Detecting climate signals using explainable AI with single-forcing large ensembles

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

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

Key Points:

- Using explainable AI methods with artificial neural networks (ANN) reveals climate patterns in large ensemble simulations
- An ANN trained using a large ensemble simulation without time-evolving aerosols makes more accurate predictions of real world data
- A metric is proposed for quantifying the uncertainty of an ANN visualization method that extracts signals from different external forcings

---

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

Abstract

It remains difficult to disentangle the relative influences of aerosols and greenhouse gases on regional surface temperature trends in the context of global climate change. To address this issue, we use a new collection of initial-condition large ensembles from the Community Earth System Model version 1 that are prescribed with different combinations of industrial aerosol and greenhouse gas forcing. To compare the climate response to these external forcings, we adopt an artificial neural network (ANN) architecture from previous work that predicts the year by training on maps of near-surface temperature. We then utilize layer-wise relevance propagation (LRP) to visualize the regional temperature signals that are important for the ANN's prediction in each climate model experiment. To mask noise when extracting only the most robust climate patterns from LRP, we introduce a simple uncertainty metric that can be adopted to other explainable artificial intelligence (AI) problems. We find that the North Atlantic, Southern Ocean, and Southeast Asia are key regions of importance for the neural network to make its prediction, especially prior to the early-21st century. Notably, we also find that the ANN performs better on inputs of observational data after training on the large ensemble experiment with industrial aerosols held fixed to 1920 levels. This work illustrates the sensitivity of regional temperature signals to changes in aerosol forcing in historical simulations. By using explainable AI methods, we have the opportunity to improve our understanding of (non)linear combinations of anthropogenic forcings in state-of-the-art global climate models.

Plain Language Summary

Using a machine learning method called artificial neural networks, we explore how human-caused climate drivers can affect regional patterns of surface temperature. Here we use a climate model with different combinations of greenhouse gases and industrial aerosols (particles in the atmosphere) to understand their influence on climate change and variability. By employing visualization tools to see how the artificial neural network makes its predictions, we can better recognize how these climate drivers influence global temperature in the past, present, and future. For instance, we find that aerosols emitted in the 20th century and early 21st century have obstructed our view of global warming in some areas of the world, such as over the North Atlantic Ocean. Machine learn-

ing accompanied by new visualization methods have the potential to bring new insights into understanding the effects of global climate change in observations and models.

1 Introductions

Separating human-induced climate forcing from internal variability remains a key challenge for attributing and communicating the impacts of global climate change on regional scales. While state-of-the-art global climate models (GCMs) include anthropogenic (e.g., greenhouse gases and aerosols) and natural (e.g., volcanoes) radiative forcings, it remains difficult to understand their combined interactions and associated effects on climate variability (Stocker et al., 2013). The chaotic noise of the atmosphere (internal variability) also gives rise to additional uncertainties on seasonal to multi-decadal timescales (Deser et al., 2012; Kay et al., 2015). For this reason, it still is difficult to constrain and reduce the uncertainty in Earth's equilibrium climate sensitivity over the last several decades (Sherwood et al., 2020). Moreover, these complex interactions between internal and external climate forcings make it challenging to interpret the physical mechanisms driving regional and even global-scale temperature variability (Stott et al., 2006; Knutti et al., 2010; Maher et al., 2014; D. M. Smith et al., 2016; Medhaug et al., 2017; Haustein et al., 2019; Mankin et al., 2020).

While greenhouse gas forcing dominates the overall climate change signal (net warming), an abundance of anthropogenic aerosols can also influence Earth's surface temperature (net cooling) by scattering or absorbing incoming solar radiation (Bellouin et al., 2020). Further, recent studies have found an influence of anthropogenic aerosols on tropospheric temperatures (e.g., Santer et al., 2019; Mitchell et al., 2020), oceanic internal variability (e.g., Haustein et al., 2019; Dagan et al., 2020; Meehl, Hu, et al., 2020; Qin et al., 2020), the hydrologic cycle (e.g., Marvel et al., 2019; Bonfils et al., 2020), and the large-scale atmospheric circulation (e.g., Allen & Sherwood, 2011; Wang et al., 2020). Meanwhile, less attention has been given to comparing regional climate trends to individual anthropogenic external forcings relative to the influence of internal variability (see examples by Polvani et al., 2011; Santer et al., 2019; Bonfils et al., 2020; Chemke et al., 2020; Deser, Phillips, et al., 2020). For instance, after using an initial-condition large ensemble, Oudar et al. (2018) found a larger role for internal variability than suggested by earlier Coupled Model Intercomparison Project Phase 5 (CMIP5) studies (e.g., D. M. Smith et al., 2016) when attributing the impact of anthropogenic aerosols to the global mean surface temperature trend in the early 21st century.

In addition to the influence of internal variability, the effective radiative forcing from anthropogenic aerosol emissions also remains uncertain over the historical period (Booth et al., 2018; Bellouin et al., 2020; Thorsen et al., 2020). In a novel experiment design, Dittus et al. (2020) assessed the sensitivity of a climate model to a plausible range of historical aerosol forcings. They found better agreement between the observed global mean surface temperature record and an experiment with smaller net aerosol forcing than the standard configuration of the GCM. Consequently, this suggests that temperature signals may be highly sensitive to small changes in aerosols, even when the aerosol forcing in GCMs is constrained to fall within observational estimates (Dittus et al., 2020). This also could be one explanation for the higher climate sensitivities found in CMIP6 models (Flynn & Mauritsen, 2020; Meehl, Senior, et al., 2020).

Recent advances in computational power have led to the development of a growing number of initial-condition large ensembles for assessing climate change and variability (Deser, Lehner, et al., 2020; Deser, 2020). Within a single large ensemble GCM simulation, one can obtain the forced response (i.e., climate signal) by averaging across individual ensemble members that differ by only a small random perturbation error. Thus, if the model is correct, observations of the real world should fall within the ensemble spread in order to reflect both a common forced signal (climate change) and the unpredictable noise of the atmosphere. In other words, the statistical characteristics of internal variability should be similar between the real world and the individual model ensemble members. However, although numerous statistical methods have been proposed to further extract the forced response from internal variability (e.g., Hegerl et al., 1996; Deser et al., 2016; Barnes et al., 2019; Santer et al., 2019; Sippel et al., 2019; Barnes et al., 2020; Sippel et al., 2020; Wills, Battisti, et al., 2020), the problem of climate pattern attribution still remains difficult (Wills, Sippel, & Barnes, 2020).

To improve our understanding of the forced signals from individual anthropogenic climate drivers amidst the noise of internal variability, we implement a method of explainable artificial intelligence (XAI) using data from a novel set of single-forcing large ensemble experiments. The adoption of machine learning applications for geoscience issues continues to rapidly grow (Ebert-Uphoff et al., 2019; McGovern et al., 2019; Rasu et al., 2019; Boukabara et al., 2020; Toms et al., 2020; Watson-Parris, 2020), especially due to an increasing number of XAI methods (Samek et al., 2017; Montavon et al., 2018; Samek et al., 2020). Recently, machine learning models have been used for diverse applications in mesoscale meteorology (e.g., Gagne et al., 2019; Lagerquist et al., 2020), numerical weather prediction (e.g., Rasp et al., 2020; Weyn et al., 2020), simulating cloud and radiation processes in GCMs (e.g., Rasp et al., 2018), turbulence and convection parameterizations (e.g., Beucler et al., 2019; Zanna & Bolton, 2020), attribution of global climate change (e.g., Barnes et al., 2019; Mansfield et al., 2020; Sippel et al., 2020), and reconstructions of historical temperature trends (Kadow et al., 2020). To explore how machine learning models are making their predictions, we focus on using XAI techniques in order to gain new scientific insights for climate science.

In this study, we use artificial neural networks (ANN) in association with an explainability method called layer-wise relevance propagation (LRP) on data from climate model simulations. By investigating the LRP results, we identify key climate pattern responses that are driven by distinct external forcings, namely, greenhouse gases and industrial aerosols. Finally, we assess the skill of the ANN by training on real word observations and introduce a metric to mask noise in assessing the LRP visualizations.

2 Data and Methods

2.1 Climate Model Simulations

For all climate model data, we use large ensemble simulations performed by the Community Earth System Model version 1 (CESM1; Hurrell et al., 2013) covering 1920 to 2080. CESM1 is a fully coupled GCM and is run with 30 vertical levels and a horizontal resolution of 1°. The atmospheric model is the Community Atmosphere Model version 5 (CAM5; Neale et al., 2012), which is coupled to interactive land, ocean, and sea ice components.

Here, we first analyze the widely-used 40-member large ensemble as described in Kay et al. (2015), which we refer to as "ALL" (for all-forcing). The large number of ensemble members is useful for characterizing atmospheric internal variability (or noise) in the climate system (Maher et al., 2019; Deser, Lehner, et al., 2020). Each of the ensemble members have the same external forcing, but are generated from a small random round-off difference in the atmospheric initial conditions. Historical forcing is imposed from 1920 to 2005, and thereafter Representative Concentration Pathway 8.5 (RCP8.5; Vuuren et al., 2011) is used to simulate a worst-case climate scenario through the end of the 21st century (Peters & Hausfather, 2020). Land use/land cover changes, biomass burning, and stratospheric ozone concentrations also evolve with time in the ALL simulation. Although large uncertainties exist, CESM1's total aerosol effective radiative forcing falls within one standard deviation of observational evidence (Zelinka et al., 2014; Bellouin et al., 2020; Deser, Phillips, et al., 2020). We will return to this last point later in the study.

In addition, we also use a set of two new single-forcing simulations from CESM1 that are both run with 20 ensemble members (Deser, Phillips, et al., 2020). These large ensembles have the same GCM, initialization protocol, and external forcing as ALL, but differ by one time-evolving forcing agent that is withheld per simulation. In particular, greenhouse gas concentrations are held fixed to 1920 levels in one experiment (AER+), and industrial aerosols are held fixed to 1920 levels in another (GHG+). While our notation in this study reflects the dominant external forcing agent per simulation (either greenhouse gases (GHG) or industrial aerosols (AER)), we do note that there are other important climate feedbacks and natural variability included in each experiment (hence, the "+" sign) that may contribute to our interpretation of the ANN results (e.g., Luyssaert et al., 2014; Hawkins et al., 2017; Deng et al., 2020; Lehner et al., 2020; Maher et al., 2020; Milinski et al., 2020). Since we only focus on one GCM (CESM1) with historical and RCP8.5 forcing, we can neglect the impact of any structural model and emissions scenario uncertainties that would arise from using, for instance, CMIP5/6 (Hawkins & Sutton, 2009; Knutti & Sedlacek, 2013; Lehner et al., 2020).

