THESIS

DETECTING FORCED CHANGE WITHIN COMBINED CLIMATE FIELDS USING EXPLAINABLE NEURAL NETWORKS

Submitted by

Jamin Rader

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2021

Master's Committee:

Advisor: Elizabeth Barnes

Maria Rugenstein
Jessica Witt

ABSTRACT

DETECTING FORCED CHANGE WITHIN COMBINED CLIMATE FIELDS USING EXPLAINABLE NEURAL NETWORKS

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 visualization 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 visualization tools can be harnessed to identify patterns of the forced signal within combined fields.

ACKNOWLEDGEMENTS

I would like to thank my adviser, Dr. Elizabeth Barnes, my committee, Dr. Maria Rugenstein and Dr. Jessica Witt, and my co-contributors, Dr. Imme Ebert-Uphoff and Dr. Chuck Anderson, for their support in this thesis work, and my family for providing a place to write. 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 NOAA MAPP grant NA19OAR4310289. Data access: https://esgf-node.llnl.gov/projects/cmip6/ (CMIP6), http://berkeleyearth.org/data/ (BEST), https://psl.noaa.gov/data/gridded/data.gpcp.html (GPCP), https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5 (ERA5), https://climatedataguide.ucar.edu/climate-data/jra-55 (JRA55). We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we 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. All data used in this study is publicly available and referenced throughout the paper.

PREFACE

The following materials will be submitted to a peer-reviewed publication.

TABLE OF CONTENTS

Chapter 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. 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 and Sutton, 2009; Mahony and Cannon, 2018; Scaife and 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 and Newman, 2012), to name a 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 and 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 and 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 CMIP5 models. Barnes et al. (2020) demonstrated that neural networks, and their visualization methods, are powerful tools for extracting forced patterns from climate data. Since then, neural networks have 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 datasets (Labe and 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 and Connon, 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 and 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 datasets by identifying the multivariate patterns that enhance the signal of change relative to the underlying noise (e.g., Barnett

et al., 2008; Marvel and 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.

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 are explored here. 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 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 highlights the advantages and disadvantages of using neural networks and LRP for forced response detection and scientific exploration.

Chapter 2

Data

2.1 CMIP6 Climate Models

We use climate model output from the sixth phase of the Coupled Model Intercomparison Project (CMIP6; Eyring et al. 2016). Specifically we focus on monthly-, seasonal-, and annual-mean fields of 2-meter 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 A.1 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 address the issue of climate model "drizzle" (e.g., Chen et al., 2021), we only include days that simulated at least 1 mm of precipitation when calculating the 95th percentile of all days with precipitation.

One ensemble member is selected for each of the 37 CMIP6 climate models analyzed. Since daily output is required to calculate very wet days, we are limited to 32 models for extreme precipitation (Figure A.2). We analyze the climate model data from 1920 to 2098 under historical forcing (1920–2014) and the SSP585 scenario (2015–2098). SSP585 represents the highest development pathway within CMIP6 scenarios (O'Neill et al., 2016), combining shared socioeconomic pathway 5 (SSP5) and representative concentration pathway 8.5 (RCP8.5).

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). For temperature and precipitation, the data is regridded to 4 degrees latitude by 4 degrees longitude. For extreme precipitation, the data is regridded to 1.5 degrees latitude by 1.5 degrees 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. 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). Each map of temperature and precipitation has 948 unique data points, while each map of extreme precipitation has 2314 unique data points.

2.2 Observations

While this work largely focuses on the results of the neural network fed climate model data as input, we show that these methods can be extended to observational data as well. For temperature, we use the Berkeley Earth Surface Temperature (BEST) dataset (Rohde and Hausfather, 2020). This dataset 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 GPCP precipitation observations are only available back to October 1996, we use the European Centre for Medium-Range Weather Forecasts' ERA5 global reanalysis (Hersbach et al., 2020) at 6-hour resolution to construct observational monthly mean extreme precipitation fields from 1980 to the present. All observations are regridded in the same way as the climate model data for each respective variable.

Chapter 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 a single simulated year from 1920 to 2098, are tasked to predict the year that is being simulated. This follows the framework established by Barnes et al. (2019, 2020) and extends it to combined fields (multiple maps) of input. The results for neural networks trained on ten 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 3.1. Each unit of the input layer vector to the neural network 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 ten nodes each, which are subsequently followed by the output layer. The output layer 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. More details on the neural network design can be found in the supplementary materials.

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) as in Barnes et al. (2020) such that each year can be mapped

to one or more neighboring classes with varying degrees of membership (encoded as likelihood). 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 Fig. 2 of Barnes et al. (2020).

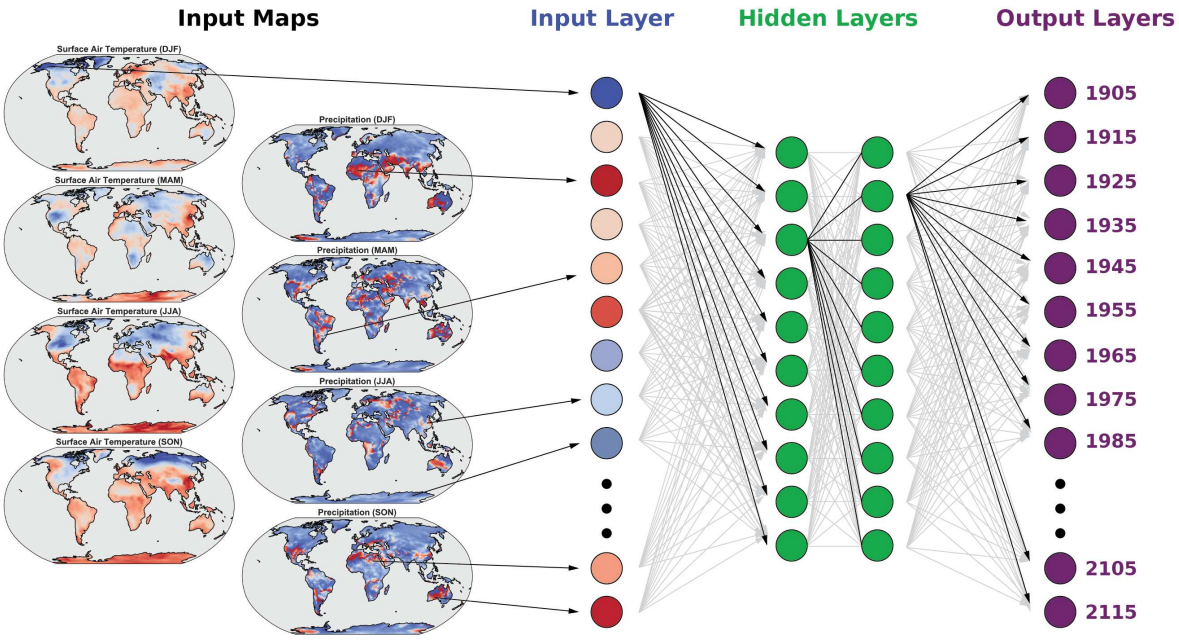

Figure 3.1: Schematic of the fully connected neural network architecture. Inputs from multiple maps of data are flattened into an input layer vector. These inputs are fed through two hidden layers with ten nodes each. The neural network is optimized to predict the year that the data came from, outputting 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.

