10.1002/QJ.4174

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., ... Ting, M. (2020, mar). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, 1–10. Retrieved from http://www.nature.com/articles/s41558-020-0731-2 doi: 10.1038/s41558-020-0731-2

Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010, dec). GLOBAL SURFACE TEMPERATURE CHANGE. *Reviews of Geophysics*, *48*(4), 4004. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2010RG000345 doi: 10.1029/2010RG000345

Hazeleger, W., Severijns, C., Semmler, T., Ştefănescu, S., Yang, S., Wang, X., ... Willén, U. (2010, oct). EC-Earth: A Seamless Earth-System Prediction Approach in Action. *Bulletin of the American Meteorological Society*, *91*(10), 1357–1364. Retrieved from https://journals.ametsoc.org/view/journals/bams/91/10/2010bams2877{_}1.xml doi: 10.1175/2010BAMS2877.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020, may). The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803 doi: 10.1002/qj.3803

Jeffrey, S., Rotstayn, L., Collier, M., Dravitzki, S., Hamalainen, C., Moeseneder, C., ... Syktus, J. (2013). Australia's CMIP5 submission using the CSIRO-Mk3.6 model. *Australian Meteorological and Oceanographic Journal*, *63*(1). doi: 10.22499/2.6301.001

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., ... Vertenstein, M. (2015, aug). 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. Retrieved from http://journals.ametsoc.org/doi/10.1175/BAMS-D-13-00255.1 doi: 10.1175/BAMS-D-13-00255.1

Kirchmeier-Young, M. C., Zwiers, F. W., & Gillett, N. P. (2017, jan). Attribution of Extreme Events in Arctic Sea Ice Extent. *Journal of Climate*, *30*(2), 553–571. Retrieved from https://journals.ametsoc.org/view/journals/clim/30/2/jcli-d-16-0412.1.xml doi: 10.1175/JCLI-D-16-0412.1

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., & Zyss, D. (2019, jun). Improvements in the GISTEMP Uncertainty Model. *Journal of Geophysical Research: Atmospheres*, *124*(12), 6307–6326. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2018JD029522 doi: 10.1029/2018JD029522

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., ... Marotzke, J. (2019, jul). The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2050–2069. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019MS001639 doi: 10.1029/2019MS001639

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., ... Simpson, I. R. (2021, feb). An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set. *Journal of Geophysical Research: Atmospheres*, *126*(3), e2019JD032361. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019JD032361 doi: 10.1029/2019JD032361

NCAR. (2020). *US CLIVAR Multi-Model LE Archive.* Retrieved from https://www.cesm.ucar.edu/projects/community-projects/MMLEA/

Rodgers, K. B., Lin, J., & Frölicher, T. L. (2015, jun). Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences*, *12*(11), 3301–3320. doi: 10.5194/BG-12-3301-2015

Rohde, R. A., & Hausfather, Z. (2020, dec). The Berkeley Earth Land/Ocean Temperature Record. *Earth System Science Data*, *12*(4), 3469–3479. Retrieved from https://essd.copernicus.org/articles/12/3469/2020/ doi: 10.5194/essd-12-3469-2020

Slivinski, L. C., Compo, G. P., Sardeshmukh, P. D., Whitaker, J. S., McColl, C., Allan, R. J., ... Wyszynski, P. (2021, feb). An Evaluation of the Performance of the Twentieth Century Reanalysis Version 3. *Journal of Climate*, *34*(4), 1417–1438. Retrieved from https://journals.ametsoc.org/view/journals/clim/34/4/JCLI-D-20-0505.1.xml doi: 10.1175/JCLI-D-20-0505.1

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., ... Wyszyński, P. (2019, oct). Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 2876–2908.

Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3598 doi: 10.1002/qj.3598

Sun, L., Alexander, M., & Deser, C. (2018, oct). Evolution of the Global Coupled Climate Response to Arctic Sea Ice Loss during 1990–2090 and Its Contribution to Climate Change. *Journal of Climate*, *31*(19), 7823–7843. Retrieved from https://journals.ametsoc.org/view/journals/clim/31/19/jcli-d-18-0134.1.xml doi: 10.1175/JCLI-D-18-0134.1