After taking into account the smaller ensemble size of the single-forcing runs, we only consider the first 20 members of ALL. However, this does not affect the skill of the ANN for training and testing data (not shown). We apply a bilinear interpolation to the three sets of large ensembles so that they share a slightly coarser latitude by longitude global grid (1.9° x 2.5°). We only consider fields of monthly near-surface air temperature (TREFHT; °C) to calculate seasonal and annuals means from model output. An overview of the climate model simulations used in this study can be found in Table S1.

2.2 Observations

To understand the effect of training on climate model simulations with different external forcing, we test the ANN on observations using the new National Oceanic and Atmospheric Administration/Cooperative Institute for Research in Environmental Sciences/Department of Energy (NOAA-CIRES-DOE) Twentieth Century Reanalysis (20CR) version 3 (20CRv3; also referred to here as 'observations') (Slivinski et al., 2019). Updates to 20CRv3 include an 80-member ensemble size for confidence estimation, a four-dimensional incremental analysis data assimilation scheme (4DIAU), and a higher resolution (T254) core model (described in Slivinski et al., 2019). These improvements lead to a reduction in biases of near-surface temperature, sea surface temperature, and sea level pressure compared to older versions of 20CR, especially in the early to mid-20th century (Compo et al., 2011; Giese et al., 2016). Further, 20CRv3 was found to be in close agreement with other independently derived reanalysis data sets, including the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-20C and CERA-20C (Slivinski et al., 2019, 2020).

We analyze monthly fields of 2-m air temperatures (°C) from 20CRv3 after interpolating (bilinear) onto a common grid of 1.9° latitude by 2.5° longitude for consistency with the climate model simulations. 20CRv3 was selected for our analysis due to its temporally and spatially complete fields of 2-m temperature that are available globally from 1920 to 2015. Similar results were also obtained from the ANN after evaluating on the ECMWF ERA5 reanalysis (Hersbach et al., 2020) for the more recent 1979 to 2019 period. However, in this study, we focus our attention on 20CRv3 for consistency with the historical climate model output. A summary of the observations can be found in Table S2.

2.3 Neural Network Framework

In this analysis, we adopt a neural network architecture that was first introduced in Barnes et al. (2020) and is further illustrated here in Figure 1. We compare the impact of time-evolving greenhouse gases and industrial aerosols on a classification task of predicting the decade (year) from input maps of temperature. Each unit of the ANN input layer represents one grid point from a 2-m temperature map (13824 units per map with dimensions of 96 latitudes by 144 longitudes), and our output layer represents the probabilities of a particular decade class (e.g., 2000-2009). For the time being, the current method of ANN interpretation we use here (layer-wise relevance propagation; LRP) has only been optimized for neural networks with classification algorithms (Bach et al., 2015; Toms et al., 2020). Since we are interested in how the ANN determines its output (i.e., looking into the black box), rather than the output itself, we frame our problem as a classification task for use of LRP.

Our ANN is set up with two hidden layers that each contain 20 hidden units (relatively shallow). We find that increasing the number of layers does not improve the skill of the model, and this architecture supports the interpretability of the fully connected neural network for scientific discovery. In particular, we apply the Rectified Linear Unit (ReLU; Agarap, 2018) activation function to all hidden layer nodes before the output layer, which is defined as $f(x) = max(0, x)$. ReLU is well equipped for use in LRP visualization, since it tests whether individual neurons have been activated (Toms et al., 2020). We also apply a soft-max function to the output layer, which normalizes the decadal class probabilities so that they add up to one. Both ReLU and soft-max functions are common in ANN classification problems such as ours (e.g., Lecun et al., 2015; Goodfellow et al., 2016; Samek et al., 2020).

Before the maps are fed into the ANN, all training data are standardized by their standard deviation across all ensemble members and years at each grid point. Each ANN is then trained using a randomly selected subset of 80% of the climate model simulation data (16 ensemble members) and tested on the remaining 20% (4 ensemble members). During training, our loss function uses binary cross-entropy/log loss, which acts to penalize the ANN when the prediction is wrong, but the model confidence is still high. The ANN are trained using the Nesterov method (momentum = 0.9) for stochastic gradient descent (SGD; Ruder, 2016) for 500 epochs. While the interpretability results are not sensitive to our selection in hyperparameters, we set our learning rate to 0.01 and a batch size to 32 for each ANN used to generate the following figures.

To overcome the problem of overfitting the input data, we use L_2 ridge regularization (Friedman, 2012). The L_2 parameter is set to 0.01 and applied to the weights of the first hidden layer. L_2 regularization imposes a penalty on the model by adding a coefficient to the loss function that is proportional to the sum of the squares of the feature weights. Thus, L_2 regularization leads to weights that are more smoothly distributed across

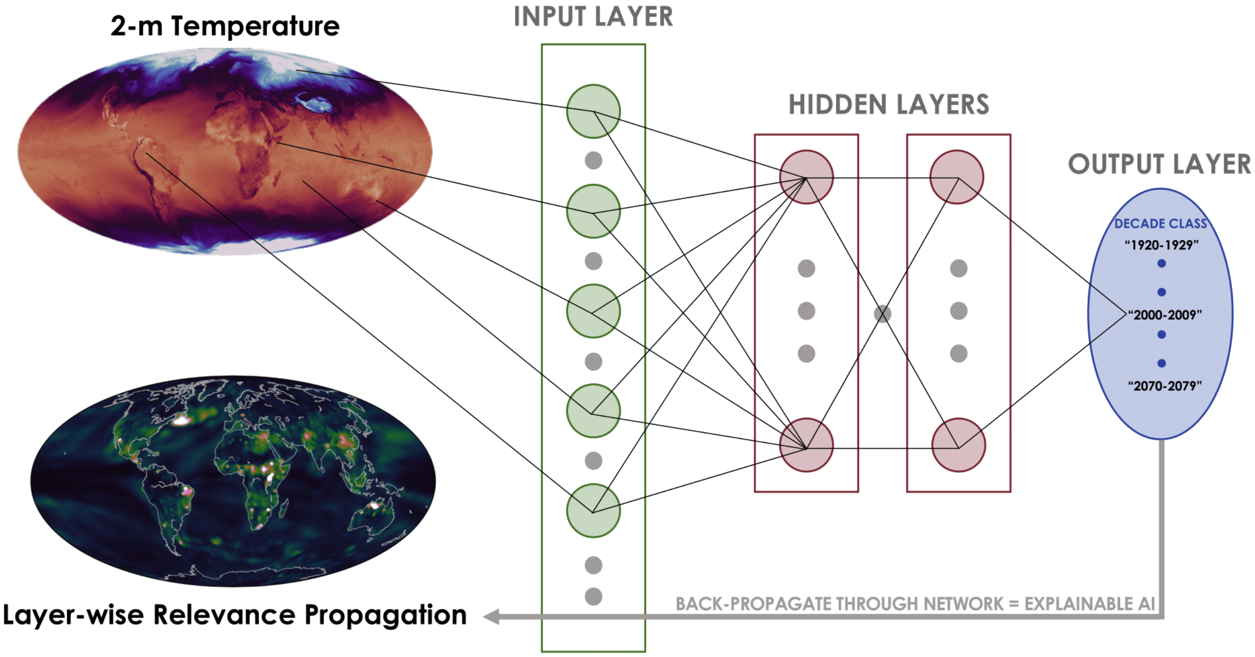

Figure 1. Schematic of the artificial neural network (ANN) used in this study for predicting the decade/year from global maps of 2-m air temperature (input layer). The shallow ANN features two hidden layers that both contain 20 hidden units. The output layer uses fuzzy classification (Zadeh, 1965) to assign each prediction year to the probability of it occurring in a single decade (e.g., within 2000-2009) (Barnes et al., 2020). An example heatmap using layer-wise relevance propagation (LRP; Bach et al., 2015) is also illustrated here. LRP highlights the regions of greater relevance for the ANN to predict the year by propagating an output sample backward through the frozen nodes of the ANN until it reaches the input layer (Toms et al., 2020). Thereby, LRP can attempt to learn the prediction decision by the ANN.

the model and are not as sensitive to outliers in the input data. Importantly, and in relation to standard climate science tools, the inclusion of this parameter accounts for spatial autocorrelation that can exist in the 2-m temperature fields. L_2 also improves the interpretation of the LRP heatmaps for identifying key regions that are relevant for the ANN to make its prediction (e.g., see Figure 3 in Barnes et al., 2020).

Finally, to retrieve the predicted year (output) by the ANN from the maps of 2-m temperature (input), we use a method called fuzzy classification encoding and decoding (Zadeh, 1965; Amo et al., 2004). This occurs during the ANN's output layer (see Barnes et al. (2020)). From this approach, each decade is identified by its central year (e.g., 2005 for 2000 to 2009). The ANN is then designed to assign an input map to the probability of it falling under a particular decade class (encode). Finally, fuzzy classification determines the particular year by computing the weighted sum of the decadal class probabilities (decode). For instance, the year 2008 would be encoded with the probability of 0.7 of belonging to class center 2005 (for 2000 to 2009) and 0.3 of belonging to class center 2015 (for 2010 to 2019). Thus, we can compute the exact year as follows: $0.7{\cdot}2005 + 0.3{\cdot}2015 = 2008$. Additional examples are depicted in Figure 2 of Barnes et al. (2020). Utilizing this setup, we hold onto each particular year while still performing a classification task for use in computing the LRP visualizations. Given our approach using both LRP and fuzzy classification, we do not explore the more typical method of multiple linear regression in this work. However, that approach has been explored in Barnes et al. (2019, 2020) for CMIP temperature and precipitation data.

2.4 Layer-wise Relevance Propagation

The motivation for this work is to reveal the underlying climate patterns that are learned by the ANN from climate model simulations with different combinations of external forcing. As we will show, using XAI tools alongside existing climate science methods have the potential to bring new insights for interpreting projections of climate change in GCMs.

For this work, we use an interpretation method called layer-wise relevance propagation (LRP; Bach et al., 2015; Montavon et al., 2018) for tracing the decisions determined by the ANN. While there are an increasing number of LRP routines, we use a form here (alpha-beta rule) that works well for ReLU networks and is related to Taylor series expansion (Montavon et al., 2017). By propagating information backward until the first layer of the ANN is reached, we learn about the individual input units (features) that are “relevant” to make the ANN’s prediction.

While a detailed overview of using LRP in the geosciences in provided in Toms et al. (2020), we briefly describe the method here: (1) the weights and biases of the ANN are frozen after training, (2) a single prediction output (prior to the soft-max function) is conserved and propagated backward through each node of the ANN based on the frozen weights and biases, (3) the feature relevance is learned until the propagation reaches the input layer, and (4) the final output of LRP retains the original dimensions of the input data by showing the relevance for each pixel (i.e., gridded latitude by longitude points on a map). This process is repeated for every sample. Hence, we are left with a spatial heatmap (unitless) showing the regions of importance for the ANN to determine the decade (see Figure 1).

