JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2021MS002941

Special Section:
Machine learning application to Earth system modeling

Key Points:

- Neural networks and their explainability tools can be harnessed to identify patterns of forced change within combined fields
- Combined fields of input allow for earlier detection of the emergence of a forced climate response
- Explainable AI techniques can be used to identify patterns that describe the emergence and evolution of forced climate change

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. K. Rader,
jamin.rader@colostate.edu

Citation:

Rader, J. K., Barnes, E. A., Ebert-Uphoff, I., & Anderson, C. (2022). Detection of forced change within combined climate fields using explainable neural networks. *Journal of Advances in Modeling Earth Systems*, *14*, e2021MS002941. https://doi.org/10.1029/2021MS002941

Received 6 DEC 2021
Accepted 15 JUN 2022

Author Contributions:

Conceptualization: Jamin K. Rader, Elizabeth A. Barnes
Formal analysis: Jamin K. Rader
Funding acquisition: Jamin K. Rader, Elizabeth A. Barnes
Investigation: Jamin K. Rader
Methodology: Jamin K. Rader, Elizabeth A. Barnes, Imme Ebert-Uphoff

Detection of Forced Change Within Combined Climate Fields Using Explainable Neural Networks

Jamin K. Rader[1], Elizabeth A. Barnes[1], Imme Ebert-Uphoff[2,3], and Chuck Anderson[4]

[1]Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA, [2]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA, [3]Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA, [4]Department of Computer Science, Colorado State University, Fort Collins, CO, USA

Abstract Assessing forced climate change requires the extraction of the forced signal from the background of climate noise. Traditionally, tools for extracting forced climate change signals have focused on one atmospheric variable at a time, however, using multiple variables can reduce noise and allow for easier detection of the forced response. Following previous work, we train artificial neural networks to predict the year of single- and multi-variable maps from forced climate model simulations. To perform this task, the neural networks learn patterns that allow them to discriminate between maps from different years—that is, the neural networks learn the patterns of the forced signal amidst the shroud of internal variability and climate model disagreement. When presented with combined input fields (multiple seasons, variables, or both), the neural networks are able to detect the signal of forced change earlier than when given single fields alone by utilizing complex, nonlinear relationships between multiple variables and seasons. We use layer-wise relevance propagation, a neural network explainability tool, to identify the multivariate patterns learned by the neural networks that serve as reliable indicators of the forced response. These "indicator patterns" vary in time and between climate models, providing a template for investigating inter-model differences in the time evolution of the forced response. This work demonstrates how neural networks and their explainability tools can be harnessed to identify patterns of the forced signal within combined fields.

Plain Language Summary Using machine learning tools called neural networks, we identify patterns of the changing climate within climate model data. Changes in the climate can be identified earlier when detecting patterns within maps of multiple variables and seasons than for single maps alone. By visualizing the patterns learned by the neural networks, we can identify which regions, variables, and seasons are most important for detecting climate change. These patterns offer insight into how climate change is represented in different climate models, and how the patterns of climate change will evolve over time.

1. Introduction

Changes in the climate system comprise the Earth system's response to anthropogenic external forcings (e.g., greenhouse gas and aerosol emissions), natural external forcings (e.g., variations in the solar cycle, volcanic activity), internal variability (natural variations in the climate due to internal processes), and the interactions between them. Distinguishing which features of climate change are the product of external forcings, rather than a byproduct of internal variability, is critical for mitigation and adaptation science (Field et al., 2014; Maher et al., 2021; Mankin et al., 2020; Sanderson et al., 2018). To identify the forced response to external forcings, changes in the climate are often simplified into "signal" and "noise" components (e.g., Hawkins & Sutton, 2009; Mahony & Cannon, 2018; Scaife & Smith, 2018). The signal of climate change captures all anthropogenic and natural external forcings, which we refer to as the forced signal or forced response in this study. Climate noise, a combination of internal variability (natural variations in the climate due to internal processes) and climate model disagreement in the magnitude of the response, often acts to obscure the forced signal (Santer et al., 2011).

Innovative methods are required to determine which behaviors of the climate are the result of the forced signal and which are the result of climate noise. Decades of research have provided a diverse toolkit for this task (North & Stevens, 1998) which includes linear regression (e.g., Mudelsee, 2019; Santer et al., 1996; Sippel et al., 2020; Solow, 1987), empirical orthogonal functions and linear discriminant analysis (e.g., Santer et al., 2019; Schneider & Held, 2001; Wills et al., 2018, 2020), and linear inverse models (e.g., Solomon & Newman, 2012), to name a

Project Administration: Elizabeth A. Barnes
Software: Jamin K. Rader, Elizabeth A. Barnes, Imme Ebert-Uphoff, Chuck Anderson
Supervision: Elizabeth A. Barnes
Validation: Jamin K. Rader
Visualization: Jamin K. Rader
Writing – original draft: Jamin K. Rader
Writing – review & editing: Jamin K. Rader, Elizabeth A. Barnes, Imme Ebert-Uphoff, Chuck Anderson

few. Recently, neural networks have also entered the fold. Neural networks are machine learning algorithms that are able to detect complex, nonlinear relationships between input and output data (Abiodun et al., 2018). Because neural networks are able to detect highly complex relationships, they are useful for many high dimensional problems and have become prevalent in several atmospheric science research fields, such as weather forecasting (e.g., Lagerquist et al., 2019; Lee et al., 2021; Weyn et al., 2020), climate model parameterizations (e.g., Brenowitz & Bretherton, 2018; Gettelman et al., 2021; Silva et al., 2021), and, most relevant to the focus of this study, detection of a forced climate response (e.g., Barnes et al., 2019, 2020; Labe & Barnes, 2021; Madakumbura et al., 2021). To detect patterns of forced change, Barnes et al. (2020) trained a neural network to predict the year label of maps of annual-mean temperature (or precipitation) from climate model simulations for forced historical and future scenarios. Given that the internal variability in any given year differs between the various climate models, the neural network had to learn patterns of the forced climate response. Using neural network explainability methods, they then visualized the regions that were most reliable indicators for identifying change across the Coupled Model Intercomparison Project (CMIP5) models. Barnes et al. (2020) demonstrated that neural networks, and their explainability methods, are powerful tools for extracting forced patterns from climate data. This neural network method is a natural approach for isolating the forced climate response. While many other methods require assumptions to be made about the time evolution of the forced signal and internal variability within the system, neural networks do not (Barnes et al., 2019). Following Barnes et al. (2020), neural networks have since been used to explore the sensitivity of regional temperature signals to aerosols and greenhouse gases using single-forcing large ensembles, and to detect the signal of extreme precipitation in observational data sets (Labe & Barnes, 2021; Madakumbura et al., 2021).

Though many climate signal detection studies focus on single variables, such as annual-mean temperature or a single season of precipitation (Gaetani et al., 2020; Li et al., 2017; Santer et al., 1996, 2019), there are benefits to studying climate change through a multivariate lens (Bindoff et al., 2013; Bonfils et al., 2020; Mahony & Cannon, 2018). Many variables in our atmosphere are closely interconnected, so when the variables are intelligently selected signals of change within multiple variables may be detected earlier than in single variables alone. For example, departure from natural variability can be seen decades earlier in bivariate maps of summertime temperature and precipitation than in either variable alone (Mahony & Cannon, 2018). Similarly, Fischer and Knutti (2012) found that climate model biases in the signal of relative humidity and temperature are negatively correlated such that climate model simulations of their combined quantity, heat stress, have considerably less spread. Combined variables have also been used to identify the impacts of anthropogenic forcings on climate in observational data sets by identifying the multivariate patterns that enhance the signal of change relative to the underlying noise (e.g., Barnett et al., 2008; Marvel & Bonfils, 2013). Understanding how the patterns of the forced response take shape through multiple atmospheric variables also allows for a deeper understanding of the physics at play, as in Bonfils et al. (2020). They explored the evolution of the climate fingerprint by analyzing the leading combined empirical orthogonal functions of temperature, precipitation, and climate moisture index. This multivariate approach illuminated two cross-variable patterns of change: intensification of wet-dry patterns and meridional shifts in the ITCZ associated with interhemispheric temperature contrasts. Neither pattern can be fully explained by a single variable which highlights the utility of combining variables when identifying patterns of the forced response.

Combining fields can be useful for identifying patterns of forced change that do not reveal themselves in single fields alone, but this added information does not come without its drawbacks. Many variables covary in complex and nonlinear ways, such as sea surface temperature and precipitation (Lu et al., 2015), drought indices (Wu et al., 2017), and snowpack, soil moisture and flood risk (Swain et al., 2020), often requiring complex statistics to isolate these interactions. Identifying nonlinear correlations within climate fields introduces another issue, namely in explaining the complex interplay between fields. These drawbacks highlight the need for methods that are both complex and explainable in multivariate climate analyses.

Providing a method for both nonlinear and multi-variable analysis of the forced response, this study extends the neural-network approach of Barnes et al. (2020) to combined fields of input. Combined fields could mean the same variable for different temporal segments (e.g., seasons), or different geophysical variables, both of which are explored here. For the sake of consistency and comparability, this study largely follows the methodology of Barnes et al. (2020), however there are some departures. We standardize the input fields differently which improves the predictive skill of the neural networks. We also use a slightly simpler neural network architecture with fewer nodes in the hidden layers to reduce the computational expense of training a single neural network, and the results from multiple neural networks, rather than just one, are explored. Barnes et al. (2020) demonstrated the utility of neural network explainability methods, and we use these methods in tandem with a clustering technique to enhance post-hoc explanations of neural network decisions.

Section 2 outlines the climate models and observations analyzed in this study. Section 3 introduces the neural network design, the explainability technique (layer-wise relevance propagation [LRP]), and their applications to detection of the forced climate response. We then apply these methods to global temperature and precipitation over land in Section 4. Here, we investigate the benefits of combining variables and compare the results of the neural network with the classical approach of calculating signal-to-noise (S/N) ratios. In Section 5, we explore the patterns of the forced response for extreme precipitation over the Americas and investigate the applications of LRP to studying the evolution of nonlinear climate patterns across multiple climate models. Finally, Section 6 summarizes the results of this work and its implications for future work in forced change detection.

2. Data

2.1. CMIP6 Climate Models

We use climate model output from the sixth phase of the CMIP6 (Eyring et al., 2016). Specifically we focus on monthly-, seasonal-, and annual-mean fields of 2-m air temperature (K), precipitation rate (kg m^{-2} s^{-1}), and precipitation rate from very wet days (kg m^{-2} s^{-1}), hereafter referred to as temperature, precipitation, and extreme precipitation, respectively. We use the meteorological seasons of December-January-February (DJF), March-April-May (MAM), June-July-August (JJA), and September-October-November (SON) for calculating seasonal-mean fields. Defining seasons in this way allows for the earliest detection of forced change (see Figure S1 in Supporting Information S1 for more details).

Very wet days are defined as days that exceed the 95th percentile of all days with precipitation over a pre-defined baseline period (Donat et al., 2016). This is a popular index for measuring changes in extreme precipitation (Cui et al., 2019; Kim et al., 2020) and is used as an indicator of climate change in the U.S. Global Climate Research Program (USGCRP, 2018). We define the baseline as the 40 years from 1980 to 2019, a period for which daily precipitation data exists in both the climate models and the observations. To remove the instances in which climate models simulate sub-trace daily precipitation totals, we only include days that simulated at least 1 mm of precipitation when calculating the 95th percentile of all days with precipitation (Dai et al., 2007).

The neural networks are trained on CMIP6 climate model data. One ensemble member is selected for each of the 37 CMIP6 climate models analyzed so each climate model is only represented once in the training and testing data. Since daily output is required to calculate very wet days, we are limited to 32 models for extreme precipitation (Figure S3 in Supporting Information S1). We analyze the climate model data from 1920 to 2098 under historical forcing (1920–2014) and the shared socioeconomic pathway 585 (SSP585) scenario (2015–2098). SSP585 represents the highest development pathway within CMIP6 scenarios (O'Neill et al., 2016), combining SSP5 and representative concentration pathway 8.5.