3.2 Neural Network Training

For each input vector we train 100 neural networks that differ only in which simulations were split into 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 the supplementary materials.

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 patterns of the noise in the training dataset which can lead to overfitting (Srivastava et al., 2014). To reduce overfitting, 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 weight 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. 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 of the forced climate signal (see Section 3.3). By testing the sensitivity of neu-

ral network outputs to different choices in ridge penalty, we find that neural networks trained with smaller ridge penalties have a smaller MSE in the 21st century when the forced signal is easily detectable—a period where the MSE is already small—at the expense of larger MSE before the time of emergence. This leads to a later calculation of time of emergence for the testing set. Increasing the ridge penalty allows the neural networks to detect the climate change signal earlier (Figure A.3). The ridge penalty used for each input vector can be found in the supplementary materials. We use the same ridge penalty for all 100 neural networks trained on each input vector.

We standardized the input fields (for climate models and observations) to assist with the training and overall performance of the neural network. Departing from the methodology of Barnes et al. (2020), 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 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 1980–2019 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 datasets include the years 1980 to 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 time of emergence of the forced climate response (hereafter, simply "TOE") is the time in which the forced response signal is distinguishable from the background climate by the neural network. Specifically, we follow Barnes et al. (2020) and 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 3.2. First, we calculate the maximum of the neural network-predicted years over 1920–1959 for each model. We then identify the TOE as the earliest year in which a map, and all subsequent maps, permanently exceed this maximum predicted year from the baseline period. In Figure 3.2, 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.

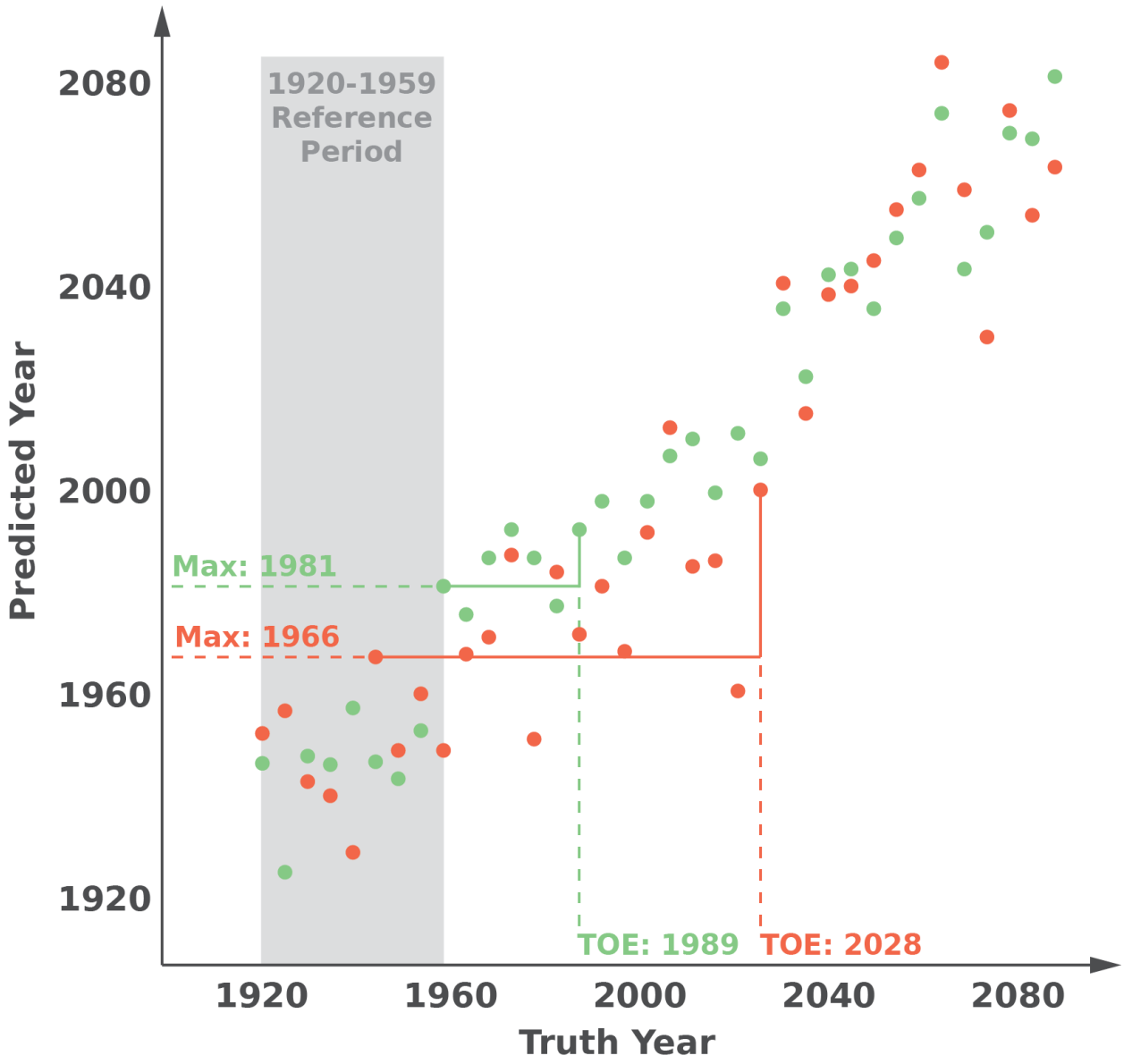

Figure 3.2: Calculation of TOE. The TOE is defined as the earliest year in which a map, and all subsequent maps, permanently exceed the maximum predicted year from the baseline period (1920-1959). Sample model 1 (red) has a baseline maximum of 1966 and permanently exceeds this threshold in 2028. Sample model 2 (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.

3.4 Layer-wise Relevance Propagation

To visualize the patterns learned by the neural network we apply layer-wise relevance propagation (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 Fig. 2 in Bach et al., 2015). We refer to this quantity as relevance. Through this backpropagation process the output value is conserved. 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. 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.

3.5 Signal-to-Noise Ratio Calculation

In Section 4, we compare the LRP relevance maps to maps of signal-to-noise ratio (S/N ratio), a more conventional method for identifying indicator patterns of the forced response. S/N ratio

consists of three distinct components: (1) the forced signal, which is divided by the sum of (2) noise due to internal variability, and (3) 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 et al. (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 A.7.

Chapter 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 20th 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 4.1). However, as the signal increases in magnitude into the late-20th and 21st 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) as well as unseen data (the testing set). These behaviors are shown in Figure 4.1 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 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 20th 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.