In this study, our heatmaps are composites of both training and testing sample data. Since our output layer can return multiple probabilities of a 2-m temperature map occurring in a particular decade (fuzzy classification encoding and decoding), we only propagate the output value with the highest probability of belonging to a particular decade. Again, LRP can only propagate one sample backwards at a time. However, previous work has found that this does not affect the interpretation of the LRP output (Barnes et al., 2020). One final note about our use of LRP is that it only returns information that positively contributes to the accuracy of the ANN prediction. Other XAI methods are exploring ways to interpret negative weights and biases that contribute to less confident predictions (e.g., Botari et al., 2020), but that is beyond the scope of this analysis. To interpret the heatmap figures in this study, the higher relevance values indicate greater importance for the ANN’s prediction. Lastly, we introduce a method to mask noise (i.e., relevance) in the LRP output (Section 3.2).

3 Results

3.1 Response to External Forcing

3.1.1 Evolution of simulated and observed trends

We first evaluate the three large ensemble experiments (AER+, GHG+, ALL) using more traditional climate science methods (i.e., trend analysis and signal-to-noise ratios) to understand the spatial patterns of the 2-m temperature response. Figure 2 shows annual maps of temperature trends over four separate 40-year periods for the ensemble mean of each experiment. In the historical period, there is an observed cooling for AER+ (time-evolving aerosols; constant greenhouse gases) for all continental regions and most of the world's oceans (Figures 2a-2b). However, there is a notable statistically significant region of warming over parts of the North Atlantic and Southern Ocean (Figure 2b). These areas of warming may be connected to a strengthened Atlantic Meridional Overturning Circulation (AMOC) (Dagan et al., 2020; Keil et al., 2020; Menary et al., 2020). The global signature of cooling prior to 2000 is associated with an increase in industrial aerosol emissions. Trends in aerosol optical depth are driven by an increase in emissions over Southeast Asia, North America, and Europe in the first half of the 20th century (see Figure 2 in Deser, Phillips, et al., 2020). However, a decrease in aerosol optical depth is observed in North America and Europe closer to present-day with the largest aerosol forcing remaining over Southeast Asia. As industrial aerosols are reduced over the 21st century, there is a net warming trend globally in AER+ through 2080 (Figure 2c-2d). Notably, the temperature trend in the North Atlantic reverses and resembles the "North Atlantic Warming Hole." In agreement with earlier studies (e.g., Dagan et al., 2020), this suggests an important role for aerosols in North Atlantic climate variability. Figure 2e-2h reveals the global warming signature due to the dominant greenhouse gas forcing in GHG+ (time-evolving greenhouse gases; constant aerosols), along with a cooling patch in the North Atlantic. Relative to GHG+, statistically significant warming trends emerge later in ALL (Figure 2i), which is due to its greater aerosol forcing prior to 1960 (net cooling effect). As trends in optical aerosol depth decrease by 2040, there are larger global temperature trends in ALL (Figure 2l) compared to GHG+ (Figure 2h).

We compare the simulated temperature trends with observations by showing the observed (using 20CRv3) 2-m temperature trend (annual mean) for two 40-year periods in Figure S1. However, we note that the observations reflect only one possible realization of internal variability. Therefore, they are not directly comparable with the ensemble mean trends presented in Figure 2. Regardless, we still find some common temperature signatures emerge. By the second half of the 20th century (Figure S1b), we find statistically significant warming across the majority of the tropics and parts of North America. We also find the cooling trend over the North Atlantic detectable in observations for the 1960 to 1999 period.

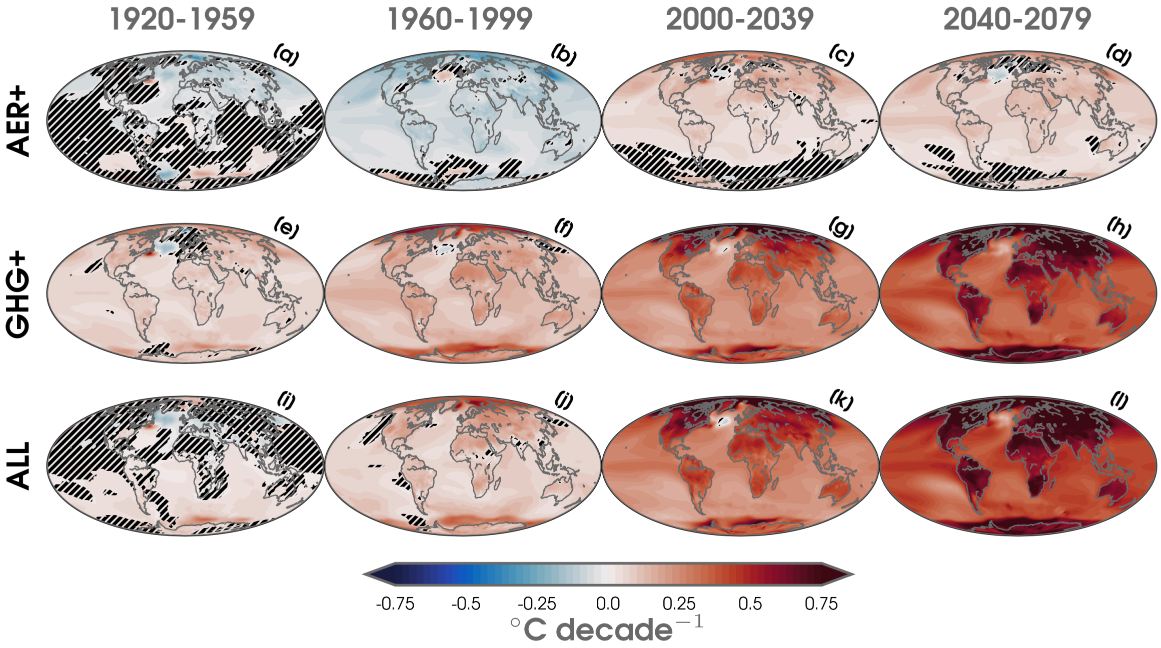

Figure 2. Annual linear least squares trends of 2-m temperature (°C per decade) over 1920 to 1959 (a,e,i), 1960 to 1999 (b,f,j), 2000 to 2039 (c,g,k), and 2040 to 2079 (d,h,l) for the ensemble means of three climate model simulations (AER+; a-d, GHG+; e-h, ALL; i-l). Statistically significant trends are shown with shaded contours at the 95% confidence level following the Mann-Kendall (MK) test (Mann, 1945; Bevan & Kendall, 1971), while those that are not are masked out using black hatch marks.

To understand the timing of emergence of forced climate signals, we compute signal-to-noise (SNR) maps in Figure S2. Here, the SNR is computed as the absolute ensemble mean trend divided by the standard deviation of the individual ensemble member trends for each 40-year period. We observe the highest SNR in the tropics, which is a result of the smaller internal variability in this region. High values of SNR (> 3) emerge as early as the 1920 to 1959 period in GHG+ from the Amazon to the Indian Ocean (Figure S2e), but do not appear until the later half of the 20th century in ALL (Figure S2j-S2k). SNR values are also high in the tropics for the AER+ simulation, but there is little to no forced response (SNR < 1) in the extratropics and polar regions (Figure S2a-S2d). While the global warming signal overwhelms internal variability in GHG+ and ALL beginning in the 2000 to 2039 period, SNR values remain lower ($\sim$1-2) in the subpolar Atlantic.

The effect of aerosols has a consequential role in identifying patterns of forced climate signals. Increases in industrial aerosol loading (e.g., prior to 1960) can mask the timing of emergence of greenhouse gas-induced warming, particularly in the extratropics (Figure S2). Therefore, to fully understand the patterns of responses that are driven by anthropogenic climate drivers, we now turn to our interpretable ANN architecture. One advantage to using our ANN is that we can address potential nonlinearities in regional responses that evolve over time, which would not be captured in the simple trend and SNR analysis.

3.1.2 Predictions by the ANN

Figure 3 shows the predictions by the ANN after separately training and testing on each of the three large ensemble experiments. Here, we use fuzzy classification decoding to show how well the ANN can predict the year from the input maps of 2-m temperature. It is clear that the ANN closely predicts the year on the climate model data, especially after 1980 (blue shading). This is particularly noticeable in ALL (Figure 3c), where the ANN is not able to discern the year in the first half of the 20th century. We also note that the ANN performs even better in AER+ compared to ALL for training and testing, despite the fact that there is no time-evolving greenhouse gas forcing and consequently smaller global mean temperature trends.

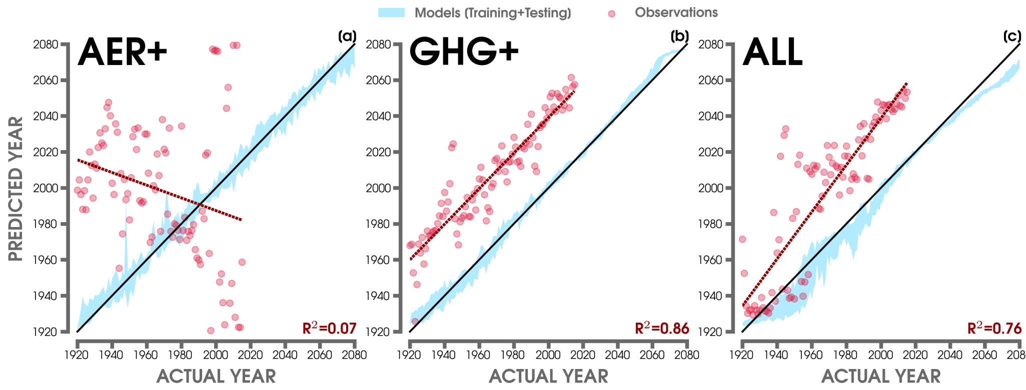

Figure 3. (a) Predictions of the year by the artificial neural network (y-axis) compared to the actual year (x-axis) from global maps of annual 2-m temperature in AER+. (b) Same as (a) but for GHG+. (c) Same as (a) but for ALL. The blue shading highlights the 5th-95th percentiles of predictions from the combined training and testing large ensemble data. The red points show the ANN predictions using 20CRv3 observations. The red dashed line shows the linear least squares fit through the predicted observations in each model, and the associated R^2 is shown in the lower right-hand corner. The 1:1 line (or perfect prediction) is overlaid in black.