Our neural network methodology requires that all climate model fields have the same shape. To accommodate this we regrid the climate model fields from their native resolutions using the second-order conservative remapping method in the Climate Data Operators package from MPI (Schulzweida, 2019). This regridding step reduces the spatial resolution of the data for most climate models. For temperature and precipitation, the data is regridded to 4° latitude by 4° longitude. We elect to use lower resolution data to reduce the computational expense of training neural networks over global maps of temperature and precipitation. Since the domain for extreme precipitation is smaller than the domain for temperature and precipitation (see the following paragraph), and higher resolution data may better capture regional extreme precipitation patterns, the data for extreme precipitation is regridded to a slightly higher resolution: 1.5° latitude by 1.5° longitude.

Two spatial domains are considered in the results of this paper. For temperature and precipitation, the neural networks are trained on all land north of 60°S. Here, we choose to focus on land grid points because that is where humanity lives and will acutely feel the impacts of changing surface temperatures and precipitation. We also exclude Antarctica where climate models and reanalyses struggle to accurately simulate temperature and precipitation. Each map of temperature and precipitation has 948 unique data points. For extreme precipitation, the neural networks are trained on North and South America (land grid points bounded by 90°N, 55°S, 170°W, and 25°W). Here, we choose to narrow the regional scope to show that neural networks are powerful tools for identifying the forced response even when the spatial domain, and thus the available data, is limited. Each map of extreme precipitation has 2,314 unique data points.

2.2. Observations

While this work largely focuses on the results of neural networks trained and tested on climate model data, we show that neural networks trained on climate model data can be applied to observational data as well. For temperature, we use the Berkeley Earth Surface Temperature data set (Rohde & Hausfather, 2020). This data set provides both a temperature climatology and the anomalies at monthly resolution from 1850 to the present. We added the anomalies to the climatology to reconstruct the absolute temperature (K) at each grid point for all months between 1920 and 2019. Monthly observational precipitation fields are obtained from the NOAA Global Precipitation Climatology Project (GPCP), version 2.3, for 1979 to the present (Adler et al., 2018). Since daily precipitation fields are required to calculate extreme precipitation, and daily GPCP precipitation observations are only available back to October 1996, we elected to calculate observed extreme precipitation using the European Centre for Medium-Range Weather Forecasts' ERA5 global reanalysis (Hersbach et al., 2020) at 6-hr resolution from 1980 to present. All observations are regridded in the same way as the climate model data for each respective variable.

3. Forced Change Detection Framework

3.1. Neural Network Design

To identify indicator patterns of the forced response for combined fields we first develop artificial neural networks that, given maps of CMIP6 climate model output from every simulated year from 1920 to 2098, are tasked to predict the year that is being simulated. The results for neural networks trained on 10 different input vectors are explored in the following two sections. The input vectors include annual-, seasonal-, and monthly-mean data for temperature, precipitation, and temperature and precipitation combined, as well as seasonal-mean maps for extreme precipitation over the Americas. We use this diverse selection of input vectors to compare neural network performance and indicator patterns for single-field and combined-field inputs.

The neural network architecture is illustrated in Figure 1. Each unit of the input layer corresponds to a different grid point in the input fields. For example, a neural network that uses seasonal-mean maps of temperature and precipitation as input (two variables and four seasons for a total of eight maps, 948 grid points per map) would have an input vector with 7,584 units. In all cases, this input layer is followed by two fully connected hidden layers with 10 nodes each. The hidden layers are followed by an output layer that consists of 22 classes, one corresponding to each decade midpoint between 1905 and 2115 (e.g., 1905, 1915, 1925, …, 2115). A softmax function is applied to the outputs to convert them to units of likelihood, where the sum of the output vector is one.

Neural networks with this architecture learn the patterns of forced change well, and more complicated architectures do not substantially improve neural network performance (see Figure S2 in Supporting Information S1). It is also notable that this neural network architecture performs better than multiple linear regression, especially when trained on precipitation, and thus using nonlinear techniques improves our ability to detect the year via patterns of forced change (Figure S2 in Supporting Information S1). This architecture is also widely accessible to most in the climate science community as it can be trained on a personal laptop–highly complex architectures can be prohibitively computationally expensive (Chen et al., 2020). These neural networks were trained on a standard desktop computer with 16 GB of RAM and a 3.1 GHz, 6-core processor. Training a single network took anywhere between 2 and 10 min depending on the size of the input field. More details on the neural network design and hyperparameter tuning can be found in Supporting Information S1.

The neural network is tasked with "predicting the year" rather than "predicting the decade" as the output layer may suggest. To translate between decade midpoints and individual year labels, we use fuzzy encoding (Zadeh, 1965) such that each year can be mapped to one or more neighboring classes with varying degrees of membership (encoded as likelihood). This is different than traditional methods that would map each year to a single decade midpoint. In the traditional case, 2040 and 2049 would be considered to be members of the same class since

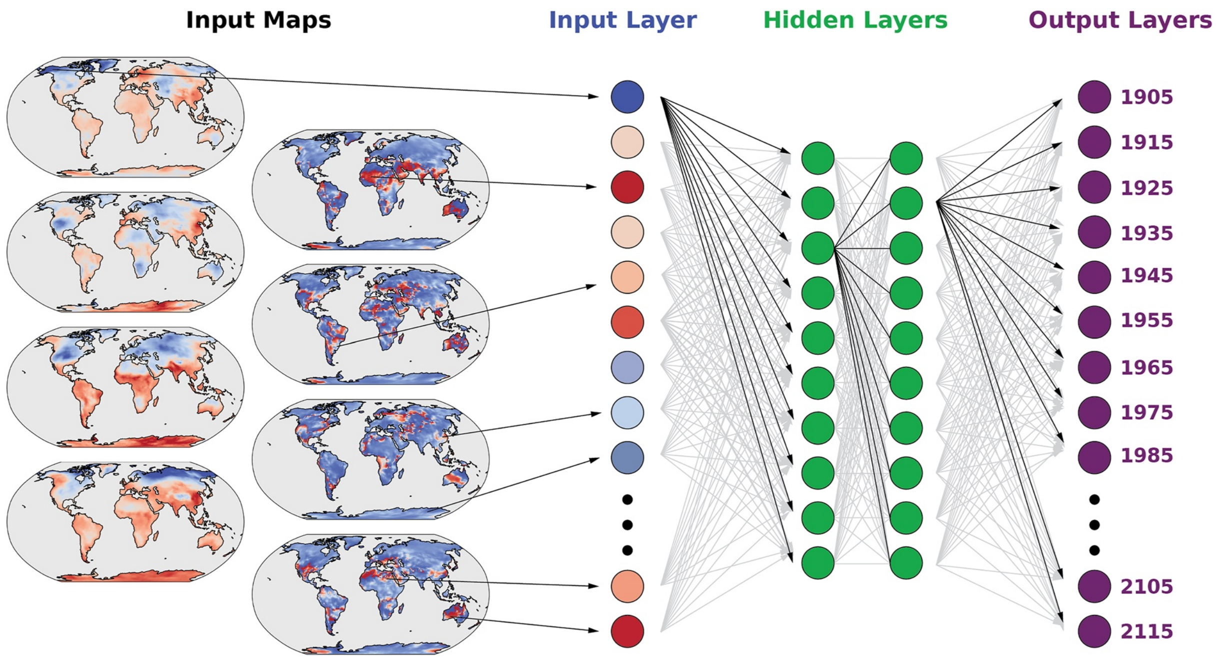

Figure 1. Schematic of the fully connected neural network architecture. Inputs from multiple maps of data are flattened into an input layer vector (size of the input layer ranges from 948 to 22,752). These inputs are fed through two hidden layers with 10 nodes each. The neural network is trained to predict the year that the data came from, outputting the likelihood that the input data came from each decade midpoint between 1905 and 2115. This is then converted to a year via fuzzy classification.

they are in the same decade, and information would be lost as there is no way to distinguish whether the samples come from the beginning or the end of the decade. Using fuzzy encoding, this information of where a sample lies in each decade is retained. We use a triangular membership function (Zadeh, 1965) with a width equal to one decade such that each year has partial membership in one or two neighboring decade classes, and the total membership sums to one. Following this method, any year directly on a decade midpoint has membership in that class only while years that fall between decade midpoints have membership in the two neighboring classes. The year 1925, for example, is mapped to a likelihood of one for the class 1925 and a likelihood of zero in all other classes. The year 2078 is mapped to a likelihood of 0.7 for the 2075 class and a likelihood of 0.3 for the 2085 class. Note that decoding class likelihoods back to their year is simply the decade-weighted sum of the likelihood: $0.7 \times 2075 + 0.3 \times 2085 = 2078$. A visualization of the encoding/decoding process can be found in Figure 2 of Barnes et al. (2020).

3.2. Neural Network Training

For each input vector we train 100 neural networks that differ only in which climate models are randomly split into the training and testing sets. Partitioning so that each climate model's samples are all part of either the training set or the testing set avoids issues with autocorrelation (i.e., near-identical data appearing in both the training and testing sets). One hundred neural networks provide a range of results across multiple combinations of training and testing simulations and offer confidence that the results are consistent across CMIP6 climate models and do not overfit to any one training set. Each neural network is trained over the entire 1920–2098 period on 80% of the climate model simulations, and then tested on the remaining 20%. This leads to a training set of 30 simulations and a testing set of 7 simulations for temperature and precipitation fields, and a training set of 26 simulations and a testing set of 6 simulations for extreme precipitation fields. We train the neural networks using the binary cross-entropy loss (see Barnes et al., 2020) between the predicted class likelihoods and the correct class membership weights, such that the loss function is minimized when the two are equal. Properties of the neural network training process, such as the learning rate and activation functions, can be found in Supporting Information S1.

The neural networks have several hidden nodes which enable them to learn complicated relationships between the input and output data. However, with limited training data, many of these learned relationships will capture

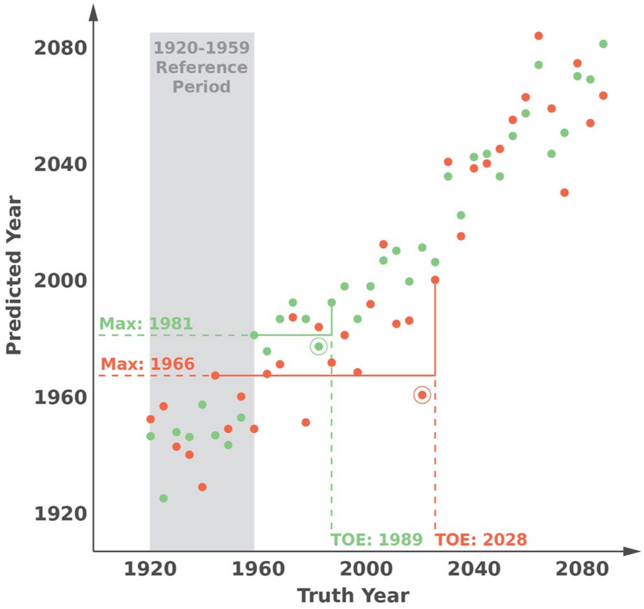

Figure 2. Calculation of time of emergence (TOE). The TOE is defined as the earliest year in which a map, and all subsequent maps, permanently exceed the maximum predicted year from the baseline period (1920–1959). The baseline maximum for each model is indicated by the horizontal lines, the last year that falls below the baseline maximum is circled, and the TOE is indicated by the vertical lines. Sample model 1 (dark red) has a baseline maximum of 1966 and permanently exceeds this threshold in 2028. Sample model 2 (light green) has a baseline maximum of 1981 and permanently exceeds this threshold in 1989. Thus, the TOE for sample models 1 and 2 are estimated as 2028 and 1989, respectively.

patterns of the noise in the training data set which can lead to overfitting (Srivastava et al., 2014). Atmospheric science data is also highly correlated in space and this collinearity can cause complications in the interpretation of the learned weights (Newell & Lee, 1981). Thus, to reduce overfitting and address these issues, we apply ridge regularization (L_2 regularization, see Barnes et al., 2020) to the weights of the first hidden layer. Ridge regularization adds a penalty (called the ridge penalty) to the square of the weights so the solution is penalized for having large weights. Through training, this acts to shrink the largest weights, thus spreading the weight out more evenly across multiple grid points. In our application this results in a more even distribution of importance across regions with strong spatial correlation and improves the performance of the neural networks when given data they were not trained on, namely those models in the testing set (elaborated on in Figure 3, Section 4 of Barnes et al., 2020).