Given that the predictive skill is similar for the training and testing sets we are confident that these neural networks, and thus their learned patterns, are generalizable across CMIP6 models. 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 4.1). 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 4.2. All correlations are positive, indicating that the years predicted by the neural networks increase with time. These correlations are most strong 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 A.4), 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 dataset of choice, which is expanded on in Appendix 3 and Figures A.4 and A.5.

The average TOEs, calculated from the climate models in the testing sets of all 100 trained neural networks for each input field (Figure 4.3), reveal that the forced response can be detected earlier in maps of temperature than in maps of precipitation (Figure 4.3a-c). 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 (Figure 4.3b,f). The TOE is generally earlier for the neural networks trained on seasonal-mean data than for the neural networks trained on annual-mean data (Figure 4.3d-f). This is most notable for precipitation fields, likely because there are large seasonal precipitation responses muted by taking the annual mean (Tabari and Willems, 2018; Zappa et al., 2015). The TOEs are earlier for temperature and precipitation combined than temperature alone when using seasonal-mean maps (Figure 4.3b), but are approximately equal when using annual-mean or monthly-mean maps (Figure 4.3a,c), which suggests that precipitation only improves upon the detectability of the forced temperature signal when seasonal-mean fields are used (discussed further in Figure A.4). The neural networks identify the earliest TOEs when trained on seasonal-mean temperature and precipitation combined (Figure 4.3b,f). The TOE results for all 100 seasonal-mean neural networks are summarized in the box plots in Figure A.6.

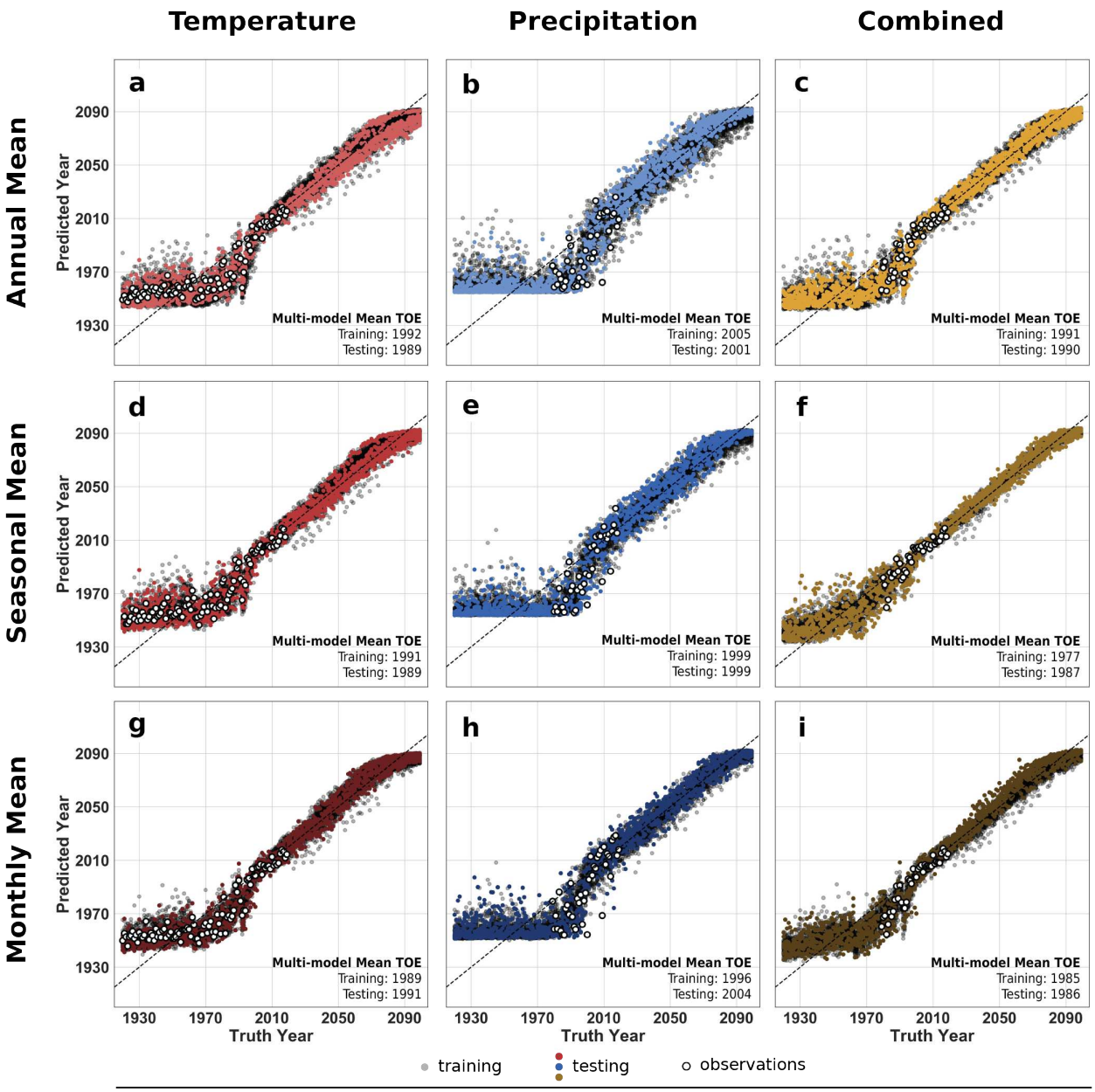

Figure 4.1: Neural network output for temperature and precipitation. Year predicted by the neural network (y-axis) versus the truth year (x-axis) for temperature (a, d, g), precipitation (b, e, h), and temperature and precipitation combined (c, f, i). Input maps include annual-mean data (a, b, c), seasonal-mean data (d, e, f), and monthly-mean data (g, h, i). Training data is shown in gray, testing data is shown in color, and observations are shown in white.