To assess the utility of our ANNs that are trained only on climate model data, we test their performance on observations by inputing 2-m temperature maps from 20CRv3. By testing on observational data, we find striking differences between the ANN predictions. The ANN has no skill in predicting the year for observations after training on AER+ (Figure 3a). Since the real world features a large greenhouse gas-induced warming signal, the ANN does not learn regional indicators that are in common with observations. The ANN also performs worse on observations for the ALL model prior to 1980 (Figure 3c). Considering that a forced temperature response has not clearly emerged from the background noise (see Figure S2i-S2j), we infer that this is why the ANN is unable to detect the year.

In contrast, the ANN performs quite well after training on GHG+ for predicting the order of the years in observations (Figure 3b). Since the real world does consist of both direct and indirect effects of greenhouse gases and aerosols, it is somewhat surprising to see that the ANN trained on GHG+ performs better on predicting the year in observations than ALL. In fact, the observations approximately parallel the 1:1 line in GHG+, but are offset by about four decades. This means that the patterns of forced responses

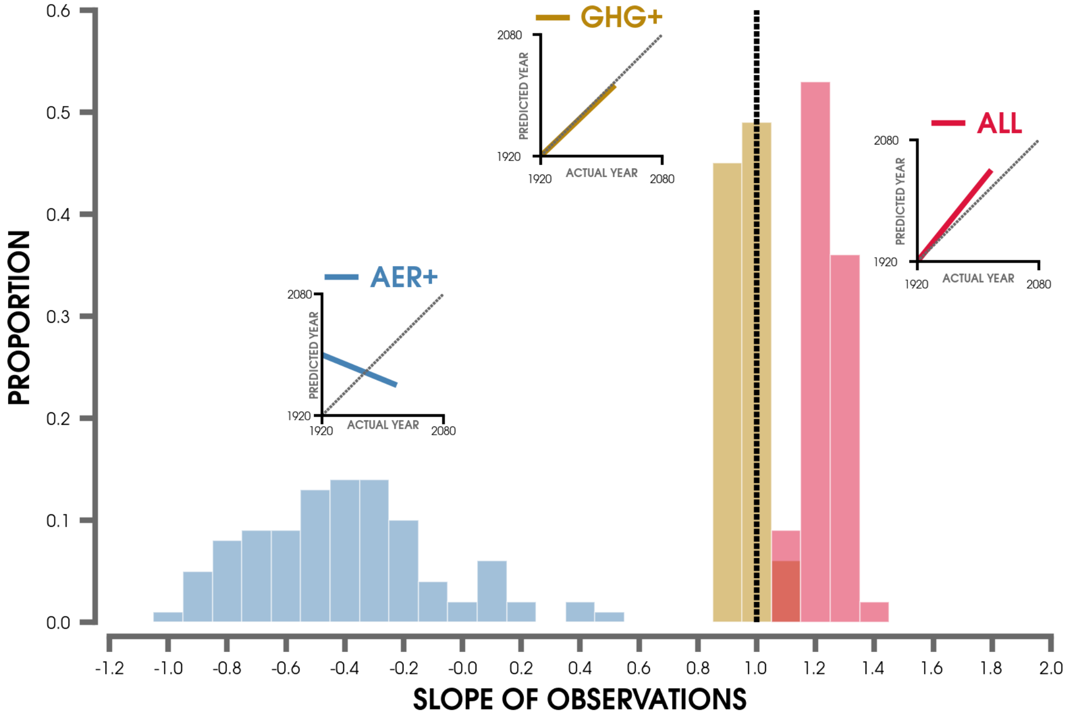

Figure 4. Histogram of the possible slopes of predicted 20CRv3 observations after considering different combinations of training and testing data for each of the AER+ (blue), GHG+ (brown), and ALL (red) artificial neural networks. An example graph of the median slope of predicted observations is shown next to every histogram for the actual year (x-axis) relative to the predicted year (y-axis) (similar to Figure 3). The 1:1 is highlighted by the dashed gray lines.

are similar, but may emerge later in the climate model data compared to observations. This offset could also arise from a difference in Earth's mean temperature that is common between climate models and reanalysis data sets (Hawkins & Sutton, 2016). Therefore, we compare our results in Figure 3 to ANNs trained using input data with the global mean temperature removed from each map (not shown). The results are quite similar. The ANN is still more skillful in predicting the year of the observations on the ANN trained using GHG+. This evidence suggests that the ANN is learning regional temperature signals and not just differences in the global mean temperature to make its predictions, as discussed further in Section 3.3.

We investigate the robustness of our observational predictions in Figure 3 by using 100 unique ANNs trained on different combinations of training and testing data sets (i.e., individual ensemble members). We then test our observational data on each of these 100 iterations and plot a histogram of their predicted (linear) slopes in Figure 4. In agreement with our single trained ANNs in Figure 3, we find that the observations tested on the ANN using GHG+ performs the closest to the 1:1 (or perfect prediction) line with little variability between each iteration. Once again, there is no skill in predicting the year of the observations for the ANN trained on the AER+ simulation. In ALL, the median slope is greater than the 1:1 line likely due to the fact that a forced temperature signal does not emerge until after the middle of the 20th century.

While the results in Figures 3 and 4 show predictions based on maps of annual mean 2-m temperature, we also investigate differences by calculating seasonal means before training and testing the ANN. Figure S3 show the results of predicting the year for boreal winter (January-February-March; JFM) and boreal summer (July-August-September; JAS) in the ANNs using GHG+ and ALL+, respectively. Once more, we find that the predicted year of the observations is improved by testing on the GHG+ model. Notably, we also find a slight improvement in ANN skill (both GHG+ and ALL) on observations from JAS relative to JFM. This may be a result of greater internal variability of 2-m temperatures in the Northern Hemisphere during JFM. In other words, the indicator patterns in common between observations and the climate model data may be weaker in boreal winter compared to summer.

To understand how the ANN is making its predictions, we utilize LRP for evaluating regional climate patterns of interest. In particular, we investigate why the ANN predictions of observations appear more accurate after training on a climate simulation without time-evolving aerosols. As a reminder, the LRP heatmaps indicate areas of “relevance” (or importance) for the ANN to make an accurate prediction. Therefore, greater relevance does not necessarily correspond to the locations of greatest climate forcing. Additionally, the locations of higher relevance may change over time.

3.2 Uncertainty in Layer-wise Relevance Propagation

The LRP algorithm employed here provides output (relevance) for all grid points of every sample. However, it can be difficult to distinguish physically meaningful regions of importance to the ANN, especially for identifying known climate signals. To limit noise in our LRP maps, we compute a threshold (or statistical significance) using a baseline relevance value. In other words, we determine the maximum feature relevance that could be expected from an ANN that is trained on random noise. While other uncertainty metrics for LRP have been proposed (e.g., Bykov et al., 2020; Fabi & Schneider, 2020), our simple method can be employed without modifying the existing ANN architecture or LRP algorithm and takes a common approach applied by climate scientists.

We compute this baseline relevance threshold as follows: (1) we randomly shuffle the individual ensemble member and year dimensions of the ALL input data while keeping the true year fixed (not shuffling), (2) we proceed with training and testing using the same ANN architecture and hyperparameters as Section 2.3, (3) each output sample is then propagated backward into the ANN to compute the relevance map, (4) we repeat steps 1-3 for 500 iterations of the ANN by using unique random initialization seeds and taking different combinations of the training and testing data, and (5) finally, we compute the 95th percentile from the distribution of LRP values at all grid points that are obtained from this procedure. Thus, this bootstrapping-like method determines the distribution of LRP values that could be expected from climate data with no serial autocorrelation or temporal trends from forced signals.

Figure 5 displays a histogram of this distribution of LRP values after 500 unique iterations of the shuffled ANN. We also test our observations (20CRv3) on the ANN trained by the shuffled ensemble from steps (1)-(5). As expected, the ANN cannot predict the year (median linear slope near 0), since it is unable to learn any forced climate signals from the shuffled data. Figure S4 shows a histogram of possible R^2 values from the linear fit of observations compared to the median R^2 of observations trained on either AER+, GHG+, or ALL (Section 3.1.2). We also show an example of a LRP map from a single iteration of the ANN trained on the shuffled ensemble, which highlights the lack of relevant regions for the ANN to make a decision on this synthetic data.

As an additional check of our methodology, we create a “large ensemble” of random numbers drawn from a normal distribution. This large ensemble of random noise has the same dimensions as our real data (20 ensembles, 161 years, 96 by 144 spatial grid points). After repeating steps (2)-(5), we find that the 95th percentile of the random noise LRP is in close agreement with our baseline calculated from Figure 5 (not shown).

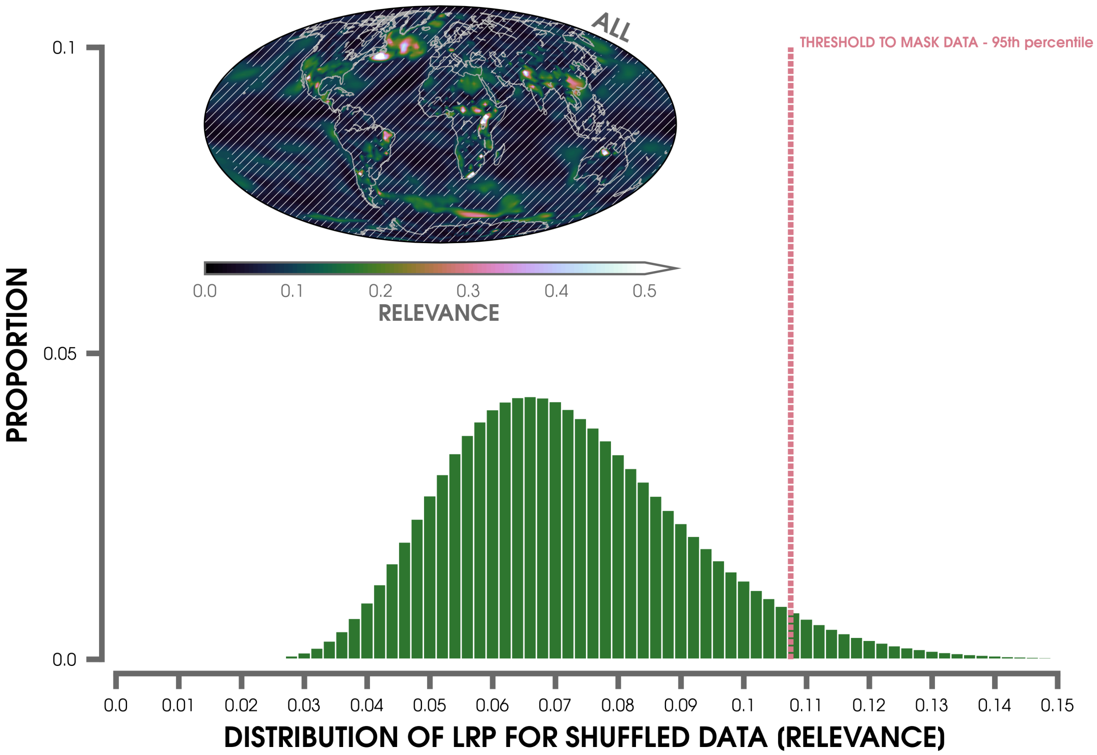

Figure 5. Histogram of the possible relevance values from layer-wise relevance propagation (LRP) after randomly shuffling the ensemble members and years of the input data using the ALL experiment. The 95th percentile LRP threshold is shown by the dashed vertical red line. The LRP composite heatmap is shown for the ALL experiment (training and testing data) using each year from 1920 to 2080 after masking out (white hatch marks) statistically insignificant values (i.e., below the LRP threshold from the histogram; see text for details). Higher LRP values indicate greater relevance for the artificial neural network's prediction.