Unlike classical approaches which tune the neural network to reduce the mean squared error (MSE) between the predicted and truth outputs in the testing set (in our case this would be the MSE between the truth and predicted years), we select the ridge penalty that minimizes the time of emergence (TOE) of the forced climate signal (see Section 3.3). Using TOE, rather than MSE, to identify the appropriate ridge penalty ensures that we are encouraging the neural networks to learn the patterns of the forced response across all decades. When a small ridge penalty is used, the neural networks are able to predict the year at the end of the twenty-first century almost perfectly, at the expense of the predictive skill in earlier decades. This results in a later calculation of TOE for the testing set. Slightly increasing the ridge penalty can allow the neural networks to detect the climate change signal slightly earlier (Figure S4 in Supporting Information S1). The ridge penalty used for each input vector can be found in Supporting Information S1. We use the same ridge penalty for all 100 neural networks trained on each input vector.

All input fields (for climate models and observations) are standardized to assist with the training and overall performance of the neural network. We subtracted the 1980–2019 mean at each grid point of the input fields for each climate model independently. This recasts each input field to measure the change relative to the 1980–2019 mean, rather than the raw magnitudes, which improves the predictive skill of the neural networks and is also appropriate for identifying indicator patterns of forced change. Since values for precipitation change are often on the order of 10^{-6}, while the values for temperature change are on the order of 10^{0}, we normalized the data so the inputs to the neural network all have a similar magnitude. To do this, the data from 1980 to 2019 at each grid point for each climate model are detrended using ordinary least squares linear regression. We then take the multi-model mean of the standard deviation of the detrended 1980–2019 data for each grid point. The input fields are then divided by this new field of standard deviations so the inputs are of the same magnitude and fall in a reasonable range for training the neural networks. Since all our observational data sets include the years 1980–2019, we standardize the observations as if they were additional climate models: raw observations are subtracted by their own 1980–2019 mean, and divided by the same multi-model standard deviations that were used to standardize the CMIP6 data.

3.3. Time of Emergence Calculation

The TOE of the forced climate response is the time in which the forced response signal is distinguishable from the background climate by the neural network. Specifically, we define the TOE as the year when the neural network is able to distinguish that year's map from any map over a historical baseline period. In this work, we define this baseline period as 1920–1959 and, under this definition, the earliest possible TOE estimate is 1960. The TOE is estimated for each climate model simulation independently and a schematic of how the TOE is estimated is presented in Figure 2. First, we calculate the maximum of the neural network-predicted years over 1920–1959 for each model, which is referred to as the baseline maximum. We then identify the TOE as the earliest year in which

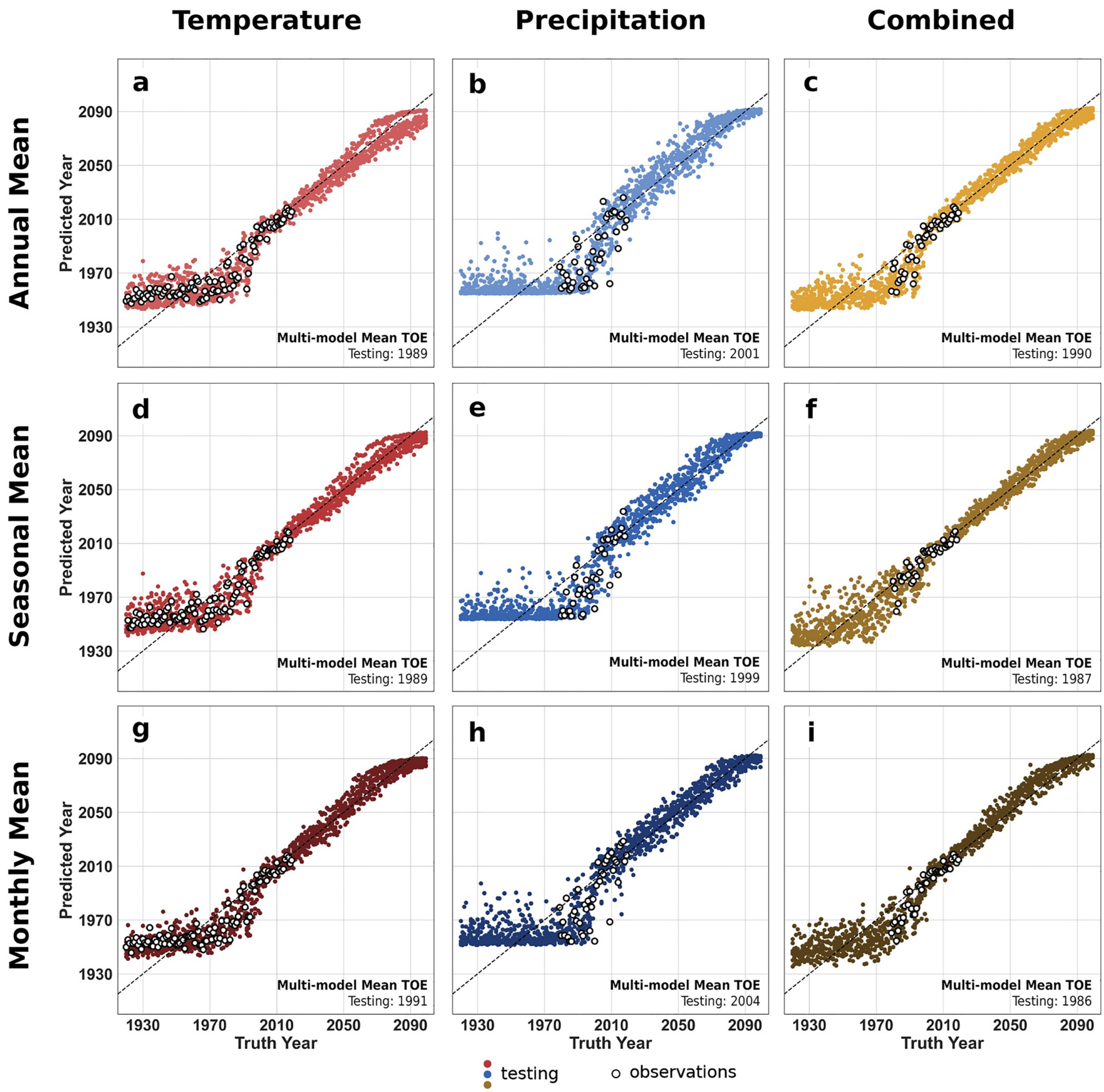

Figure 3. Neural network output for temperature and precipitation. Year predicted by the neural network (y-axis) versus the truth year (x-axis) for temperature (a, d, g), precipitation (b, e, h), and temperature and precipitation combined (c, f, i). Input maps include annual-mean data (a–c), seasonal-mean data (d–f), and monthly-mean data (g–i). Testing data is shown in color and observations are shown in white.

a map, and all subsequent maps, permanently exceed the baseline maximum. In Figure 2, sample model 1 has a baseline maximum of 1966 and permanently exceeds this prediction threshold in 2028. Sample model 2 has a baseline maximum of 1981 and permanently exceeds this threshold in 1989. Thus, the TOE for sample models 1 and 2 are estimated as 2028 and 1989, respectively. In the following sections we present the TOE for the testing set, however TOE estimates are similar for both the training and testing sets.

3.4. Layer-Wise Relevance Propagation

To visualize the patterns learned by the neural network we apply LRP which highlights the regions that were most relevant in the neural network's decision-making process (Bach et al., 2015; Montavon et al., 2019). Toms et al. (2020) discusses in detail how LRP can be used for neural network explainability in the geosciences, though the most relevant details of LRP are described here.

LRP is a neural network explainability method that traces how information flows through the pathways of a trained neural network. The values in a single-sample input vector (in our case, a single year) are passed forward through the neural network. Using the same weights and activations used in the forward pass, LRP then propagates a single-valued output back through the neural network to infer the extent to which each of the values in the input layer contribute to the output (see Figure 2 in Bach et al., 2015). We refer to this quantity as relevance. Through this backpropagation process the output value is conserved such that the sum of all relevance is equal to the output. At first order, relevance can be likened to the product of the regression weights and input map in a linear model. This quantity is natively unitless, but we convert it to a fraction by dividing by the output value. This way, we can consider the relevance of a single pixel in terms of its fractional contribution to the predicted class. Since LRP propagates only a single output value at a time, we propagate relevance only for the decade class with the highest likelihood. While the relevance maps detected by these networks evolve from year to year, this evolution is slow so we find visualizing the highest likelihood decade is sufficient.

There are several LRP decomposition rules which provide different methods of visualizing neural networks (Lapuschkin, 2019; Mamalakis et al., 2021). In our applications we use the $\alpha\beta$-rule which propagates positive relevance (regions that act to increase the class likelihood) and negative relevance (regions that act to decrease the class likelihood) separately. Using the parameters $\alpha = 1$ and $\beta = 0$ we choose to only propagate positive relevance, thus highlighting the regions that added to the likelihood of the selected decade class. We also looked at the relevance maps for $\beta = 1$ and found that propagating negative relevance did not impact the conclusions.

3.5. Signal-to-Noise Ratio Calculation

In Section 4, we compare the LRP relevance maps to maps of S/N ratio, a more conventional method for identifying indicator patterns of the forced response. S/N ratio consists of three distinct components: the forced signal, which is divided by the sum of noise due to internal variability, and noise due to climate model disagreement. A higher S/N ratio indicates that the signal of the forced response within the climate models is very large relative to the underlying noise. We evaluate the S/N ratio for each grid point separately, following the methodology in Hawkins and Sutton (2012). First, we smooth the data from 1920 to 2098 for each climate model using a fourth-order polynomial fit. The signal is defined as the difference between 2090 and 1920 in the smoothed data, while internal variability is defined as the standard deviation of the residuals from the smoothed data, and climate model disagreement is defined as the standard deviation of the signals calculated for all the climate models. S/N ratio is calculated by dividing the climate signal by the 90% confidence interval in the noise: internal variability and climate model disagreement. S/N ratio, and its components, can be seen in Figure S8 in Supporting Information S1.