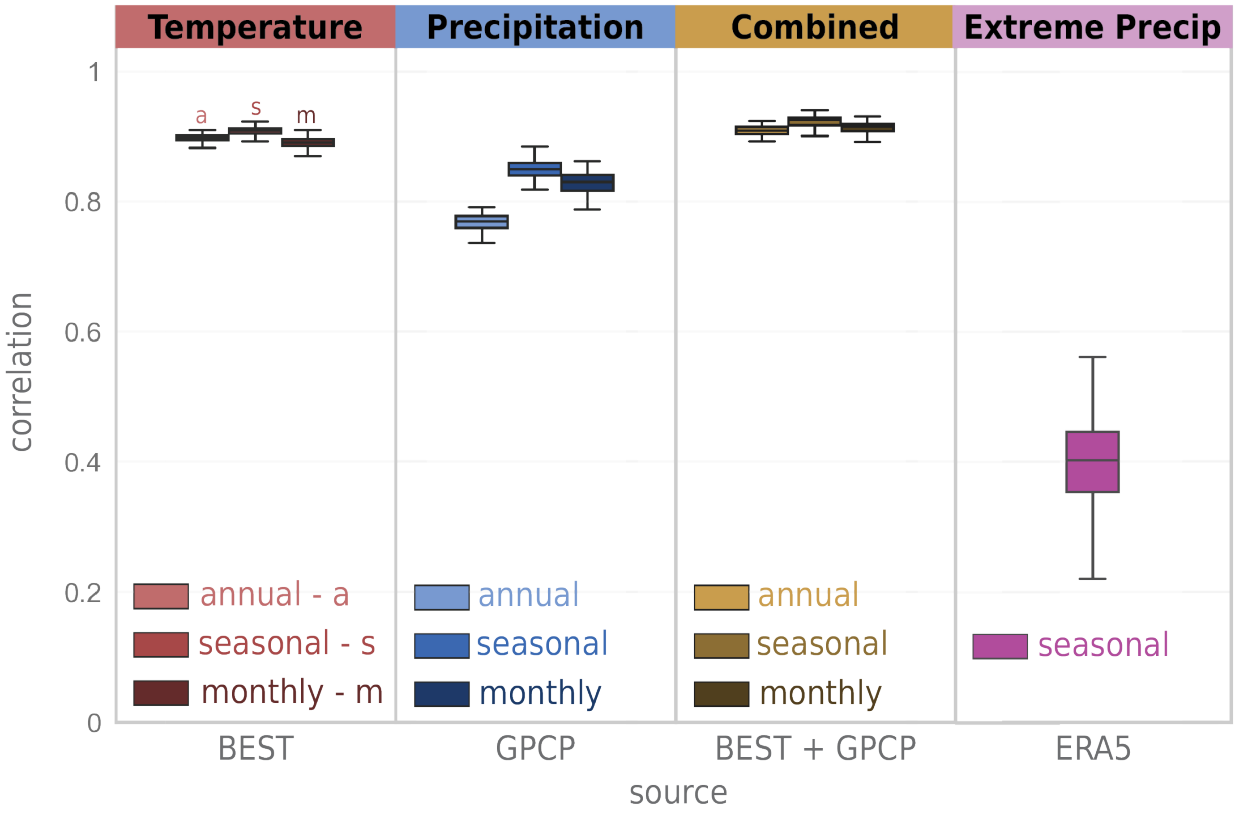

Figure 4.2: Correlation of actual years with predicted years for observations. Pearson correlations of the actual years with the years predicted by 100 trained neural networks given observations of temperature, precipitation, and extreme precipitation. Correlations were computed for all years beginning 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 minumum/maximum value, whichever is less extreme. Outliers are excluded for clarity, but can be found in Figures A.4 and A.5.

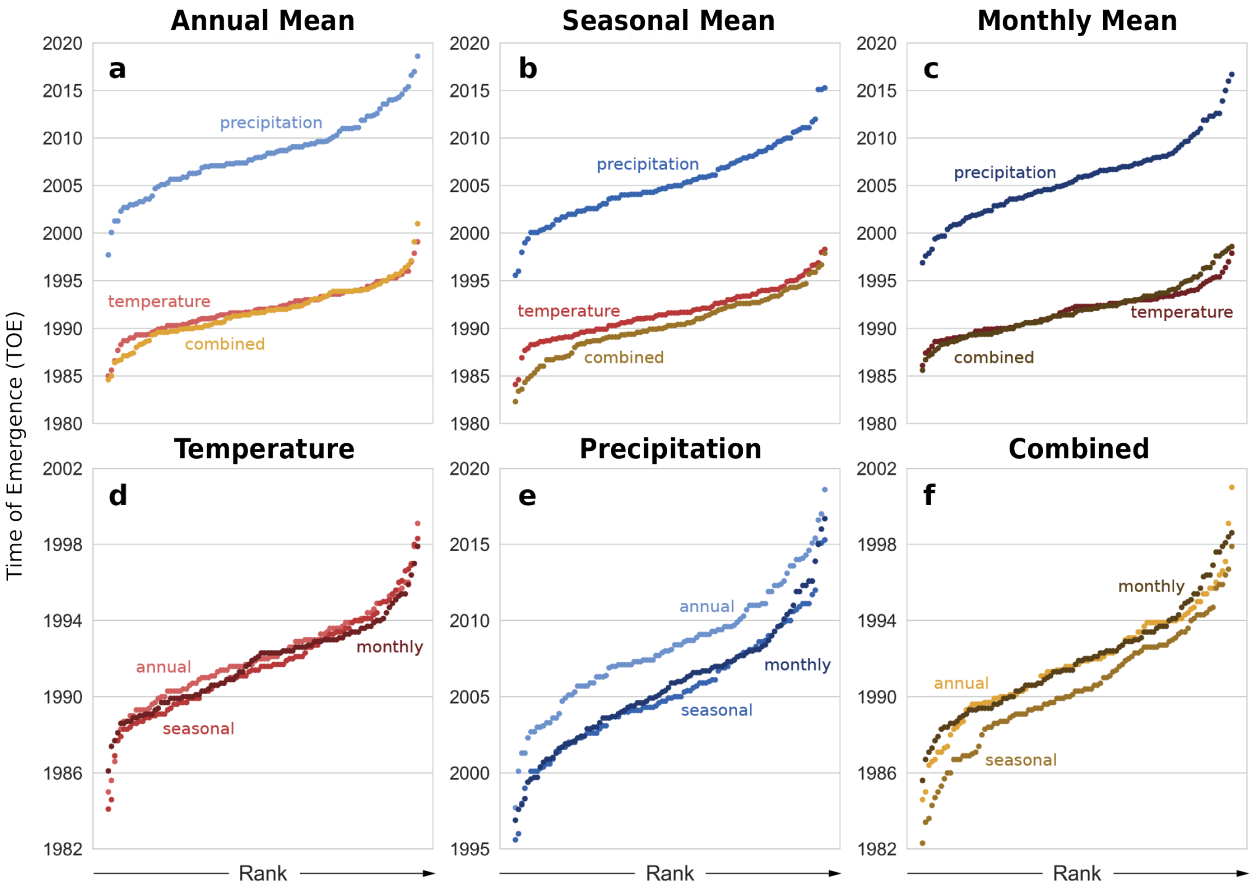

Figure 4.3: Mean TOE for each input field. Comparison of the mean time of emergence identified by neural networks trained on annual-mean (a), seasonal-mean (b), and monthly-mean (c) input fields, and neural networks trained on temperature (d), precipitation (e), and temperature and precipitation combined (f). 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.

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 4.1, 4.3), 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 2080 to 2098 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 Figure 4.4a-d (7a-d).

LRP identifies temperature over North Africa and Central Asia in JJA (Figure 4.4c) and the Andes and Central Africa in SON (Figure 4.4d) 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 (Figure 4.5a,d) and in Central Africa and India in JJA and SON (Figure 4.5c,d). 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 4.4 and 4.5, 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 in order to accurately predict the year. This is not surprising because the forced signal of temperature is clearer than the forced signal of precipitation (Fig. 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 4.3b). 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 4.4e-h, 4.5e-h, 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 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 21st 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 Figure 4.4e-h, and the corresponding maps for precipitation are shown in Figure 4.5e-h, where a higher S/N ratio (darker green) indicates a clearer forced signal. 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 4.4e-h and 4.5e-h.