3.3 Regions of Climate Signal

Figure 6 show the LRP heatmaps for the individual ANN's trained on AER+, GHG+, and ALL input data of annual mean 2-m temperature. Our LRP maps are averaged for every prediction sample (ensemble member) that is accurate to within ±2 years of the actual year (Barnes et al., 2020). In Figure 6, we show the temporal evolution of relevance for the four periods we have considered in this study (e.g., Figure 2). These LRP maps are composites after masking out the relevance below our new uncertainty threshold (see Figure 5). To compare the influence of our LRP uncertainty metric introduced in Section 3.2, we also show the same LRP heatmaps in Figure S5, but without using a mask. Comparing Figure 6 to Figure S5, we now see several climate regions of interest (e.g., North Atlantic and Southeast Asia) that are more clearly distinguishable from the background noise.

The North Atlantic is a key region of relevance between all three large ensembles, but is largest in GHG+ during the 1960 to 2039 period (Figure 6g). The LRP maps also reveal Southeast Asia as an important region for the AER+ and ALL neural networks. The relevance is largest in Southeast Asia for AER+ during the early 20th (Figure 6a) and early 21st centuries (Figure 6c). Again, although the regions of relevance do not directly correspond to surface forcing, we infer that the emissions of anthropogenic aerosols over Southeast Asia and India are important indicators for the ANN to predict the year in the AER+ and ALL large ensembles. We also find that the Southern Ocean is a significant region of relevance for the large ensembles that observe time-evolving greenhouse gases (GHG+ and ALL). Notably, this Southern Ocean signal appears along the Antarctic sea-ice edge. However, in agreement with Barnes et al. (2020), we find that the Arctic is not a region of importance for predicting the year in any of the large ensemble simulations. Despite the effects of Arctic amplification, the lack of relevance to the ANN prediction is likely a result of the large atmospheric internal variability in the high latitudes relative to the tropics (Figure S2).

To compare the differences in LRP maps between seasonal and annual mean input data, we show their relevance composites over 1960 to 2039 in Figure 7. This period is selected due to the greater differences in the timing of emergence of forced signals between the three large ensembles (Section 3.1.1). For the LRP maps based on the annual mean data (Figures 7a,f,k), we observe higher relevance in the North Atlantic for

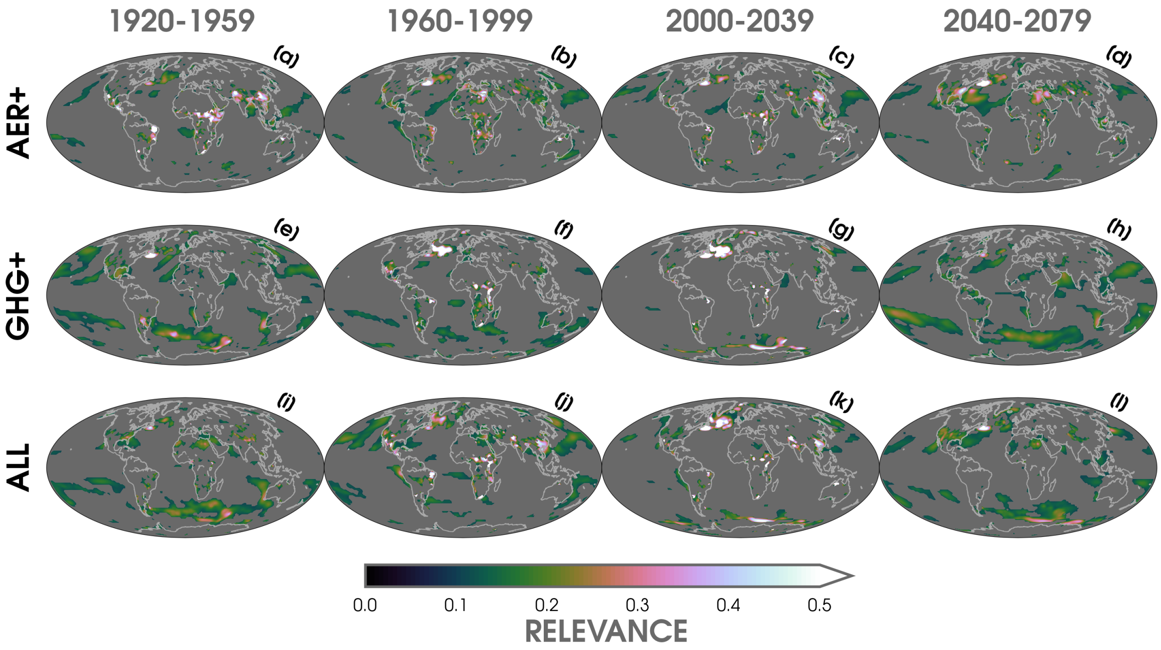

Figure 6. Layer-wise relevance propagation (LRP) composite heatmaps averaged over 1920 to 1959 (a,e,i), 1960 to 1999 (b,f,j), 2000 to 2039 (c,g,k), and 2040 to 2079 (d,h,l) for the three large ensemble experiments (AER+; a-d, GHG+; e-h, ALL; i-l). Higher LRP values indicate greater relevance for the artificial neural network's prediction. Relevance values less than the 95th percentile threshold (see text) have been masked out (gray shading).

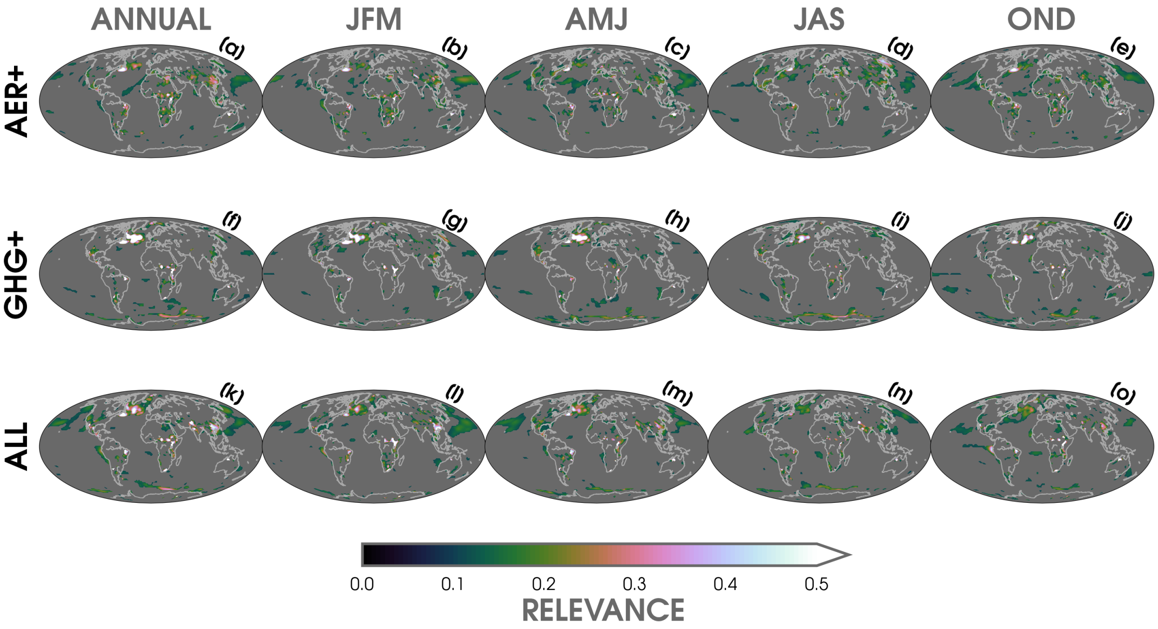

Figure 7. Layer-wise relevance propagation (LRP) heatmaps for ANNs trained separately on annual (a,f,k), January-March (JFM; b,g,l), April-June (AMJ; c,h,m), July-September (JAS; d,i,n), and October-December (OND; e,j,o) input data of 2-m temperature using the three large ensemble experiments (AER+; a-e, GHG+; f-j, ALL; k-o). Every LRP map is composited over the 1960 to 2039 period for the annual data and in each season. Higher LRP values indicate greater relevance for the ANN's prediction. Relevance values less than the 95th percentile threshold (see text) have been masked out (gray shading).

AER+, GHG+, and ALL neural networks. This area of relevance is largest in the ANN trained on GHG+ and is somewhat consistent between seasons. In agreement with Figure 6, this shows that the North Atlantic is a particularly important region for the neural network to predict the year. For AER+ and ALL, we observe a relevance hotspot over India and Southeast Asia, which is distinct during JFM and OND. This is likely due to the local influence of time-evolving aerosols in these climate model simulations, which are absent in the ANN trained on GHG+. Although there are some regional and seasonal differences in Figure 7, the primary climate indicators ("relevance hotspots") remain similar. Thus, we focus on the annual mean input data for the rest of our analysis.

As previously discussed (e.g., in Figure 4), we test the robustness of our results by running 100 unique iterations of each large ensemble ANN for different combinations of training and testing data. Figure S6 shows a composite LRP heatmap that is averaged over all 100 possible iterations of the ANN for only the mean 1960 to 2039 period. The regions of greatest relevance are consistent with Figure 7 and point to the North Atlantic and portions of Southeast Asia (only in AER+ and ALL) as essential to the ANN's predictions. This highlights that the regional signals are robust, even after considering different combinations of individual ensemble members. Moreover, the patterns of higher relevance are also similar when averaging across all climate model years (1920-2080; Figure 5).