4. Global Precipitation and Temperature

4.1. Time of Emergence

Across all input vectors of temperature and precipitation, the neural networks are able to learn patterns of the forced response. In the early-to-mid twentieth century, the forced signal is small and undetectable by the neural networks amidst the noise of internal variability and model disagreement, which leads to poor predictive skill (Figure 3). However, as the signal increases in magnitude into the late-twentieth and twenty-first centuries, the neural networks are able to detect the patterns of the forced response and distinguish between maps in different years. These patterns of the forced response detected by the neural networks are generalizable across CMIP6 models, and as a result the neural network has predictive skill for seen data (the training set, see Supporting Information S1) as well as unseen data (the testing set). These behaviors are shown in Figure 3 which presents the predicted years from one trained neural network for each combination of global precipitation and temperature input fields. Across all input vectors, a similar story of the forced signal unfolds. Prior to the TOE, the neural

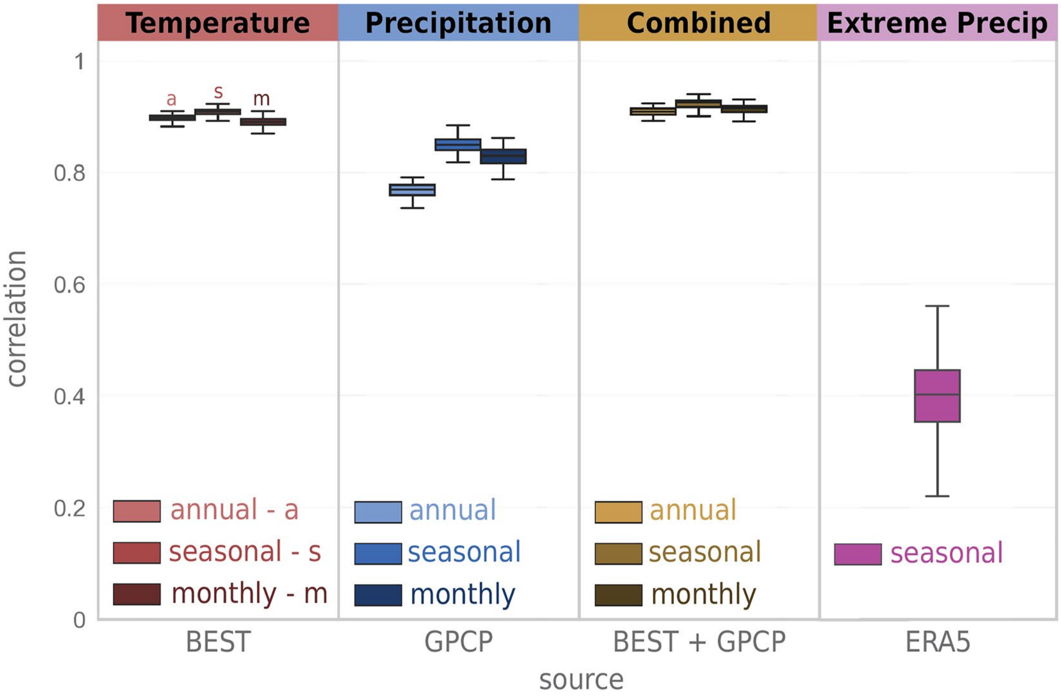

Figure 4. Correlation of actual years with predicted years for observations. Pearson correlations of the actual years with the years predicted by 100 trained neural networks given observations of temperature, precipitation, and extreme precipitation. Correlations were computed for all years beginning in 1980 where observational data exists for all variables. The box plots indicate the first, second, and third quartile statistics, and the whiskers denote 1.5 times the interquartile range, or the minimum/maximum value, whichever is less extreme. Outliers are excluded for clarity, but can be found in Figures S5 and S6 in Supporting Information S1.

network is unable to identify patterns that allow it to accurately predict the year. As a result, the neural network is equally confident (or unconfident) about which year, between 1920 and the TOE, each input came from, so it predicts years right around the middle of the twentieth century. After the TOE, the predicted years tend to follow a 1:1 line with the truth years, indicating that the neural network has identified reliable indicators of change for this period.

Although the neural networks are trained on climate model simulations, their learned patterns can be used to predict the year for observational data as well. When observations are used as input, the predicted years increase with time, just as they do for climate model input (Figure 3). This means that the indicators of change derived by the neural networks trained on climate models simulations are largely consistent with the real world. Pearson correlations (r) of the actual years with the years predicted by each neural network are shown in Figure 4. All correlations are positive, indicating that the years predicted by the neural networks increase with time. These correlations are strongest for temperature and combined observations ($r \approx 0.9$), but still quite high for precipitation ($r \approx 0.8$). Correlations of actual years with predicted years are slightly higher for the combined temperature and precipitation observations than for temperature observations alone (Figure S5 in Supporting Information S1), suggesting that the multivariate indicator patterns derived from climate model data are useful for understanding trends in the present-day climate. Across all variables, the highest observational correlations are found by the neural networks trained on seasonal-mean data. The correlation of actual years with predicted years for precipitation observations are sensitive to the data set of choice, which is expanded on in Section S4 and Figures S5 and S6 in Supporting Information S1.

The average TOEs, calculated from the climate models in the testing sets of all 100 trained neural networks for each input field (Figure 5), reveal that the forced response can be detected earlier in maps of temperature than in maps of precipitation (Figures 5a–5c). When presented with combined fields the neural networks are, in many cases, able to detect the forced signal even earlier than when given single fields alone (Figures 5b and 5f). The TOE is generally earlier for the neural networks trained on seasonal-mean data than for the neural networks

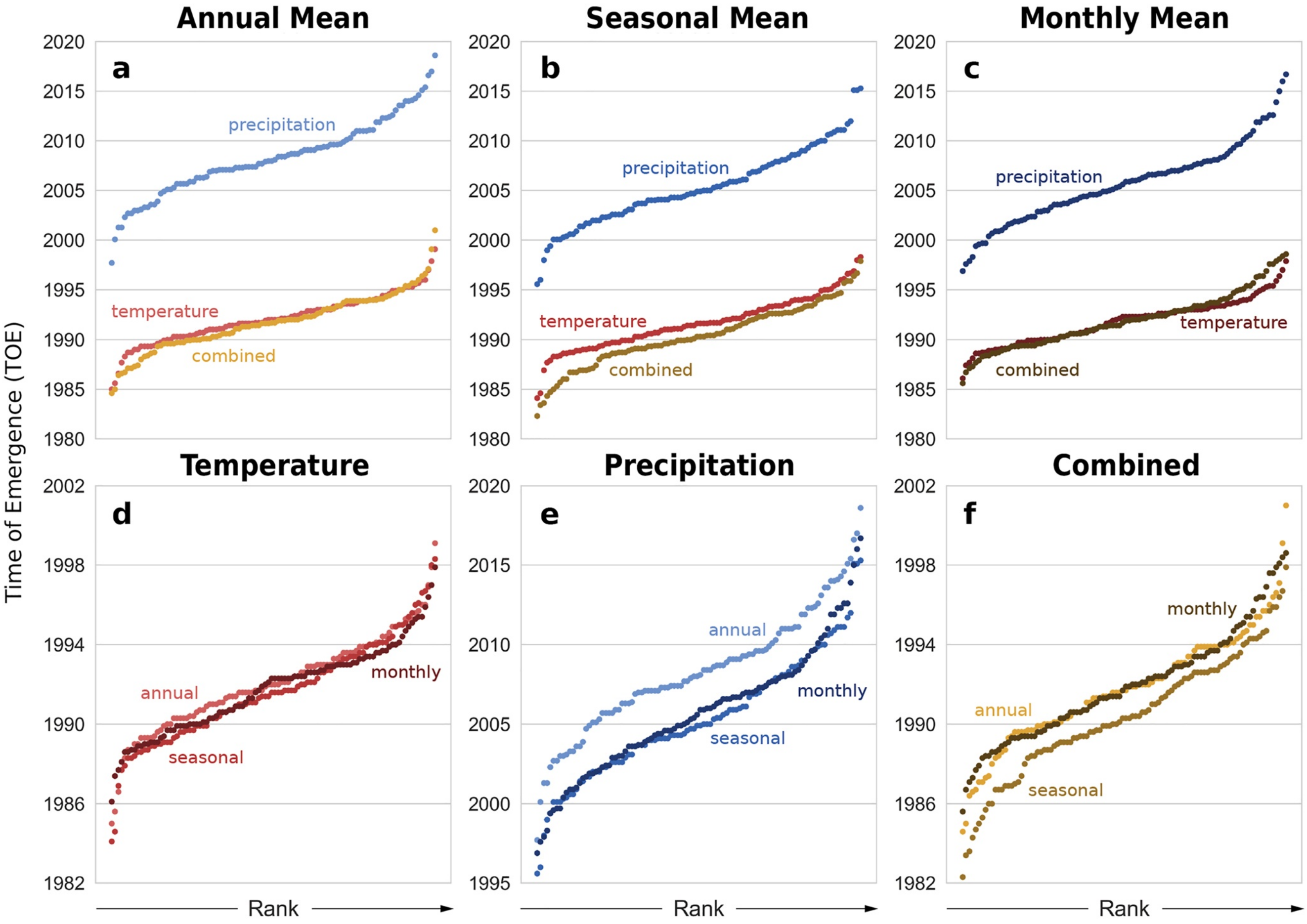

Figure 5. Mean time of emergence (TOE) for each input field. Comparison of the mean time of emergence identified by neural networks trained on annual-mean (a), seasonal-mean (b), and monthly-mean (c) input fields, and neural networks trained on temperature (d), precipitation (e), and temperature and precipitation combined (f). A total of 100 neural networks with different train-test splits were trained for each input field. Each dot represents the mean TOE for all climate models in the testing set for a single trained neural network, ranked from earliest to latest. Note the change in the *y*-axes between panels, and that the TOE results for each set of neural networks appear once in the panels (a–c), and once in panels (d–f).

trained on annual-mean data (Figures 5d–5f). This is most notable for precipitation fields, likely because there are large seasonal precipitation responses muted by taking the annual mean (Tabari & Willems, 2018; Zappa et al., 2015). The TOEs are earlier for temperature and precipitation combined than temperature alone when using seasonal-mean maps (Figure 5b), but are approximately equal when using annual-mean or monthly-mean maps (Figures 5a and 5c), which suggests that precipitation only improves upon the detectability of the forced temperature signal when seasonal-mean fields are used. While annual-mean precipitation may mute seasonal precipitation signals, monthly-mean precipitation is noisy. In this case, seasonal means emerge as the appropriate temporal segments for detecting precipitation change, underlining the importance for the intentional and intelligent selection of neural network inputs.

The neural networks identify the earliest TOEs when trained on seasonal-mean temperature and precipitation combined (Figures 5b and 5f). The TOE results for all 100 seasonal-mean neural networks are summarized in the box plots in Figure S7 in Supporting Information S1. While the improvement in forced response detection is small when precipitation is combined with temperature, it is still notable given that the forced signal of temperature is much clearer than the forced signal of precipitation. We use these variables as an initial example for employing this neural network methodology. We anticipate that more robust results might be found for combinations of variables that have more distinct combined signals, such as humidity and temperature (Fischer & Knutti, 2012).

4.2. Indicator Patterns for Combined Variables

Having shown that the neural networks are able to predict the year given seasonal means of temperature and precipitation (Figures 3 and 5), we now identify and explore the spatial indicator patterns used by the neural networks to make correct predictions. By understanding the neural networks' decision-making process, we can identify which regions act as combined (multi-seasonal and multi-variable) indicators of forced change amidst a background of internal variability and climate model disagreement. To identify these indicator patterns, we apply LRP to all climate model samples in the training and testing sets from the year 2090 for the seasonal-mean combined neural networks. Averaging the LRP results for each season and variable, we highlight the regions that have the highest mean relevance across the 37 CMIP6 climate models and 100 trained neural networks. The relevance maps for temperature (precipitation) are shown in Figures 6a–6d (7a–7d) and indicate the importance of each region in the neural networks' predictions of the year 2090.

LRP identifies temperature over North Africa and Central Asia in JJA (Figure 6c) and the Andes and Central Africa in SON (Figure 6d) as the most relevant regions for predicting the year. For precipitation, the regions of highest relevance can be found in Canada and Russia in DJF and SON (Figures 7a and 7d) and in Central Africa and India in JJA and SON (Figures 7c and 7d). That is to say that these are the regional patterns identified by the neural networks that indicate the presence of forced change across the CMIP6 climate models. The scale of the color bars are different between Figures 6 and 7, such that the darkest regions in the temperature maps are approximately one order of magnitude more relevant than the darkest regions in the precipitation maps. Hence, the neural network is relying more heavily on the temperature inputs than the precipitation inputs to accurately predict the year. This is not surprising because the forced signal of temperature is clearer than the forced signal of precipitation (Figure SPM.7 in Field et al., 2014). Even so, including seasonal precipitation allows the neural networks to detect forced change earlier within combined fields than in temperature fields alone (Figure 5b). The improvement in neural network performance provided by precipitation (alongside temperature) is particularly noteworthy given that the S/N ratio for temperature is larger than the S/N ratio for precipitation in all seasons and regions (Figures 6e–6h and 7e–7h, discussed further in this section). In other words, the forced temperature signal is always more pronounced than the forced precipitation signal, but the precipitation signal is still useful for detecting forced change.