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, Figure 4.5c,g or South Africa in SON, Figure 4.5d,h). 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, Figure 4.4a,e), 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, Figure 4.4d,h). These regions confirm that the indicator patterns identified by LRP capture more than the local S/N ratio. They also capture which regions improve detection of forced change only when coupled with signals in other regions, variables, and seasons. 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.

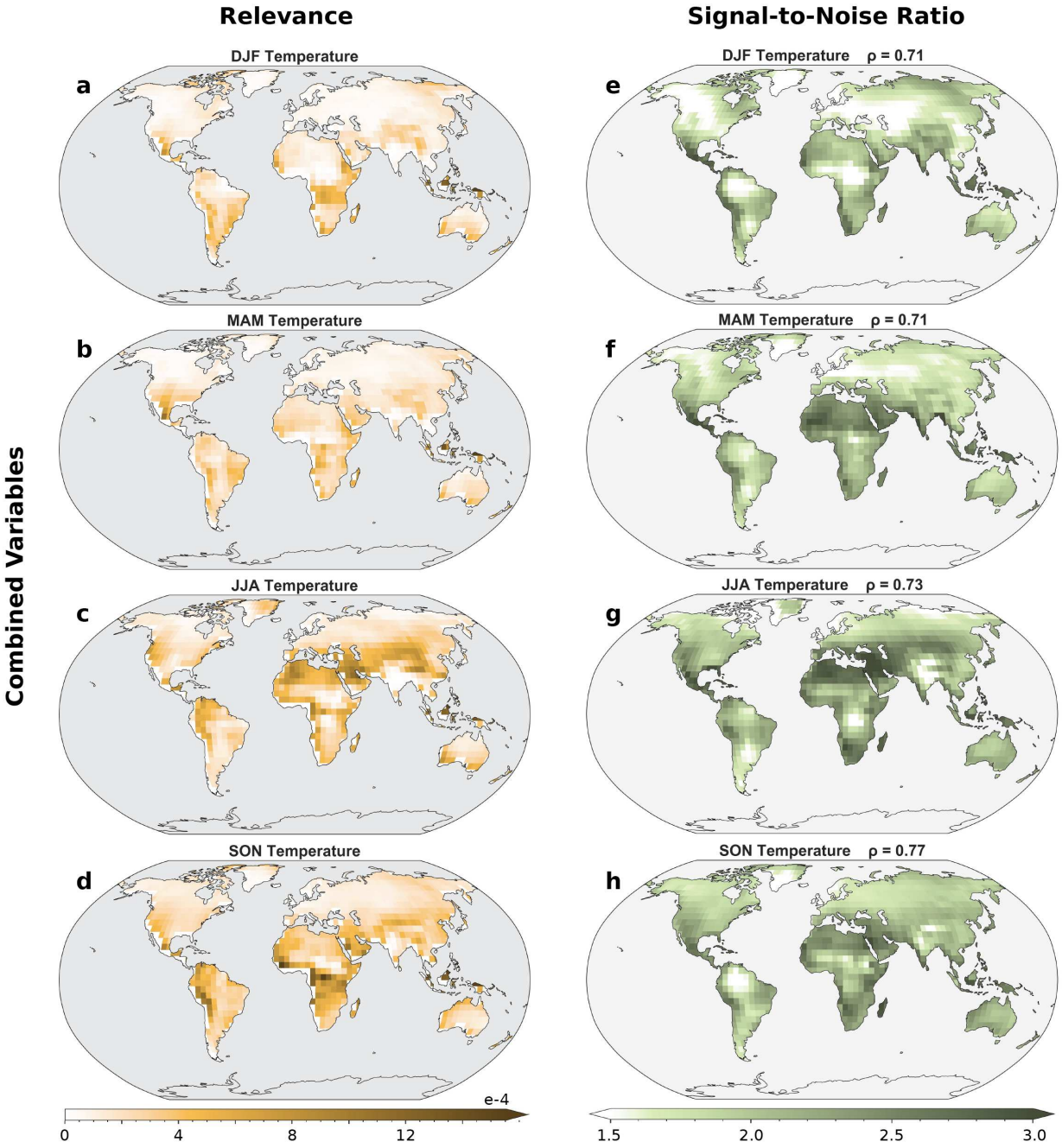

Figure 4.4: Combined indicator patterns of the forced response (temperature). Average temperature LRP results for the seasonal-mean combined neural networks when given maps 2080-2098 (left, in yellow) and S/N ratio for 2090 (right, in green). Darker shading indicates regions of temperature that are more relevant for the neural network's prediction or have a higher S/N ratio.