Figure 8 shows the distribution of relevances from the 100 unique ANN iterations for the mean relevance value (1960-2039) in five regions (Southeast Asia, India, North Atlantic, Central Africa, and a portion of the Southern Ocean). The small variance in all of the distributions further reinforces the importance of these areas as key climate indicator patterns that are learned by our nonlinear ANN. We find weaker relevance over Southeast Asia (Figure 8a) and India (Figure 8b) for GHG+, which is likely a result of its industrial aerosols being held fixed to 1920 levels. Thus, the temperature signals in these regions (e.g., absence of local cooling due to aerosols) are not as important for the ANN prediction. In contrast, GHG+ observes the greatest relevance in the North Atlantic, while AER+ observes the smallest relevance in this same area (Figure 8c). Interestingly, the North Atlantic distribution for ALL falls between AER+ and GHG+. The relevance signals across Central Africa (Figure 8d) and the Southern Ocean (Figure 8e) are mostly consistent between large ensemble simulations. Nevertheless, we note that there is a slight tendency for the Southern Ocean to be more important for the ANN when there is a larger relative contribution from greenhouse gas forcing (GHG+ and ALL). These LRP results highlight the key importance of the North Atlantic and Southeast Asia for the ANNs to make their predictions.

Finally, to understand where the ANN focuses its attention when making predictions on real world data, Figure 9 shows LRP maps for the observations that are input into the ANNs. Similar to the previous LRP maps of the climate model training and testing data, we find several common relevance regions emerge (e.g., North Atlantic and Southeast Asia). However, recall that the prediction of the years for observations are strikingly different between each large ensemble ANN (Figure 3). In particular, the GHG+ neu-

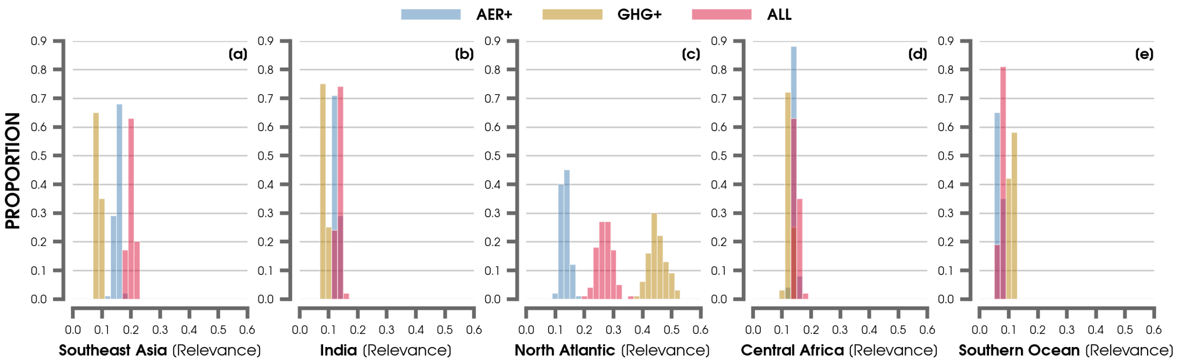

Figure 8. Histograms of mean relevance from layer-wise relevance propagation (LRP) over Southeast Asia (a), India (b), the North Atlantic warming hole region (c), Central Africa (d), and the Southern Ocean (e) for 100 unique iterations of the AER+ (blue), GHG+ (brown), and ALL (red) models. Mean LRP values are averaged over each year from 1960 to 2039.

ral network is more skillful in predicting the order of the years than by ALL. While there is somewhat greater relevance using observations across the North Atlantic and Southern Ocean for the ANN trained on GHG+ (Figure 9c-9d) compared to ALL (Figure 9e-9f), the general patterns between the LRP maps are similar. This indicates that the neural networks are learning different combinations of these regional temperature signals to predict the observations. This also suggests that the GHG+ network may be more skillful by focusing on greenhouse gas-induced responses that are closer to real world data, rather than the temperature patterns which are modulated by industrial aerosol forcing in the AER+ and ALL large ensembles. Hence, the LRP maps reveal how industrial aerosols can either mask or augment detection of greenhouse gas-induced warming signals on local to regional scales.

4 Discussion and Conclusions

Due to complex interactions between internal and external forcings in the climate system, it remains difficult to estimate the local and regional influence of human-induced climate change on surface air temperatures (Schneider & Held, 2001; Deser et al., 2012; McKinnon & Deser, 2018). Our work demonstrates the utility of explainable artificial intelligence (XAI) methods for extracting patterns of climate signals due to varying external forcing, which adds to an existing set of statistical techniques for evaluating signal-

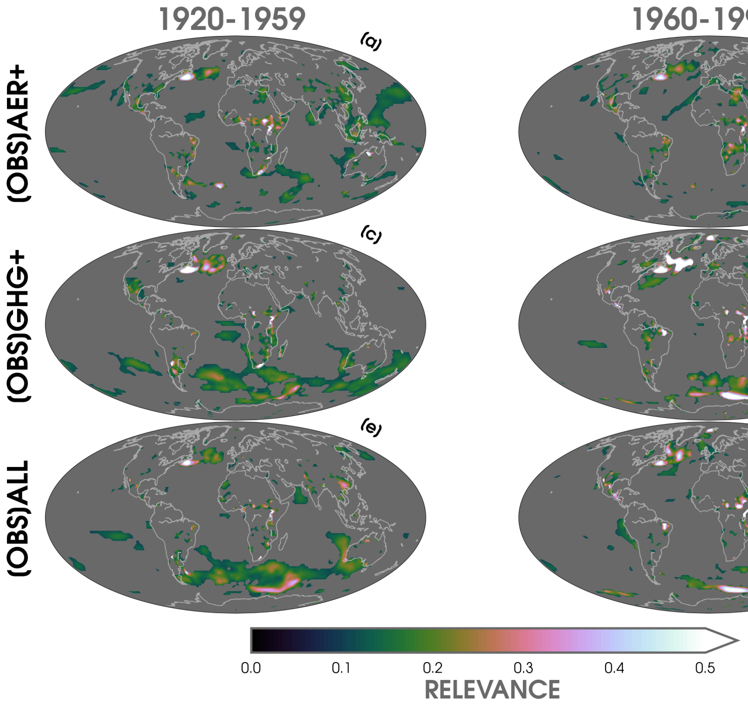

Figure 9. Layer-wise relevance propagation (LRP) composite heatmaps (annual mean) averaged over 1920 to 1959 (a,c,e) and 1960 to 1999 (b,d,f) for observations (OBS) tested separately on each large ensemble ANN (AER+; a-b, GHG+; c-d, ALL; e-f). Higher LRP values indicate greater relevance for the artificial neural network's prediction. Relevance values less than the 95th percentile threshold (see text) have been masked out (gray shading).

to-noise in the Earth system (e.g., Wills, Sippel, & Barnes, 2020). By leveraging a XAI tool as a novel pattern recognition method, we aim to understand how a nonlinear artificial neural network (ANN) makes a prediction by learning regional climate signals.

We build off of ANN results from Barnes et al. (2019, 2020) by investigating the role of different anthropogenic external forcings on temperature patterns relative to the influence of atmospheric internal variability. Using climate model data from a new set of large ensemble experiments, we compare different combinations of human-induced climate drivers (greenhouse gases and industrial aerosols) on forced temperature signals over the 20th and 21st centuries. The large number of ensemble members from one fully-coupled climate model (CESM1) allow us to disentangle forced changes from internal variability. In particular, we use layer-wise relevance propagation (LRP) to investigate how the ANN learns regional climate patterns in order to predict the year from inputs of 2-m air temperatures. Importantly, LRP allows us to investigate the time-evolving relevance (from 1920 to 2080) of input features (maps of 2-m temperature) for the ANN to make an accurate prediction. We also introduce a simple metric to further extract the key relevance regions from the LRP maps. Lastly, we test our nonlinear ANN on observations from a new 20th century atmospheric reanalysis data set (20CRv3) in order to understand how the effect of different external climate forcings impact the prediction of our ANN after testing on real world data.

While efforts are underway to constrain observational uncertainties for the effective radiative forcing of aerosols (e.g., Yoshioka et al., 2019; Bellouin et al., 2020; Bender, 2020; C. Smith et al., 2020), the net influence of aerosols on regional temperature variability remains highly uncertain in historical and future climate model simulations (Bauer et al., 2020; Dittus et al., 2020; Peace et al., 2020). Surprisingly, we found that our ANN trained on a climate model simulation with fixed industrial aerosols (set to 1920 levels; GHG+) made the most accurate predictions after testing on real world observations of temperature. In contrast, the ANN trained on a large ensemble with the most realistic external forcing configuration (ALL) was unable to make an accurate prediction for observations until after the mid-20th century. The LRP maps based on observations indicate that the temperature signal in the North Atlantic is particularly relevant for the more skillful predictions by the ANN trained on GHG+ compared to ALL.

Our ANN results suggests that CESM1 is highly sensitive to combinations between external forcings when simulating the variability and timing of emergence of global climate signals, such as the North Atlantic Warming Hole, compared to observations. While we focus on only one set of single-forcing large ensembles, we recommend that additional experiments are conducted to fully understand the sensitivity of GCMs to aerosol radiative forcing and subsequently simulate realistic temperature trends and variability.

Acknowledgments

This work was supported by NOAA MAPP grant NA19OAR4310289. We would also like to acknowledge the CESM Large Ensemble Community Project and CISL supercomputing resources provided by NCAR (`doi:10.5065/D6RX99HX`). Support for the 20CRv3 dataset is provided by the U.S. Department of Energy (DOE) Office of Science Biological and Environmental Research (BER), by the NOAA CPO, and by the NOAA PSL.

Data Availability Statements

The CESM1 Large Ensemble simulations used in this study are freely available (`https://www.cesm.ucar.edu/projects/community-projects/LENS/data-sets.html`). Monthly 20th Century Reanalysis V3 (20CRv3) data are provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their website at `https://psl.noaa.gov/`. Monthly reanalysis data for ERA5 are also freely available available (`https://climate.copernicus.eu/climate-reanalysis`). Figures and data analysis were completed using Python v3.7.6, Numpy v1.19 (Harris et al., 2020), SciPy v1.4.1 (Virtanen et al., 2020), Matplotlib v3.2.2 (Hunter, 2007), and colormaps provided by cmocean v2.0 (Thyng et al., 2016). Additional Python packages used for development of the ANN and LRP visualizations include Keras/TensorFlow (Abadi et al., 2016) and iNNvestigate (Alber et al., 2019). References for the data sets are provided throughout the study.

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*.