LRP is designed to highlight the regions that were most relevant for predicting the correct class (in our case, the correct decade class). These LRP indicator patterns for 2090 are not the time-mean patterns of the forced response, they are the patterns used by the neural network to distinguish the end of the twenty-first century from all other decades. This is distinctly different from S/N ratio which identifies the regions where the forced change from 1920 to 2090 is largest relative to internal variability and climate model spread. Maps of S/N ratio for temperature are shown in Figures 6e–6h, and the corresponding maps for precipitation are shown in Figures 7e–7h, where a higher S/N ratio (darker green) indicates a clearer forced signal. These regions of high S/N ratio are consistent with other related studies (e.g., Hawkins et al., 2020). For the most part, the indicator patterns identified by LRP correspond with the regions with the highest S/N ratios. Calculating the Spearman's rank correlation (ρ) between each map of relevance and S/N ratio, we find that there is generally a strong positive correlation ($0.71 \leq \rho \leq 0.77$) between the LRP indicator patterns and the S/N ratios for temperature, and a moderate positive correlation ($0.30 \leq \rho \leq 0.56$) for precipitation. The exact correlation coefficients between each map are displayed in the subtitles for Figures 6e–6h and 7e–7h.

Given that precipitation only contributes a small amount of relevance compared to temperature, it is perhaps unsurprising that there are several regions where the S/N ratio for precipitation is high, but the relevance is low (e.g., Alaska in JJA, Figures 7c and 7g or South Africa in SON, Figures 7d and 7h). Most likely, the forced signal of temperature is clear enough that these regions do not add to the predictive skill of the neural networks. Regions also exist where the S/N ratio for temperature is high despite low relevance (e.g., North Africa in DJF, Figures 6a and 6e), although these are more rare, as hinted by the strong correlation between the temperature maps of S/N ratio and relevance. In contrast, there are fewer regions with high relevance despite low S/N ratios, but they do occur (e.g., SON temperatures in northern South America, Figures 6d and 6h). These high-relevance, low-S/N ratio regions confirm that the indicator patterns identified by LRP capture more than the local S/N ratio. Some reasons a region/variable/season may be important in terms of LRP, but not in terms of S/N ratio, are: (a) LRP may be identifying places in our data where a signal exists only in the combination of regions/seasons/variables, which would not be captured by this definition of S/N ratio. (b) Since LRP highlights the patterns the neural

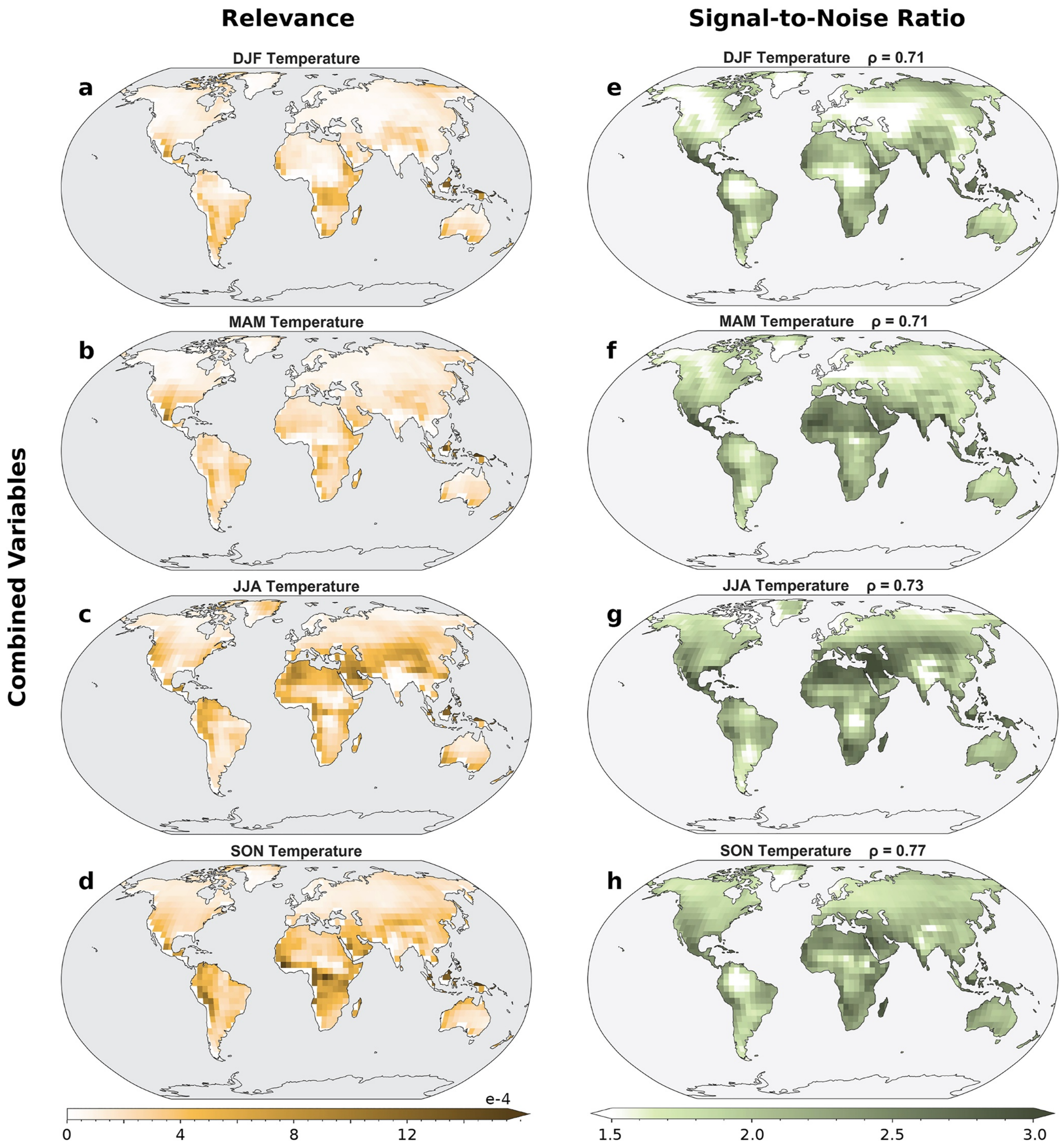

Figure 6. Combined indicator patterns of the forced response (temperature). Average temperature LRP results for the seasonal-mean combined neural networks (left, in yellow) and signal-to-noise (S/N) ratio (right, in green) for 2090. Darker shading indicates regions of temperature that are more relevant for the neural network's prediction or have a higher S/N ratio. The Spearman's rank correlation (ρ) between corresponding maps of relevance and S/N ratio are shown in the subtitles of panels (e–h).

networks use to predict the correct decade over all other decades, it may be capturing abrupt nonlinear changes in the climate that are filtered out by the century-long analysis of S/N ratio In the next section, we discuss further applications of neural network-derived indicator patterns and task the network with the much harder problem of identifying changes in extreme precipitation over the Americas.

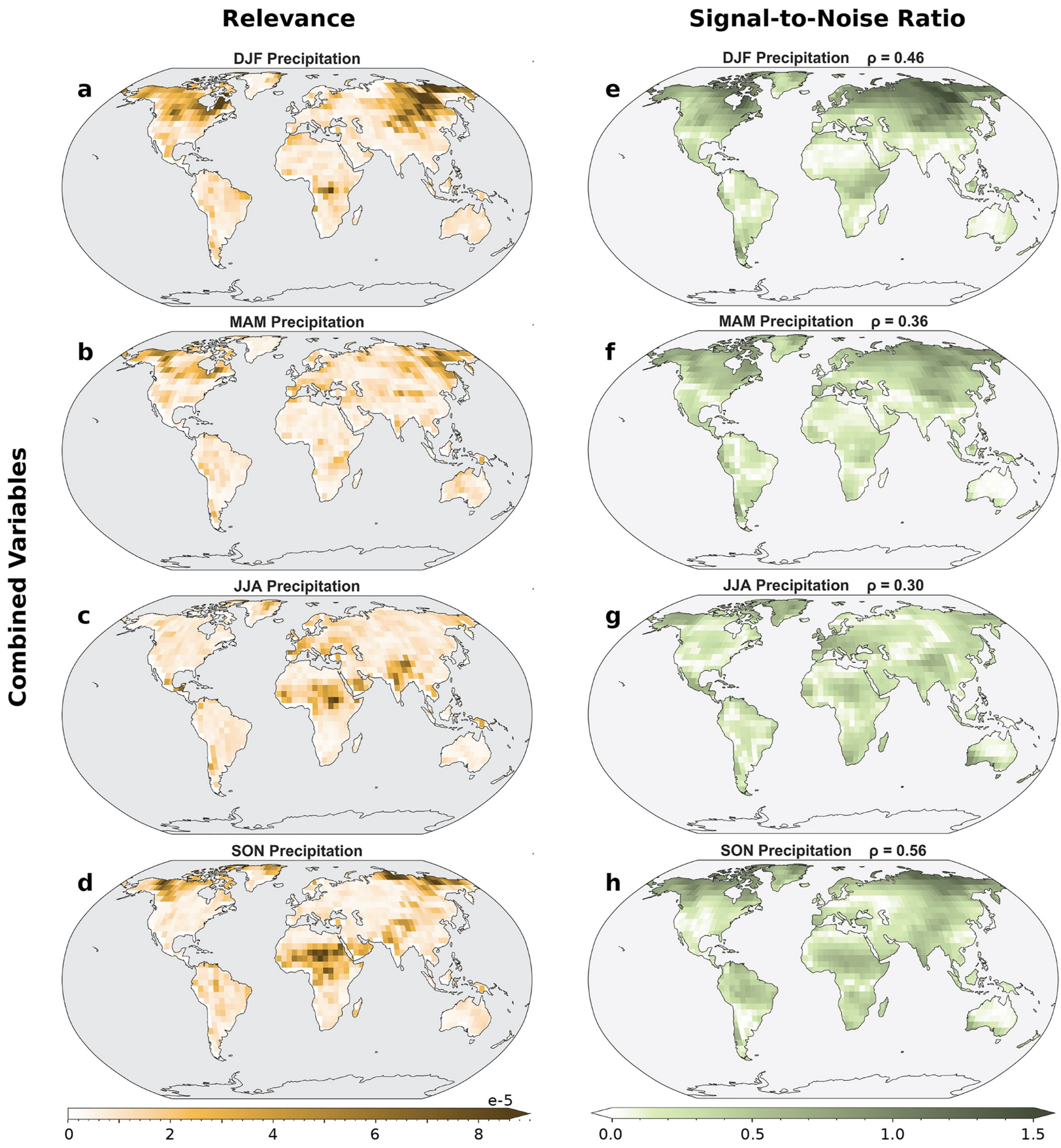

Figure 7. Combined indicator patterns of the forced response (precipitation). Average precipitation layer-wise relevance propagation results for the seasonal-mean combined neural networks (left, in yellow) and signal-to-noise (S/N) ratio (right, in green) for 2090. Darker shading indicates regions of precipitation that are more relevant for the neural network's prediction or have a higher S/N ratio. The Spearman's rank correlation (ρ) between corresponding maps of relevance and S/N ratio are shown in the subtitles of panels (e–h).