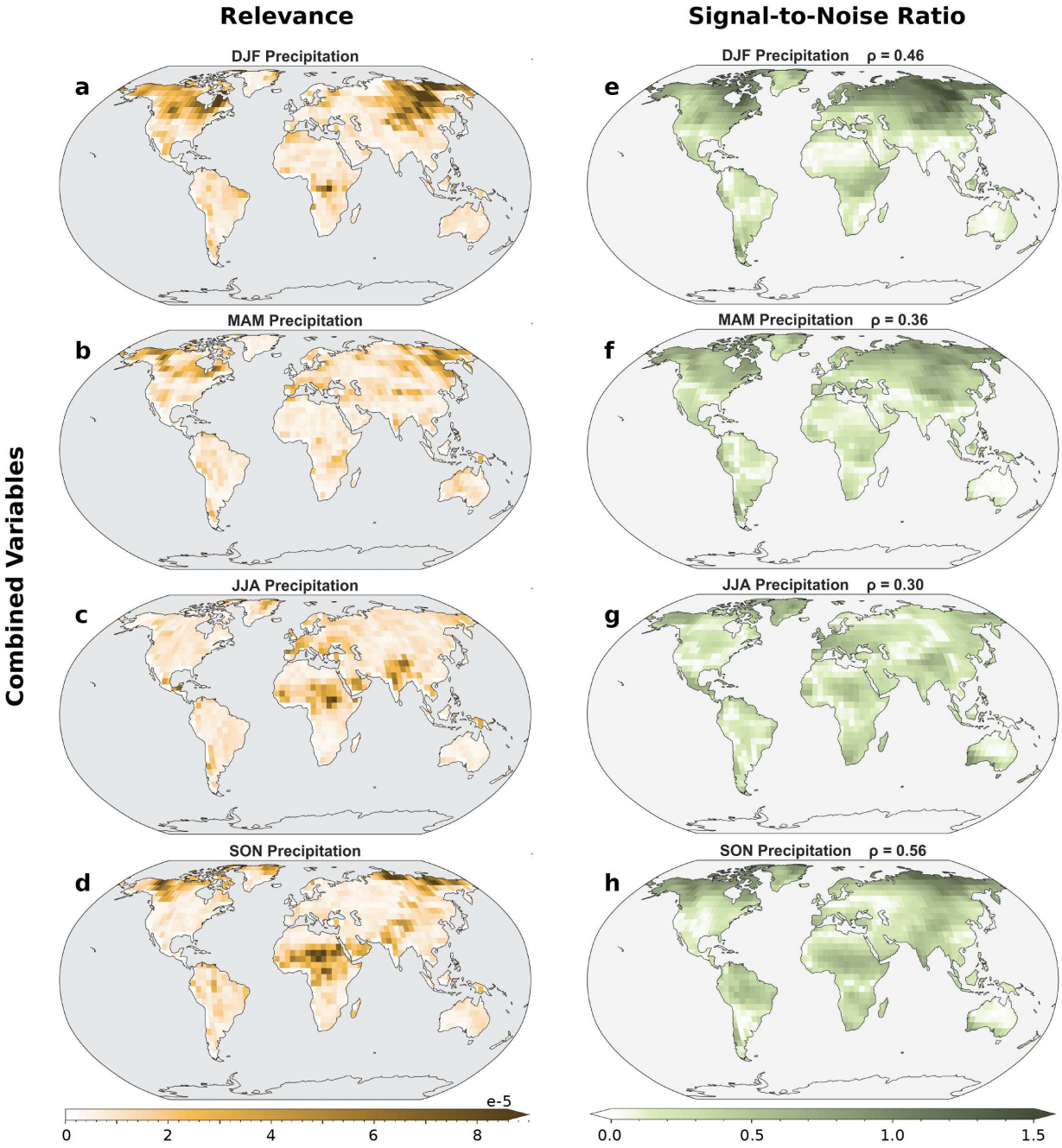

Figure 4.5: Combined indicator patterns of the forced response (precipitation). Average precipitation LRP results for the seasonal-mean combined neural networks when given maps 2080-2098 (left, in yellow) and S/N ratio for 2090 (right, in green). Darker shading indicates regions of precipitation that are more relevant for the neural network's prediction or have a higher S/N ratio.

Chapter 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 that this neural network approach can be extended to variables that have considerable noise (like extreme precipitation, see Figure A.7), and datasets 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 A.7). 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 5.1 depicts the years predicted by one neural network trained on seasonal-mean extreme precipitation. As in Figure 4.1, 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 4.2). 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 dataset used: ERA5. As shown in Figure A.5, 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 re-

gions 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 A.8). 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 (Figure 5.2a-d). These LRP maps exhibit a more even distribution in relevance across each region and season than the end-of-the-21st-century LRP maps of global temperature and precipitation (Figures 4.4a-d, 4.5a-d). 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 (Figure 5.2e-l, see the supplementary materials for more details on k-means processing). 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 (Figure 5.2m-p). 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 5.3).

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 and Barnes, 2021; Madakumbura et

al., 2021). Comparing the LRP maps at the TOE (Figure 5.4a-d) with those at the end of the 21st century (Figure 5.4e-h) highlights the regions that become more important for predicting the year over time. The difference plots in Figure 5.4i-l 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 21st century than the early 21st century, which may indicate abrupt change in the local climate at high latitudes, though further exploration is required. These time-varying patterns support the idea that combined indicators are effective for identifying dynamically evolving patterns of forced change.

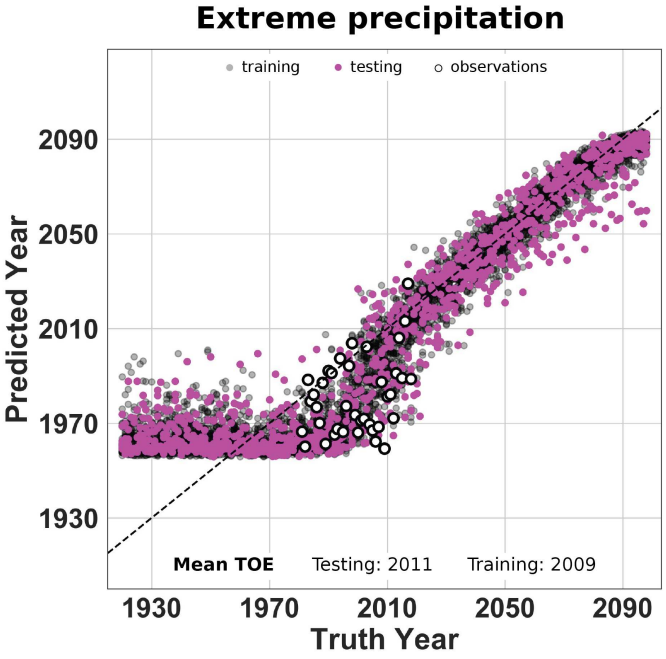

Figure 5.1: Neural network output for extreme precipitation. Year predicted by the neural network (y-axis) versus the truth year (x-axis) given seasonal-mean maps of extreme precipitation. Training data is shown in gray, testing data is shown in pink, and observations are shown in white.