Allen, R. J., & Sherwood, S. C. (2011, may). The impact of natural versus anthropogenic aerosols on atmospheric circulation in the Community Atmosphere Model. *Climate Dynamics*, *36*(9-10), 1959–1978. Retrieved from `https://link.springer.com/article/10.1007/s00382-010-0898-8` doi: 10.1007/s00382-010-0898-8

Amo, A., Montero, J., Biging, G., & Cutello, V. (2004, jul). Fuzzy classification systems. *European Journal of Operational Research*, *156*(2), 495–507. doi: 10.1016/S0377-2217(03)00002-X

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

Bauer, S. E., Tsigaridis, K., Faluvegi, G., Kelley, M., Lo, K. K., Miller, R. L., … Wu, J. (2020, aug). Historical (1850–2014) Aerosol Evolution and Role on Climate Forcing Using the GISS ModelE2.1 Contribution to CMIP6. *Journal of Advances in Modeling Earth Systems*, *12*(8). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019MS001978` doi: 10.1029/2019MS001978

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., … Stevens, B. (2020, mar). *Bounding Global Aerosol Radiative Forcing of Climate*

Change (Vol. 58) (No. 1). Blackwell Publishing Ltd. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019RG000660` doi: 10.1029/2019RG000660

Bender, F. A. (2020). Aerosol Forcing: Still Uncertain, Still Relevant. *AGU Advances*, *1*(3). doi: 10.1029/2019av000128

Beucler, T., Rasp, S., Pritchard, M., & Gentine, P. (2019, jun). Achieving Conservation of Energy in Neural Network Emulators for Climate Modeling. *arXiv*. Retrieved from `http://arxiv.org/abs/1906.06622`

Bevan, J. M., & Kendall, M. G. (1971). Rank Correlation Methods. *The Statistician*. doi: 10.2307/2986801

Bonfils, C. J., Santer, B. D., Fyfe, J. C., Marvel, K., Phillips, T. J., & Zimmerman, S. R. (2020). Human influence on joint changes in temperature, rainfall and continental aridity. *Nature Climate Change*, *10*(8). doi: 10.1038/s41558-020-0821-1

Booth, B. B., Harris, G. R., Jones, A., Wilcox, L., Hawcroft, M., & Carslaw, K. S. (2018). *Comments on "Rethinking the lower bound on aerosol radiative forcing"* (Vol. 31) (No. 22). doi: 10.1175/JCLI-D-17-0369.1

Botari, T., Hvilshøj, F., Izbicki, R., & de Carvalho, A. C. P. L. F. (2020, sep). MeLIME: Meaningful Local Explanation for Machine Learning Models. *arXiv*. Retrieved from `http://arxiv.org/abs/2009.05818`

Boukabara, S.-A., Krasnopolsky, V., Penny, S. G., Stewart, J. Q., & Mcgovern, A. (2020). Outlook for Exploiting Artificial Intelligence in the Earth and Environmental Sciences *. *Bulletin of the American Meteorological Society*, 1–53. Retrieved from `http://journals.ametsoc.org/bams/article-pdf/doi/10.1175/BAMS-D-20-0031.1/5018772/bamsd200031.pdf` doi: 10.1175/BAMS-D-20-0031.1

Bykov, K., Höhne, M. M. C., Müller, K.-R., Nakajima, S., & Kloft, M. (2020, jun). How Much Can I Trust You? – Quantifying Uncertainties in Explaining Neural Networks. *arXiv*. Retrieved from `http://arxiv.org/abs/2006.09000`

Chemke, R., Zanna, L., & Polvani, L. M. (2020). Identifying a human signal in the North Atlantic warming hole. *Nature Communications*, *11*(1). doi: 10.1038/s41467-020-15285-x

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., . . . Worley, S. J. (2011, jan). *The Twentieth Century Reanalysis Project* (Vol. 137) (No. 654). John Wiley and Sons Ltd. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.776` doi: 10.1002/qj.776

Dagan, G., Stier, P., & Watson-Parris, D. (2020, nov). Aerosol forcing masks and delays the formation of the North-Atlantic warming hole by three decades. *Geophysical Research Letters*, *47*(22). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020GL090778` doi: 10.1029/2020gl090778

Deng, J., Dai, A., & Xu, H. (2020, jan). Nonlinear climate responses to increasing CO2 and anthropogenic aerosols simulated by CESM1. *Journal of Climate*, *33*(1), 281–301. Retrieved from `www.ametsoc.org/PUBSReuseLicenses` doi: 10.1175/JCLI-D-19-0195.1

Deser, C. (2020, nov). Certain uncertainty: The role of internal climate variability in projections of regional climate change and risk management. *Earth's Future*. Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2020EF001854` doi: 10.1029/2020EF001854

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

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012, feb). Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, *38*(3-4), 527–546. Retrieved from `http://link.springer.com/10.1007/s00382-010-0977-x` doi: 10.1007/s00382-010-0977-x

Deser, C., Phillips, A. S., Simpson, I. R., Rosenbloom, N., Coleman, D., Lehner, F., . . . Stevenson, S. (2020, sep). Isolating the Evolving Contributions of Anthropogenic Aerosols and Greenhouse Gases: A New CESM1 Large Ensemble Community Resource. *Journal of Climate*, *33*(18), 7835–7858. Retrieved from `https://doi.org/10.1175/JCLI-D-20-` doi: 10.1175/JCLI-D-20

Deser, C., Terray, L., & Phillips, A. S. (2016). Forced and internal components of winter air temperature trends over North America during the past

50 years: Mechanisms and implications. *Journal of Climate*, *29*(6). doi: 10.1175/JCLI-D-15-0304.1

Dittus, A. J., Hawkins, E., Wilcox, L. J., Sutton, R. T., Smith, C. J., Andrews, M. B., & Forster, P. M. (2020, jul). Sensitivity of Historical Climate Simulations to Uncertain Aerosol Forcing. *Geophysical Research Letters*, *47*(13). Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2019GL085806` doi: 10.1029/2019GL085806

Ebert-Uphoff, I., Samarasinghe, S., & Barnes, E. (2019). Thoughtfully Using Artificial Intelligence in Earth Science. *Eos*, *100*. doi: 10.1029/2019eo135235

Fabi, K., & Schneider, J. (2020, aug). On Feature Relevance Uncertainty: A Monte Carlo Dropout Sampling Approach. *arXiv*. Retrieved from `http://arxiv.org/abs/2008.01468http://dx.doi.org/10.5281/zenodo.3970396` doi: 10.5281/zenodo.3970396

Flynn, C. M., & Mauritsen, T. (2020). On the climate sensitivity and historical warming evolution in recent coupled model ensembles. *Atmospheric Chemistry and Physics*, *20*(13). doi: 10.5194/acp-20-7829-2020

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

Gagne, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019, aug). *Interpretable deep learning for spatial analysis of severe hailstorms* (Vol. 147) (No. 8). American Meteorological Society. doi: 10.1175/MWR-D-18-0316.1

Giese, B. S., Seidel, H. F., Compo, G. P., & Sardeshmukh, P. D. (2016, sep). An ensemble of ocean reanalyses for 1815-2013 with sparse observational input. *Journal of Geophysical Research: Oceans*, *121*(9), 6891–6910. Retrieved from `http://doi.wiley.com/10.1002/2016JC012079` doi: 10.1002/2016JC012079

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

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

Haustein, K., Otto, F. E., Venema, V., Jacobs, P., Cowtan, K., Hausfather, Z., … Schurer, A. P. (2019). A limited role for unforced internal variability in twentieth-century warming. *Journal of Climate*, *32*(16). doi: 10.1175/JCLI-D-18-0555.1

Hawkins, E., Ortega, P., Suckling, E., Schurer, A., Hegerl, G., Jones, P., . . . Van Oldenborgh, G. J. (2017, sep). Estimating changes in global temperature since the preindustrial period. *Bulletin of the American Meteorological Society*, *98*(9), 1841–1856. Retrieved from `http://journals.ametsoc.org/bams/article-pdf/98/9/1841/3747963/bams-d-16-0007{\_}1.pdf` doi: 10.1175/BAMS-D-16-0007.1

Hawkins, E., & Sutton, R. (2009, aug). The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1107. doi: 10.1175/2009BAMS2607.1

Hawkins, E., & Sutton, R. (2016, jun). Connecting climate model projections of global temperature change with the real world. *Bulletin of the American Meteorological Society*, *97*(6), 963–980. Retrieved from `https://journals.ametsoc.org/view/journals/bams/97/6/bams-d-14-00154.1.xml` doi: 10.1175/BAMS-D-14-00154.1

Hegerl, G. C., Von Storch, H., Hasselmann, K., Santer, B. D., Cubasch, U., & Jones, P. D. (1996). Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *Journal of Climate*, *9*(10). doi: 10.1175/1520-0442(1996)009⟨2281:DGGICC⟩2.0.CO;2

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

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

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., . . . Marshall, S. (2013). The community earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, *94*(9). doi: 10.1175/BAMS-D-12-00121.1

Kadow, C., Hall, D. M., & Ulbrich, U. (2020, jun). Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, 1–6. doi: 10.1038/s41561-020

-0582-5

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

Keil, P., Mauritsen, T., Jungclaus, J., Hedemann, C., Olonscheck, D., & Ghosh, R. (2020). Multiple drivers of the North Atlantic warming hole. *Nature Climate Change*, *10*(7). doi: 10.1038/s41558-020-0819-8

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*, *23*(10). doi: 10.1175/2009JCLI3361.1

Knutti, R., & Sedlacek, J. (2013, apr). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Clim. Change*, *3*(4), 369–373. doi: 10.1038/nclimate1716

Lagerquist, R., McGovern, A., Homeyer, C. R., Gagne, D. J., & Smith, T. (2020, jul). Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Monthly Weather Review*, *148*(7), 2837–2861. doi: 10.1175/MWR-D-19-0372.1

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

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E., Brunner, L., … Hawkins, E. (2020). Partitioning climate projection uncertainty with multiple Large Ensembles and CMIP5/6. *Earth System Dynamics Discussions*, 1–28. doi: 10.5194/esd-2019-93

Luyssaert, S., Jammet, M., Stoy, P. C., Estel, S., Pongratz, J., Ceschia, E., … Dolman, A. J. (2014, apr). Land management and land-cover change have impacts of similar magnitude on surface temperature. *Nature Climate Change*, *4*(5), 389–393. Retrieved from `www.nature.com/natureclimatechange` doi: 10.1038/nclimate2196

Maher, N., Gupta, A. S., & England, M. H. (2014). *Drivers of decadal hiatus periods in the 20th and 21st centuries* (Vol. 41) (No. 16). doi: 10.1002/2014GL060527

Maher, N., Lehner, F., & Marotzke, J. (2020, may). Quantifying the role of internal variability in the temperature we expect to observe in the coming decades. *Environmental Research Letters*, *15*(5), 054014. Retrieved from `https://doi.org/10.1088/1748-9326/ab7d02` doi: 10.1088/1748-9326/ab7d02

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

Mankin, J. S., Lehner, F., Coats, S., & McKinnon, K. A. (2020, oct). The Value of Initial Condition Large Ensembles to Robust Adaptation Decision-Making. *Earth's Future*, *8*(10). Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2020EF001610` doi: 10.1029/2020EF001610

Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*. doi: 10.2307/1907187

Mansfield, L. A., Nowack, P. J., Kasoar, M., Everitt, R. G., Collins, W. J., & Voulgarakis, A. (2020). Predicting global patterns of long-term climate change from short-term simulations using machine learning. *npj Climate and Atmospheric Science*. Retrieved from `https://doi.org/10.1038/s41612-020-00148-5` doi: 10.1038/s41612-020-00148-5

Marvel, K., Cook, B. I., Bonfils, C. J., Durack, P. J., Smerdon, J. E., & Williams, A. P. (2019). Twentieth-century hydroclimate changes consistent with human influence. *Nature*, *569*(7754). doi: 10.1038/s41586-019-1149-8

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019, nov). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. Retrieved from `http://journals.ametsoc.org/bams/article-pdf/100/11/2175/4876688/bams-d-18-0195{\_}1.pdf` doi: 10.1175/BAMS-D-18-0195.1

McKinnon, K. A., & Deser, C. (2018). Internal variability and regional climate trends in an observational large ensemble. *Journal of Climate*, *31*(17). doi: 10.1175/JCLI-D-17-0901.1

Medhaug, I., Stolpe, M. B., Fischer, E. M., & Knutti, R. (2017). *Reconciling controversies about the 'global warming hiatus'* (Vol. 545) (No. 7652). doi: 10.1038/nature22315

Meehl, G. A., Hu, A., Castruccio, F., England, M. H., Bates, S. C., Danabasoglu, G., ... Rosenbloom, N. (2020). Atlantic and Pacific tropics connected by mutually interactive decadal-timescale processes. *Nature Geoscience*. Retrieved from `https://doi.org/10.1038/s41561-020-00669-x` doi: 10.1038/s41561-020-00669-x

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J. F., Stouffer, R. J., ... Schlund, M. (2020). *Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models* (Vol. 6) (No. 26). doi: 10.1126/sciadv.aba1981

Menary, M. B., Robson, J., Allan, R. P., Booth, B. B., Cassou, C., Gastineau, G., ... Zhang, R. (2020). Aerosol-Forced AMOC Changes in CMIP6 Historical Simulations. *Geophysical Research Letters*, *47*(14). doi: 10.1029/2020GL088166

Milinski, S., Maher, N., & Olonscheck, D. (2020, oct). How large does a large ensemble need to be? *Earth System Dynamics*, *11*(4), 885–901. Retrieved from `https://esd.copernicus.org/articles/11/885/2020/` doi: 10.5194/esd-11-885-2020

Mitchell, D. M., Eunice Lo, Y. T., Seviour, W. J., Haimberger, L., & Polvani, L. M. (2020, oct). The vertical profile of recent tropical temperature trends: Persistent model biases in the context of internal variability. *Environmental Research Letters*, *15*(10). doi: 10.1088/1748-9326/ab9af7

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

Neale, R. B., Gettelman, A., Park, S., Chen, C.-c., Lauritzen, P. H., Williamson, D. L., . . . Taylor, M. a. (2012). Description of the NCAR Community Atmosphere Model (CAM 5.0). NCAR Technical Notes. *Ncar/Tn-464+Str*.

Oudar, T., Kushner, P. J., Fyfe, J. C., & Sigmond, M. (2018, sep). No Impact of Anthropogenic Aerosols on Early 21st Century Global Temperature Trends in a Large Initial-Condition Ensemble. *Geophysical Research Letters*, *45*(17), 9245–9252. Retrieved from `http://doi.wiley.com/10.1029/2018GL078841` doi: 10.1029/2018GL078841

Peace, A. H., Carslaw, K. S., Lee, L. A., Regayre, L. A., Booth, B. B., Johnson, J. S., & Bernie, D. (2020). Effect of aerosol radiative forcing uncertainty on projected exceedance year of a 1.5 °c global temperature rise. *Environmental Research Letters*, *15*(9). doi: 10.1088/1748-9326/aba20c

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

Polvani, L. M., Waugh, D. W., Correa, G. J., & Son, S. W. (2011, feb). Stratospheric ozone depletion: The main driver of twentieth-century atmospheric circulation changes in the Southern Hemisphere. *Journal of Climate*, *24*(3), 795–812. Retrieved from `http://journals.ametsoc.org/jcli/article-pdf/24/3/795/3979461/2010jcli3772{\_}1.pdf` doi: 10.1175/2010JCLI3772.1

Qin, M., Dai, A., & Hua, W. (2020, oct). Quantifying contributions of internal variability and external forcing to Atlantic multidecadal variability since 1870. *Geophysical Research Letters*, *47*(22). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020GL089504` doi: 10.1029/2020gl089504

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020, feb). *Weatherbench: A benchmark dataset for data-driven weather forecasting* (Vol. 12) (No. 11). American Geophysical Union (AGU). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020MS002203` doi: 10.1029/2020ms002203

Rasp, S., Pritchard, M. S., & Gentine, P. (2018, sep). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(39), 9684–9689. Retrieved from `https://gitlab.com/mspritch/spcam3.0-neural-net/tree/nn` doi:

10.1073/pnas.1810286115

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

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

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K. R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(11). doi: 10.1109/TNNLS.2016.2599820

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2020, mar). Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond. *arXiv*. Retrieved from `http://arxiv.org/abs/2003.07631`

Santer, B. D., Fyfe, J. C., Solomon, S., Painter, J. F., Bonfils, C., Pallotta, G., & Zelinka, M. D. (2019, oct). Quantifying stochastic uncertainty in detection time of human-caused climate signals. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(40), 19821–19827. Retrieved from `www.pnas.org/cgi/doi/10.1073/pnas.1922907117www.pnas.org` doi: 10.1073/pnas.1904586116

Schneider, T., & Held, I. M. (2001). Discriminants of twentieth-century changes in earth surface temperatures. *Journal of Climate*, *14*(3). doi: 10.1175/1520-0442(2001)014⟨0249:LDOTCC⟩2.0.CO;2

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., ... Zelinka, M. D. (2020, dec). An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence. *Reviews of Geophysics*, *58*(4). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019RG000678` doi: 10.1029/2019rg000678

Sippel, S., Meinshausen, N., Fischer, E. M., Székely, E., & Knutti, R. (2020, jan). *Climate change now detectable from any single day of weather at global scale* (Vol. 10) (No. 1). Nature Research. doi: 10.1038/s41558-019-0666-7

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A. G., Fischer, E., & Knutti, R. (2019). Uncovering the forced climate response from a single ensemble member using statistical learning. *Journal of Climate*, *32*(17). doi: 10.1175/JCLI-D-18-0882.1

Slivinski, L. C., Compo, G. P., Sardeshmukh, P. D., Whitaker, J. S., McColl, C., Allan, R. J., … Wyszyński, P. (2020, dec). An evaluation of the performance of the 20th Century Reanalysis version 3. *Journal of Climate*, *-1*(aop), 1–64. Retrieved from `https://journals.ametsoc.org/view/journals/clim/aop/JCLI-D-20-0505.1/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

Smith, C., Kramer, R., Myhre, G., Alterskjr, K., Collins, W., Sima, A., … M. Forster, P. (2020). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, *20*(16). doi: 10.5194/acp-20-9591-2020

Smith, D. M., Booth, B. B., Dunstone, N. J., Eade, R., Hermanson, L., Jones, G. S., … Thompson, V. (2016). Role of volcanic and anthropogenic aerosols in the recent global surface warming slowdown. *Nature Climate Change*, *6*(10). doi: 10.1038/nclimate3058

Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M. M., Allen, S. K., Boschung, J., … Midgley, P. M. (2013). *Climate change 2013 the physical science basis: Working Group I contribution to the fifth assessment report of the intergovernmental panel on climate change* (Vol. 9781107057). doi: 10.1017/CBO9781107415324

Stott, P. A., Mitchell, J. F., Allen, M. R., Delworth, T. L., Gregory, J. M., Meehl, G. A., & Santer, B. D. (2006). Observational constraints on past attributable warming and predictions of future global warming. *Journal of Climate*, *19*(13). doi: 10.1175/JCLI3802.1

Thorsen, T. J., Winker, D. M., & Ferrare, R. A. (2020, oct). Uncertainty in observational estimates of the aerosol direct radiative effect and forcing. *Journal of Climate*, *34*(1), 1–63. Retrieved from `https://journals.ametsoc.org/view/journals/clim/34/1/jcliD191009.xml` doi: 10.1175/jcli-d-19-1009.1

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

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

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, H., Xie, S., Zheng, X., Kosaka, Y., Xu, Y., & Geng, Y. (2020, oct). Dynamics of Southern Hemisphere Atmospheric Circulation Response to Anthropogenic Aerosol Forcing. *Geophysical Research Letters*, *47*(19). Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2020GL089919` doi: 10.1029/2020GL089919

Watson-Parris, D. (2020, aug). Machine learning for weather and climate are worlds apart. *arXiv*. Retrieved from `http://arxiv.org/abs/2008.10679`

Weyn, J. A., Durran, D. R., & Caruana, R. (2020, sep). Improving Data-Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere. *Journal of Advances in Modeling Earth Systems*, *12*(9). Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/`

`2020MS002109` doi: 10.1029/2020MS002109

Wills, R., Battisti, D. S., Armour, K. C., Schneider, T., & Deser, C. (2020, sep). Pattern Recognition Methods to Separate Forced Responses from Internal Variability in Climate Model Ensembles and Observations. *Journal of Climate*, *33*(20), 8693–8719. Retrieved from `https://doi.org/10.1175/JCLI-D-19-` doi: 10.1175/JCLI-D-19-0855.1

Wills, R., Sippel, S., & Barnes, E. A. (2020). Separating forced and unforced components of climate change: The utility of pattern recognition methods in Large Ensembles and observations. *US CLIVAR Variations*, *18*(2). doi: 10.5065/0DSY-WH17

Yoshioka, M., Regayre, L. A., Pringle, K. J., Johnson, J. S., Mann, G. W., Partridge, D. G., ... Carslaw, K. S. (2019). Ensembles of Global Climate Model Variants Designed for the Quantification and Constraint of Uncertainty in Aerosols and Their Radiative Forcing. *Journal of Advances in Modeling Earth Systems*, *11*(11). doi: 10.1029/2019MS001628

Zadeh, L. A. (1965, jun). Fuzzy sets. *Information and Control*, *8*(3), 338–353. doi: 10.1016/S0019-9958(65)90241-X

Zanna, L., & Bolton, T. (2020, sep). Data-Driven Equation Discovery of Ocean Mesoscale Closures. *Geophysical Research Letters*, *47*(17). Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2020GL088376` doi: 10.1029/2020GL088376

Zelinka, M. D., Andrews, T., Forster, P. M., & Taylor, K. E. (2014, jun). Quantifying components of aerosol-cloud-radiation interactions in climate models. *Journal of Geophysical Research: Atmospheres*, *119*(12), 7599–7615. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2014JD021710` doi: 10.1002/2014JD021710