5. Extreme Precipitation Over the Americas

We now task the neural networks to predict the year given combinations of seasons for a single variable: extreme precipitation over the Americas. We choose to shift our focus for a few reasons. First, we wish to demonstrate

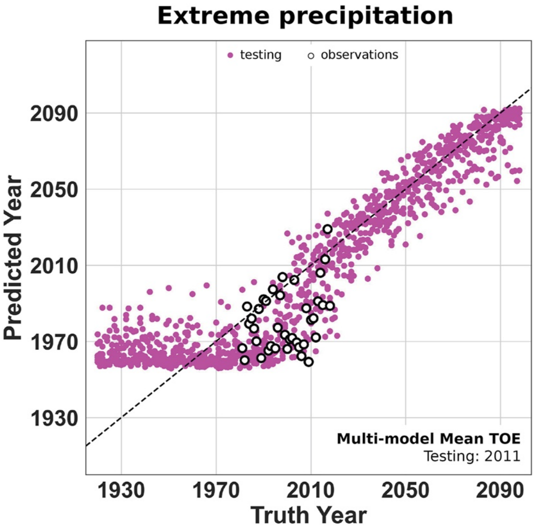

Figure 8. Neural network output for extreme precipitation. Year predicted by the neural network (y-axis) versus the truth year (x-axis) given seasonal-mean maps of extreme precipitation. Testing data is shown in pink and observations are shown in white.

that this neural network approach can be extended to variables that have considerable noise (like extreme precipitation, see Figure S8 in Supporting Information S1), and data sets that do not cover the globe. Second, extreme precipitation has major implications for human health (Ali et al., 2019; Eekhout et al., 2018; Rosenzweig et al., 2002) but there is considerable disagreement between climate models in its signal (Figure S8 in Supporting Information S1). This neural network approach can be used to identify agreed-upon patterns despite climate model spread. Further in this section, we will demonstrate that LRP maps can be used to investigate climate model differences and better understand the time evolution of the forced response.

The extreme precipitation signal is not as pronounced as the temperature signal, and using the Americas rather than the full globe limits the amount of unique information in the input field. Nevertheless, the neural networks are still able to detect patterns of forced change. Figure 8 depicts the years predicted by one neural network trained on seasonal-mean extreme precipitation. As in Figure 3, the neural network is unable to accurately predict the year given CMIP6 data prior to the TOE around 2010, whereafter the predicted years generally follow the 1:1 line with the truth years, indicating that the neural network has identified reliable indicators of change for this period. All Pearson correlations of the actual years with the predicted years for extreme precipitation in observations are positive ($r \approx 0.4$), demonstrating that the indicator patterns found in climate models can be successfully applied to observations (Figure 4). These correlations are not as strong as those for mean precipitation observations, due in part to the magnitude of climate model disagreement in extreme precipitation as well as the observational data set used: ERA5. As shown in Figure S6 in Supporting Information S1, the correlations of actual with predicted years for ERA5 precipitation observations are far smaller than those for GPCP observations. ERA5 tends to perform poorly in remote regions such as northern North America and northwestern South America (Bell et al., 2021), which may be responsible for these low correlations. The correlation between actual years and neural network-predicted years for extreme precipitation observations are explored in much more detail by Madakumbura et al. (2021).

To investigate the indicator patterns used by the neural networks to predict the year when the forced signal first emerges from the background noise, we apply LRP to all climate model samples in the training and testing sets for all 100 neural networks at the TOE (using the TOE calculated for each climate model and neural network individually, see Figure S9 in Supporting Information S1). LRP points to western South America in DJF and British Columbia in MAM and SON as the most relevant regions when the neural networks first detect the forced response (Figures 9a–9d). These LRP maps exhibit a more even distribution in relevance across each region and season than the end-of-the-twenty-first-century LRP maps of global temperature and precipitation (Figures 6a–6d and 7a–7d). Predicting the year at the TOE, when the signal has just barely emerged from the background climate, likely requires the neural networks to use all of the information available to them.

Up to this point, we have only considered the mean LRP maps across climate models. Since the neural networks are nonlinear by nature, they can identify multiple patterns that differ between climate models for a given decade. We apply k-means clustering to all 3200 LRP maps at the TOE (32 climate models samples, 100 neural networks) to identify two distinct indicator patterns that are being used by the climate models (Figures 9e–9l, see Supporting Information S1 for more details on k-means clustering). Taking the difference between the mean LRP maps for clusters one and two reveals that the Amazon in JJA is a highly relevant region in cluster one, while western Canada in DJF is a highly relevant region in cluster two (Figures 9m–9p). With the sole exception of MPI-ESM1-2-HR, all 100 LRP maps for each individual climate model fall cleanly into one cluster or the other, suggesting that there are two distinct ways in which the forced signal emerges in the CMIP6 simulations (Figure 10). Interestingly, when k-means is instructed to identify 32 unique clusters within the LRP maps, each cluster contains all 100 relevance maps for each of the 32 climate models. In other words, the pathway used by the neural networks to predict the year is unique to each climate model and distinguishable from all other climate

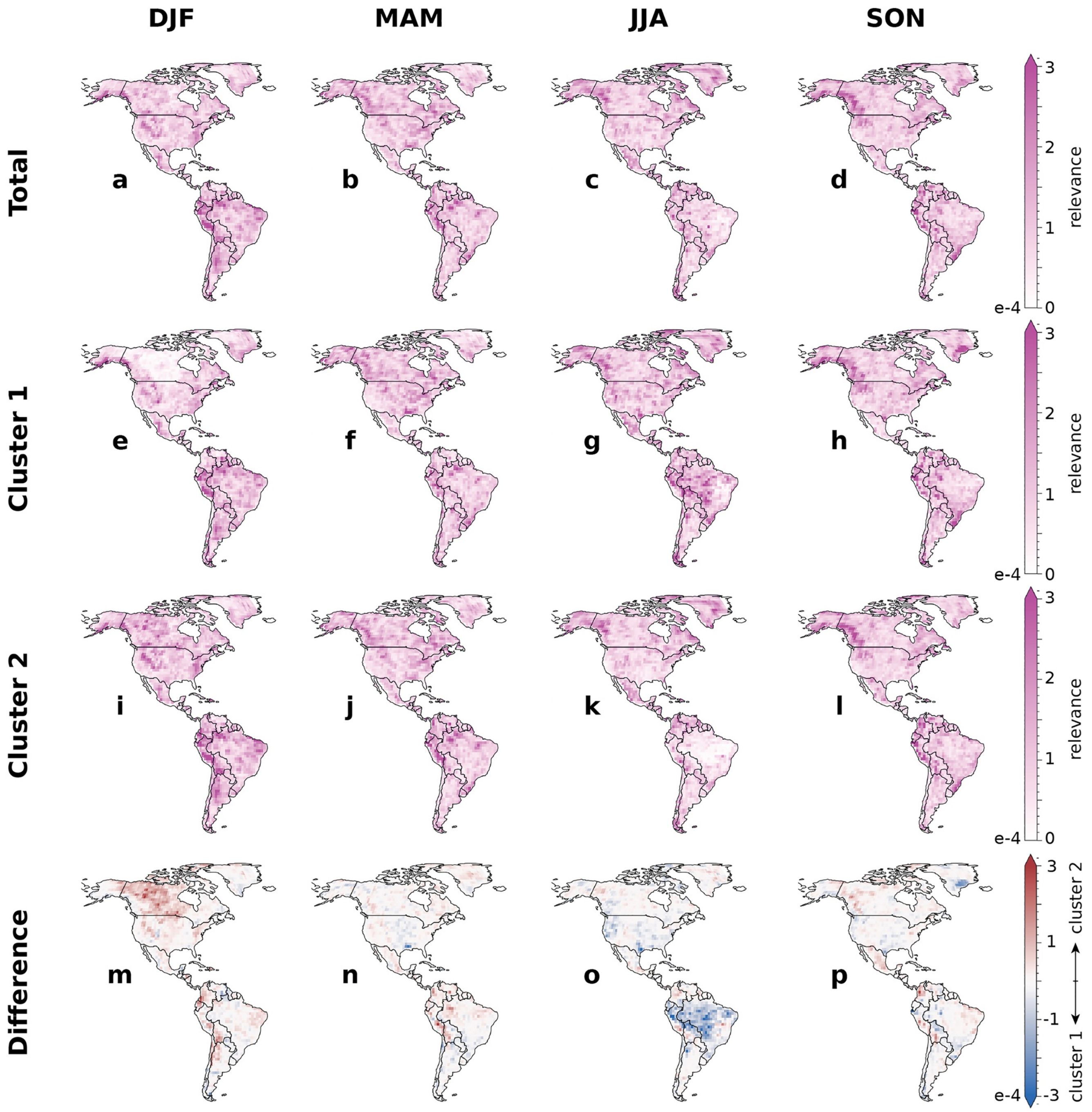

Figure 9. Relevance map clusters at the time of emergence (TOE) for extreme precipitation. Average layer-wise relevance propagation results for: extreme precipitation at the TOE (a–d), each cluster identified by k-means (e–l), and the difference between the clusters (m–p). In panels (a–l), darker shading indicates regions of extreme precipitation that are more relevant for the neural networks' prediction of the year at the TOE. In panels (m–p), blue shading indicates the regions that are more relevant in cluster 1, while red shading indicates the regions that are more relevant in cluster 2. Note that panels (a–d) are identical to panels (a–d) in Figure 11.

models, regardless of whether the climate model samples appear in the training or testing sets (further investigated by Labe and Barnes (2022)).

In the same way that indicator patterns can differ between models, indicator patterns are also able to evolve through time (e.g., Barnes et al., 2020; Labe & Barnes, 2021; Madakumbura et al., 2021). Comparing the LRP maps at the TOE (Figures 11a–11d) with those at the end of the twenty-first century (Figures 11e–11h) highlights the regions that become more important for predicting the year over time. The difference plots in Figures 11i–11l

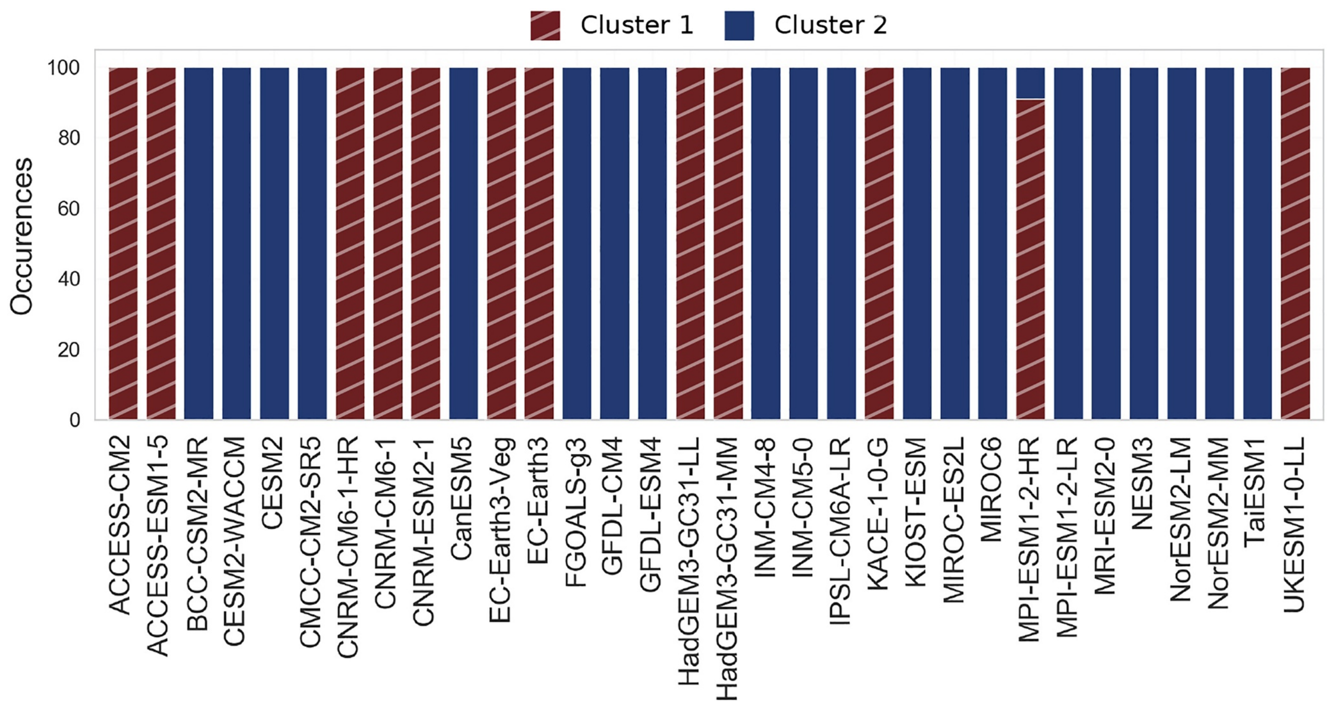

Figure 10. Climate models in each relevance map cluster at the time of emergence (TOE). The number of times each climate model appears in each cluster when *k*-means is applied to the maps of relevance at the TOE for 100 ANNs trained on extreme precipitation over the Americas. Only the relevance maps for MPI-ESM1-2-HR appear in both clusters. All other relevance maps for each climate model are found in one cluster or the other.

reveal that the neural network learns to focus on Alaska during MAM, JJA, and SON, Greenland in JJA and SON, and Quebec in MAM and SON as the forced response becomes stronger. These regions are more important for predicting the year at the end of the twenty-first century than the early twenty-first century. While further exploration is required, there are several reasons a region may become more relevant over time. For example, it may be that the region does not initially have a clear forced signal, but following some abrupt change (e.g., an ice-free Arctic) the forced signal becomes extremely pronounced. It may also be that the region has a signal that is consistently agreed upon by the majority of CMIP6 climate models, and becomes more relevant compared to other regions as climate model projections in those other regions drift apart. These time-varying patterns support the idea that combined indicators are effective for identifying dynamically evolving patterns of forced change.