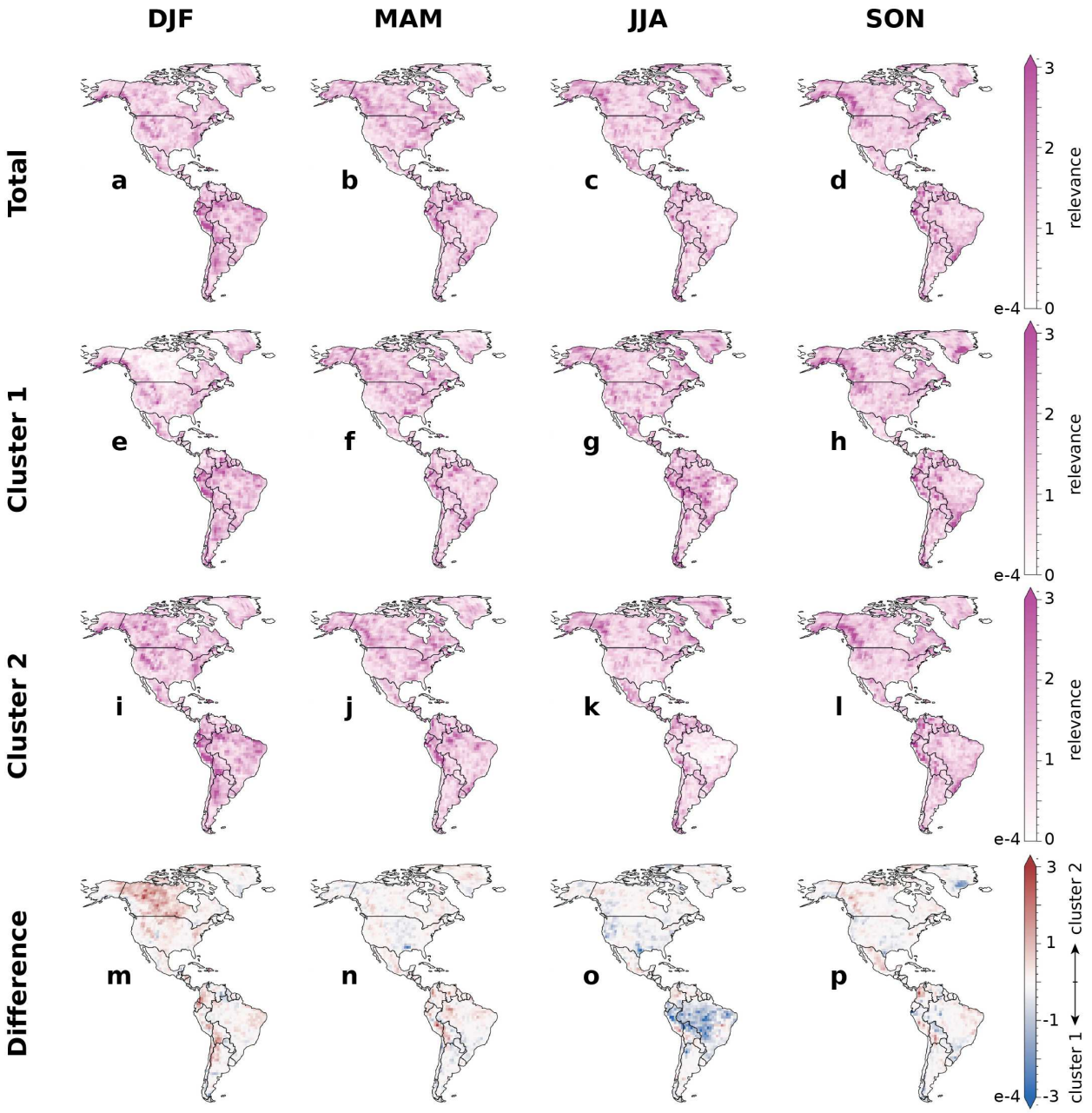

Figure 5.2: Relevance map clusters at the TOE for extreme precipitation. Average LRP results for: extreme precipitation at the TOE (a-d), each cluster identified by k-means (e-h, i-l), and the difference between the clusters (m-p). In panels a-l, darker shading indicates regions of extreme precipitation 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.

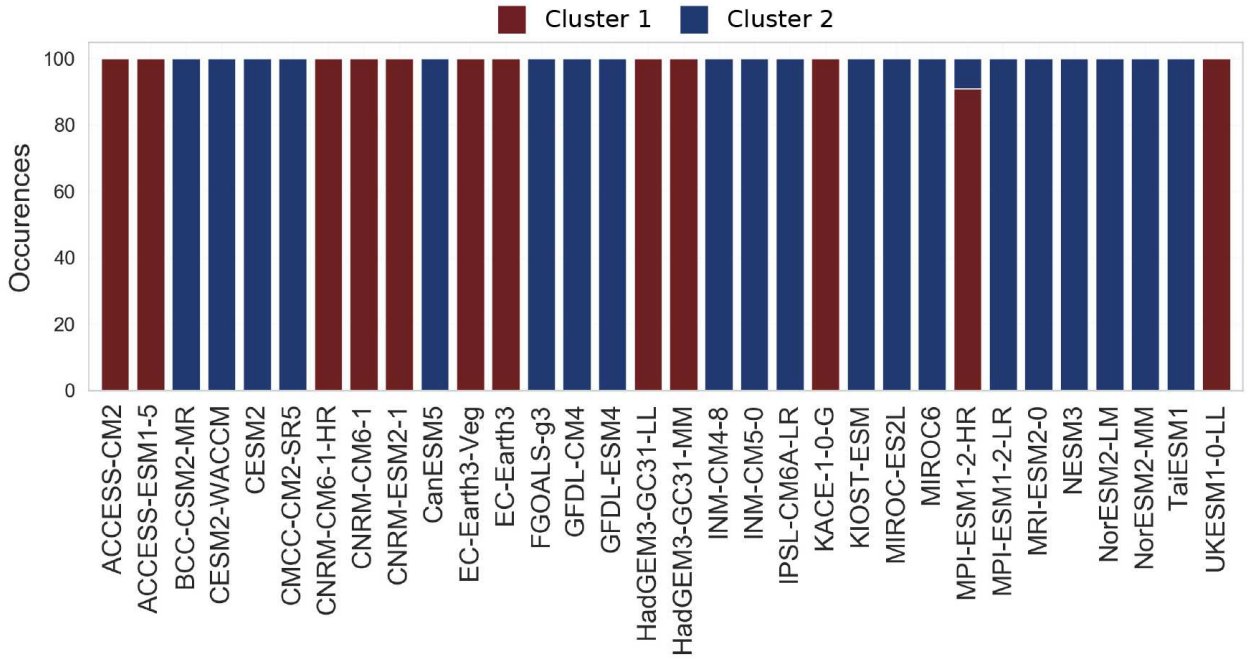

Figure 5.3: Climate models in each relevance map cluster at the TOE. The number of times each climate model appears in each cluster when k-means is applied to the maps of relevance at the TOE for 100 ANNs trained on extreme precipitation over the Americas. Only the relevance maps for MPI-ESM1-2-HR appear in both clusters. All other relevance maps for each climate model are found in one cluster or the other. Additionally, when k-means is instructed to identify 32 different clusters, each cluster contains exactly 100 relevance maps from a single climate model. This indicates that the relevance maps for a single climate model are not sensitive to whether the given model appeared in the training or testing set.