6. Conclusions

Neural networks are powerful tools for identifying patterns of forced change in the climate system. When tasked with predicting the year given climate model simulations of temperature, precipitation, or extreme precipitation, artificial neural networks can learn these patterns of forced change that allow them to distinguish between maps from different years. In combined fields, such as multiple variables, seasons, or both, the forced response can be detected earlier than in single fields alone. By visualizing the decision-making process of the neural networks with an explainability method we extracted reliable, multivariate patterns of forced change. These neural network-derived combined indicator patterns are complex and nonlinear and capture more than the local S/N ratio. Explainability methods take a huge step toward disentangling the relationships learned by neural networks by pointing out what inputs contributed most to the final prediction, but they stop short of explaining why.

Expanding on previous work by Barnes et al. (2020), we used k-means clustering in tandem with LRP to study the relationships learned by the neural networks. This approach revealed two distinct ways in which the extreme precipitation response emerges in CMIP6 data. While combining neural network explainability methods with other statistical techniques can enhance explanations of neural network decisions, there is still a large gap between what the neural network has learned and what we can explain post hoc. Some unanswered questions, such as "why does temperature in Region A combine with precipitation in Region B to improve the signal of the forced

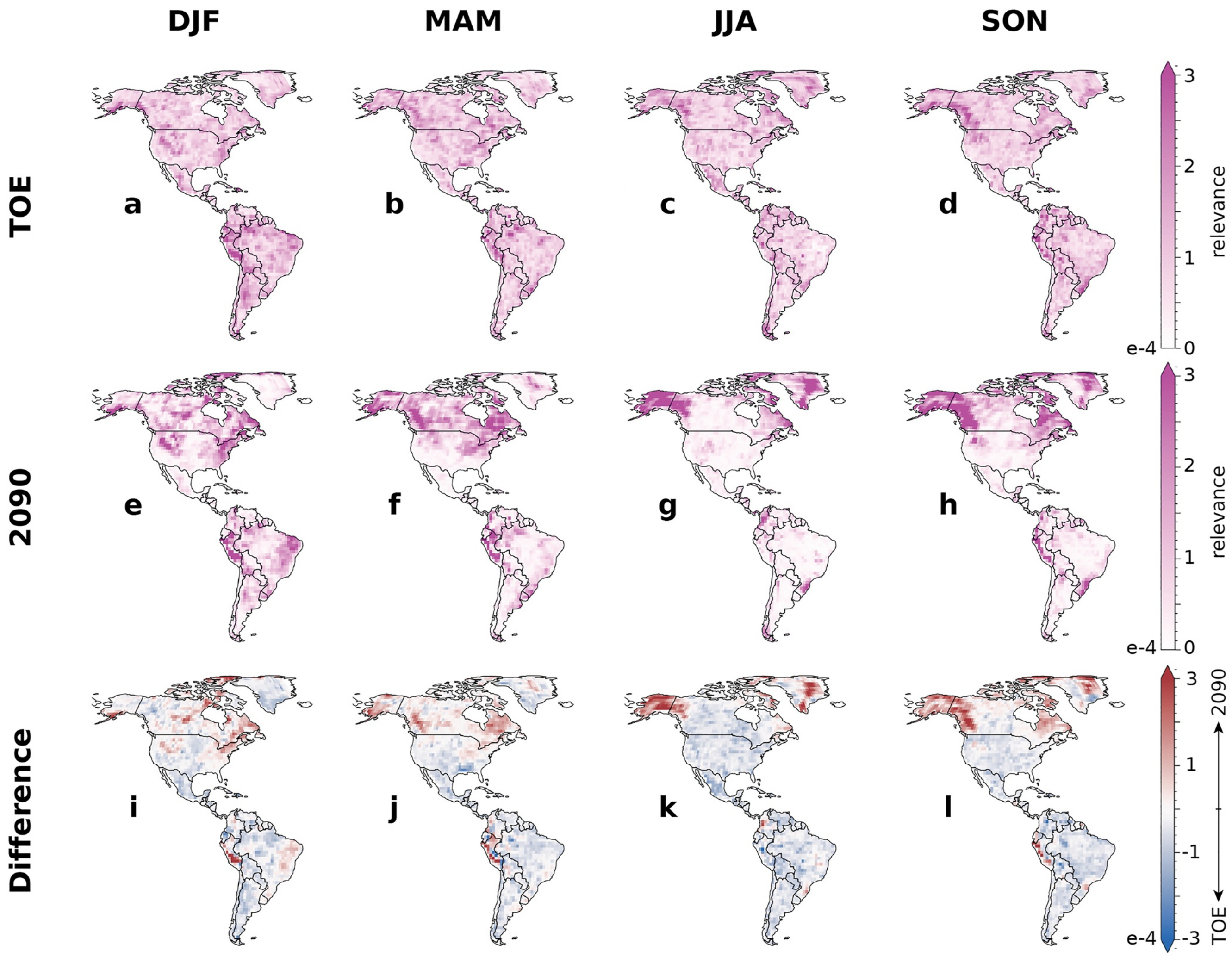

Figure 11. Time evolution of extreme precipitation relevance. Average layer-wise relevance propagation results at the time of emergence (a–d), 2090 (e–h), and the difference between (i–l). Darker shading in panels (a–h) highlights regions that were more relevant for the neural networks' prediction of the year. In panels (i–l), red shading indicates regions where the relevance has increased over time, while blue shading indicates regions where the relevance has decreased over time. Note that panels (a–d) are identical to panels (a–d) in Figure 9.

response?" may be better answered with a different architectural approach, such as neural network designs that are inherently interpretable and do not require post-hoc approaches like LRP (Rudin, 2019). This is a natural next step for future work. The flexibility and accessibility of this framework provide several other future research directions. Given that this predict-the-year approach can be applied to observational data, one possible extension of this work could involve exploring the observed features of forced change that are consistent with climate model simulations. There is also space for these methods to be used to determine which definitions of seasons are optimal for detecting forced change. While we used meteorological seasons here, there may be more appropriate definitions, such as unique definitions of the wet and dry seasons, or the shoulder seasons, that vary between variables and regions. Furthermore, this framework should be expanded to other variables, regions of focus, and climate change scenarios, to identify the combined indicators that best elucidate the forced signal. For example, extreme precipitation and extreme drought may combine to capture the increased volatility in precipitation extremes that are expected with climate change (O'Gorman, 2015). Further application of this technique to compound climate extremes, such as heat wave intensity, drought duration, and flood frequency, may reveal that explainable neural networks are useful for assessing societal impacts and improving climate change preparedness.

Data Availability Statement

All data used in this study is publicly available and referenced throughout the paper. The CMIP6 simulations used in this study can be via the Earth System Grid Federation (ESGF, https://esgf-node.llnl.gov/projects/cmip6/). Monthly temperature observations are available through Berkeley Earth (http://berkeleyearth.org/data/). Global Precipitation Climatology Project monthly global precipitation fields are available through the NOAA Physical Sciences Laboratory (https://psl.noaa.gov/data/gridded/data.gpcp.html). Monthly, daily, and sub-daily precipitation reanalyses were provided by the European Centre for Medium-Range Weather Forecasts (ERA5: https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5) and the National Center for Atmospheric Research (JRA55: https://climatedataguide.ucar.edu/climate-data/jra-55). Python code used in this work has been made publicly available at https://doi.org/10.5281/zenodo.6780638.

Acknowledgments

The authors thank the two anonymous reviewers, the Editor, Maria Rugenstein, and Jessica Witt for their constructive feedback, which substantially improved this study. JKR would also like to thank the members of the Barnes Group for all the science discussions that brought about new ideas, and his parents and siblings for their support through the pandemic. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0020347, and by the NOAA MAPP grant NA19OAR4310289. The authors acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

Adler, R. F., Sapiano, M., Huffman, G. J., Wang, J., Gu, G., Bolvin, D., et al. (2018). The Global Precipitation Climatology Project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation. *Atmosphere*, *9*(4), 138. https://doi.org/10.3390/atmos9040138

Ali, H., Modi, P., & Mishra, V. (2019). Increased flood risk in Indian sub-continent under the warming climate. *Weather and Climate Extremes*, *25*, 100212. https://doi.org/10.1016/j.wace.2019.100212

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise explanations for Non-Linear classifier decisions by Layer-Wise relevance propagation. *PLoS One*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019). Viewing forced climate patterns through an AI lens. *Geophysical Research Letters*, *46*(22), 13389–13398. https://doi.org/10.1029/2019gl084944

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002195. https://doi.org/10.1029/2020ms002195

Barnett, T. P., Pierce, D. W., Hidalgo, H. G., Bonfils, C., Santer, B. D., Das, T., et al. (2008). Human-induced changes in the hydrology of the western United States. *Science*, *319*(5866), 1080–1083. https://doi.org/10.1126/science.1152538

Bell, B., Hersbach, H., Simmons, A., Berrisford, P., Dahlgren, P., Horányi, A., et al. (2021). The ERA5 global reanalysis: Preliminary extension to 1950. *Quarterly Journal of the Royal Meteorological Society*, *147*(741), 4186–4227. https://doi.org/10.1002/qj.4174

Bindoff, N. L., Stott, P. A., AchutaRao, K. M., Allen, M. R., Gillett, N., Gutzler, D., et al. (2013). Chapter 10 – Detection and attribution of climate change: From global to regional. In *Climate Change 2013: The Physical Science $basis. IPCC Working Group I Contribution to AR5*. Cambridge University Press.

Bonfils, C. J. W., Santer, B. D., Fyfe, J. C., Marvel, K., Phillips, T. J., & Zimmerman, S. R. H. (2020). Human influence on joint changes in temperature, rainfall and continental aridity. *Nature Climate Change*, *10*(8), 726–731. https://doi.org/10.1038/s41558-020-0821-1

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. https://doi.org/10.1029/2018gl078510

Chen, C., Zhang, P., Zhang, H., Dai, J., Yi, Y., Zhang, H., & Zhang, Y. (2020). Deep learning on computational-resource-limited platforms: A survey. *Mobile Information Systems*, *2020*, 1–19. https://doi.org/10.1155/2020/8454327

Cui, L., Wang, L., Qu, S., Singh, R. P., Lai, Z., & Yao, R. (2019). Spatiotemporal extremes of temperature and precipitation during 1960–2015 in the Yangtze River basin (China) and impacts on vegetation dynamics. *Theoretical and Applied Climatology*, *136*(1), 675–692. https://doi.org/10.1007/s00704-018-2519-0

Dai, A., Lin, X., & Hsu, K.-L. (2007). The frequency, intensity, and diurnal cycle of precipitation in surface and satellite observations over low- and mid-latitudes. *Climate Dynamics*, *29*(7–8), 727–744. https://doi.org/10.1007/s00382-007-0260-y

Donat, M. G., Alexander, L. V., Herold, N., & Dittus, A. J. (2016). Temperature and precipitation extremes in century-long gridded observations, reanalyses, and atmospheric model simulations. *Journal of Geophysical Research*, *121*(19), 11174–11189. https://doi.org/10.1002/2016jd025480

Eekhout, J. P. C., Hunink, J. E., Terink, W., & de Vente, J. (2018). Why increased extreme precipitation under climate change negatively affects water security. *Hydrology and Earth System Sciences Discussions*, *22*(11), 1–16. https://doi.org/10.5194/hess-2018-161

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Field, C. B., Barros, V. R., Mastrandrea, M. D., Mach, K. J., Abdrabo, M. A.-K., Adger, N., et al. (2014). Summary for policymakers. In *Climate change 2014: Impacts, adaptation, and vulnerability. Part A: Global and sectoral aspects. Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change* (pp. 1–32). Cambridge University Press.

Fischer, E. M., & Knutti, R. (2012). Robust projections of combined humidity and temperature extremes. *Nature Climate Change*, *3*(2), 126–130. https://doi.org/10.1038/nclimate1682

Gaetani, M., Janicot, S., Vrac, M., Famien, A. M., & Sultan, B. (2020). Robust assessment of the time of emergence of precipitation change in West Africa. *Scientific Reports*, *10*(1), 7670. https://doi.org/10.1038/s41598-020-63782-2

Gettelman, A., Gagne, D. J., Chen, C.-C., Christensen, M. W., Lebo, Z. J., Morrison, H., & Gantos, G. (2021). Machine learning the warm rain process. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002268. https://doi.org/10.1029/2020ms002268

Hawkins, E., Frame, D., Harrington, L., Joshi, M., King, A., Rojas, M., & Sutton, R. (2020). Observed emergence of the climate change signal: From the familiar to the unknown. *Geophysical Research Letters*, *47*(6), e2019GL086259. https://doi.org/10.1029/2019gl086259

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/10.1175/2009BAMS2607.1

Hawkins, E., & Sutton, R. (2012). Time of emergence of climate signals. *Geophysical Research Letters*, *39*(1), L01702. https://doi.org/10.1029/2011gl050087

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Kim, Y.-H., Min, S.-K., Zhang, X., Sillmann, J., & Sandstad, M. (2020). Evaluation of the CMIP6 multi-model ensemble for climate extreme indices. *Weather and Climate Extremes*, *29*, 100269. https://doi.org/10.1016/j.wace.2020.100269

Labe, Z. M., & Barnes, E. A. (2021). Detecting climate signals using explainable AI with single-forcing large ensembles. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2021MS002464. https://doi.org/10.1029/2021ms002464

Labe, Z. M., & Barnes, E. A. (2022). Comparison of climate model large ensembles with observations in the arctic using simple neural networks. https://doi.org/10.1002/essoar.10510977.1

Lagerquist, R., McGovern, A., & Gagne, D. J. (2019). Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting*, *34*(4), 1137–1160. https://doi.org/10.1175/WAF-D-18-0183.1

Lapuschkin, S. (2019). *Opening the machine learning black box with layer-wise relevance propagation* (Doctoral dissertation). Technischen Universität Berlin. https://doi.org/10.14279/DEPOSITONCE-7942

Lee, Y., Kummerow, C. D., & Ebert-Uphoff, I. (2021). Applying machine learning methods to detect convection using Geostationary Operational Environmental Satellite-16 (GOES-16) advanced baseline imager (ABI) data. *Atmospheric Measurement Techniques*, *14*(4), 2699–2716. https://doi.org/10.5194/amt-14-2699-2021

Li, J., Thompson, D. W. J., Barnes, E. A., & Solomon, S. (2017). Quantifying the lead time required for a linear trend to emerge from natural climate variability. *Journal of Climate*, *30*(24), 10179–10191. https://doi.org/10.1175/JCLI-D-16-0280.1

Lu, E., Chen, H., Tu, J., Song, J., Zou, X., Zhou, B., et al. (2015). The nonlinear relationship between summer precipitation in China and the sea surface temperature in preceding seasons: A statistical demonstration. *Journal of Geophysical Research – D: Atmospheres*, *120*(23), 12027–12036. https://doi.org/10.1002/2015JD024030

Madakumbura, G. D., Thackeray, C. W., Norris, J., Goldenson, N., & Hall, A. (2021). Anthropogenic influence on extreme precipitation over global land areas seen in multiple observational datasets. *Nature Communications*, *12*(1), 3944. https://doi.org/10.1038/s41467-021-24262-x

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

Mahony, C. R., & Cannon, A. J. (2018). Wetter summers can intensify departures from natural variability in a warming climate. *Nature Communications*, *9*(1), 783. https://doi.org/10.1038/s41467-018-03132-z

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science*, *1*(8), 1-17. https://doi.org/10.1017/eds.2022.7

Mankin, J. S., Lehner, F., Coats, S., & McKinnon, K. A. (2020). The value of initial condition large ensembles to robust adaptation decision-making. *Earth's Future*, *8*(10), e2012EF001610. https://doi.org/10.1029/2020ef001610

Marvel, K., & Bonfils, C. (2013). Identifying external influences on global precipitation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19301–19306. https://doi.org/10.1073/pnas.1314382110

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: An overview. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700, pp. 193–209). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_10

Mudelsee, M. (2019). Trend analysis of climate time series: A review of methods. *Earth-Science Reviews*, *190*, 310–322. https://doi.org/10.1016/j.earscirev.2018.12.005

Newell, G. J., & Lee, B. (1981). Ridge regression: An alternative to multiple linear regression for highly correlated data. *Journal of Food Science*, *46*(3), 968–969. https://doi.org/10.1111/j.1365-2621.1981.tb15400.x

North, G. R., & Stevens, M. J. (1998). Detecting climate signals in the surface temperature record. *Journal of Climate*, *11*(4), 563–577. https://doi.org/10.1175/1520-0442(1998)011<0563:DCSITS>2.0.CO;2

O'Gorman, P. A. (2015). Precipitation extremes under climate change. *Current Climate Change Reports*, *1*(2), 49–59. https://doi.org/10.1007/s40641-015-0009-3

O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., et al. (2016). The scenario model intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*(9), 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

Rohde, R. A., & Hausfather, Z. (2020). The Berkeley Earth land/ocean temperature record. *Earth System Science Data*, *12*(4), 3469–3479. https://doi.org/10.5194/essd-12-3469-2020

Rosenzweig, C., Tubiello, F. N., Goldberg, R., Mills, E., & Bloomfield, J. (2002). Increased crop damage in the US from excess precipitation under climate change. *Global Environmental Change*, *12*(3), 197–202. https://doi.org/10.1016/S0959-3780(02)00008-0

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sanderson, B. M., Oleson, K. W., Strand, W. G., Lehner, F., & O'Neill, B. C. (2018). A new ensemble of GCM simulations to assess avoided impacts in a climate mitigation scenario. *Climate Change*, *146*(3), 303–318. https://doi.org/10.1007/s10584-015-1567-z

Santer, B. D., Fyfe, J. C., Solomon, S., Painter, J. F., Bonfils, C., Pallotta, G., & Zelinka, M. D. (2019). 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. https://doi.org/10.1073/pnas.1904586116

Santer, B. D., Mears, C., Doutriaux, C., Caldwell, P., Gleckler, P. J., Wigley, T. M. L., et al. (2011). Separating signal and noise in atmospheric temperature changes: The importance of timescale. *Journal of Geophysical Research*, *116*(D22), D22105. https://doi.org/10.1029/2011jd016263

Santer, B. D., Taylor, K. E., Wigley, T. M. L., Johns, T. C., Jones, P. D., Karoly, D. J., et al. (1996). A search for human influences on the thermal structure of the atmosphere. *Nature*, *382*(6586), 39–46. https://doi.org/10.1038/382039a0

Scaife, A. A., & Smith, D. (2018). A signal-to-noise paradox in climate science. *Npj Climate and Atmospheric Science*, *1*(1), 1–8. https://doi.org/10.1038/s41612-018-0038-4

Schneider, T., & Held, I. M. (2001). Discriminants of twentieth-century changes in Earth surface temperatures. *Journal of Climate*, *14*(3), 249–254. https://doi.org/10.1175/1520-0442(2001)014<0249:LDOTCC>2.0.CO;2

Schulzweida, U. (2019). *CDO user guide* (version 1.9. 6). Max Planck Institute for Meteorology.

Silva, S. J., Ma, P.-L., Hardin, J. C., & Rothenberg, D. (2021). Physically regularized machine learning emulators of aerosol activation. *Geoscientific Model Development*, *14*(5), 3067–3077. https://doi.org/10.5194/gmd-14-3067-2021

Sippel, S., Meinshausen, N., Fischer, E. M., Székely, E., & Knutti, R. (2020). Climate change now detectable from any single day of weather at global scale. *Nature Climate Change*, *10*(1), 35–41. https://doi.org/10.1038/s41558-019-0666-7

Solomon, A., & Newman, M. (2012). Reconciling disparate twentieth-century Indo-Pacific ocean temperature trends in the instrumental record. *Nature Climate Change*, *2*(9), 691–699. https://doi.org/10.1038/nclimate1591

Solow, A. R. (1987). Testing for climate change: An application of the two-phase regression model. *Journal of Applied Meteorology and Climatology*, *26*(10), 1401–1405. https://doi.org/10.1175/1520-0450(1987)026<1401:TFCCAA>2.0.CO;2

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Swain, D. L., Wing, O. E. J., Bates, P. D., Done, J. M., Johnson, K. A., & Cameron, D. R. (2020). Increased flood exposure due to climate change and population growth in the United States. *Earth's Future*, *8*(11), e2020EF001778. https://doi.org/10.1029/2020ef001778

Tabari, H., & Willems, P. (2018). Seasonally varying footprint of climate change on precipitation in the Middle East. *Scientific Reports*, *8*(1), 4435. https://doi.org/10.1038/s41598-018-22795-8

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019ms002002

USGCRP. (2018). *Impacts, risks, and adaptation in the United States: Fourth national climate assessment, volume II* (Tech. Rep.). U.S. Global Change Research Program. https://doi.org/10.7930/NCA4.2018

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002109. https://doi.org/10.1029/2020ms002109

Wills, R. C., Battisti, D. S., Armour, K. C., Schneider, T., & Deser, C. (2020). Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations. *Journal of Climate*, *33*(20), 8693–8719. https://doi.org/10.1175/JCLI-D-19-0855.1

Wills, R. C., Schneider, T., Wallace, J. M., Battisti, D. S., & Hartmann, D. L. (2018). Disentangling global warming, multidecadal variability, and El Niño in Pacific temperatures. *Geophysical Research Letters*, *45*(5), 2487–2496. https://doi.org/10.1002/2017GL076327

Wu, J., Chen, X., Yao, H., Gao, L., Chen, Y., & Liu, M. (2017). Non-linear relationship of hydrological drought responding to meteorological drought and impact of a large reservoir. *Journal of Hydrology*, *551*, 495–507. https://doi.org/10.1016/j.jhydrol.2017.06.029

Zadeh, L. A. (1965). Information and control. *Fuzzy Sets and Systems*, *8*(3), 338–353. https://doi.org/10.1002/joc.1027

Zappa, G., Hoskins, B. J., & Shepherd, T. G. (2015). Improving climate change detection through optimal seasonal averaging: The case of the North Atlantic jet and European precipitation. *Journal of Climate*, *28*(16), 6381–6397. https://doi.org/10.1175/JCLI-D-14-00823.1

References From the Supporting Information

Celisse, A., & Robin, S. (2008). Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics & Data Analysis*, *52*(5), 2350–2368. https://doi.org/10.1016/j.csda.2007.10.002

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., et al. (2015). The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Series II*, *93*(1), 5–48. https://doi.org/10.2151/jmsj.2015-001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.