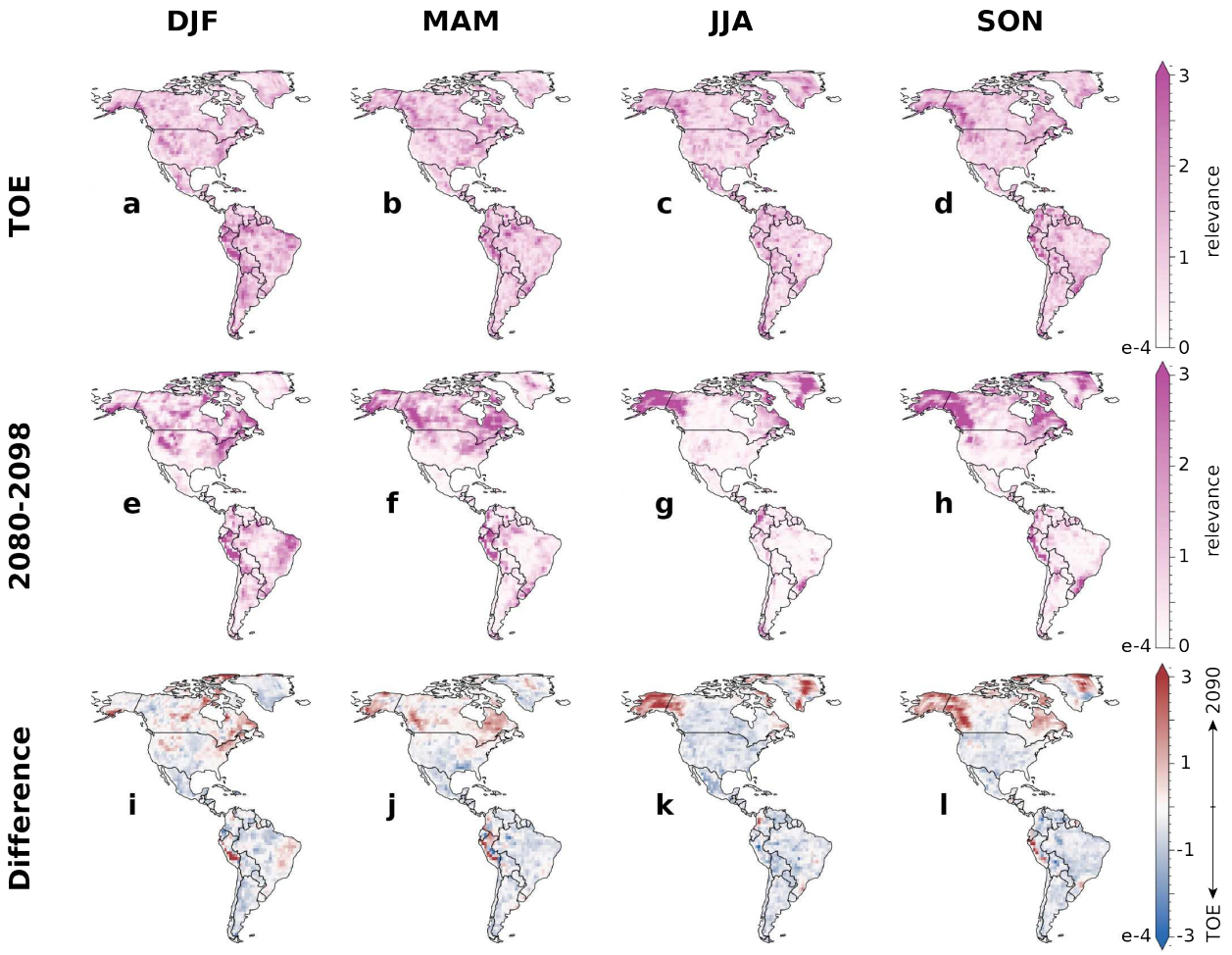

Figure 5.4: Time evolution of extreme precipitation relevance. Average LRP results at the TOE (a-d), 2080-2098 (e-h), and the difference between (i-l). Darker shading in panels a-h highlights regions that were more relevant for the neural networks' prediction of the year. In panels i-l, red shading indicates regions where the relevance has increased over time, while blue shading indicates regions where the relevance has decreased over time.

Chapter 6

Conclusions

When tasked with predicting the year given climate model simulations of temperature, precipitation, or extreme precipitation, artificial neural networks are able to learn indicator patterns of the forced response that allow them to distinguish between inputs from different years. The neural networks detected the forced response in combined fields earlier than in single fields through their ability to identify and utilize complex, nonlinear relationships between multiple variables and seasons. Using LRP, we visualized the reliable, multivariate patterns of forced change that were used by the neural networks. These indicator patterns vary both in time and between groups of climate models, suggesting that neural networks may be useful for examining inter-model differences in the nonlinear response of the climate system to external forcings. The regional indicator patterns identified by the neural networks are not necessarily the regions that will change most under a high-emissions, global warming scenario. Instead, they are the regions that combine to capture the signal of forced change most clearly across CMIP6 models. We have just begun to explore how neural networks can be used to study combined indicators of forced change. This framework is flexible, and should be expanded to other variables, regions of focus, and climate change scenarios, to identify the combined indicators that best elucidate the forced signal. Further application of this technique to climate extremes, such as heat wave frequency, drought indices, and flood risk may reveal that explainable neural networks are useful for assessing societal impacts and improving climate change preparedness as well.

References

Abiodun, O. I., A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, 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., and Coauthors, 2018: The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation. *Atmosphere* , *9*(4), https://doi.org/10.3390/atmos9040138.

Ali, H., P. Modi, and V. Mishra, 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., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, 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., J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2019: Viewing forced climate patterns through an AI lens. *Geophys. Res. Lett.*, *46*(22), 13389–13398, https://doi.org/10.1029/2019gl084944.

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

Barnett, T. P., and Coauthors, 2008: Human-induced changes in the hydrology of the west-

ern United States. *Science*, *319*(5866), 1080–1083, https://doi.org/10.1126/science.1152538.

Bell, B., and Coauthors, 2021: The ERA5 Global Reanalysis: Preliminary Extension to 1950. *Quart J Roy Meteor Soc*,.

Bindoff, N. L., and Coauthors, 2013: Chapter 10 - Detection and attribution of climate change: From global to regional. *Climate Change 2013: The Physical Science Basis. IPCC Working Group I Contribution to AR5*, Cambridge University Press.

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

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

Chen, D., A. Dai, and A. Hall, 2021: The convective-to-total precipitation ratio and the "drizzling" bias in climate models. *J. Geophys. Res.*, *126*(16), e2020JD034198, https://doi.org/10.1029/2020jd034198.

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

Donat, M. G., L. V. Alexander, N. Herold, and A. J. Dittus, 2016: Temperature and pre-

cipitation extremes in century-long gridded observations, reanalyses, and atmospheric model simulations. *J. Geophys. Res.*, *121*(19), 11,174–11,189, https://doi.org/10.1002/2016jd025480.

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

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

Field, C. B., and Coauthors, 2014: Summary for Policymakers. *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*, Cambridge University Press, 1–32.

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

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

Gettelman, A., D. J. Gagne, C.-C. Chen, M. W. Christensen, Z. J. Lebo, H. Morrison, and G. Gantos, 2021: Machine learning the warm rain process. *J. Adv. Model. Earth Syst.*, *13*(2), e2020MS002268, https://doi.org/10.1029/2020ms002268.

Hawkins, E., and R. Sutton, 2009: The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bull. Am. Meteorol. Soc.*, *90*(8), 1095–1108, https://doi.org/10.1175/2009BAMS2607.1.

Hawkins, E., and R. Sutton, 2012: Time of emergence of climate signals. *Geophys. Res. Lett.*, *39*(1), https://doi.org/10.1029/2011gl050087.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, *146*(730), 1999–2049, https://doi.org/10.1002/qj.3803.

Kim, Y.-H., S.-K. Min, X. Zhang, J. Sillmann, and M. Sandstad, 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., and E. A. Barnes, 2021: Detecting climate signals using explainable AI with single-forcing large ensembles. *J. Adv. Model. Earth Syst.*, *13*(6), e2021MS002464, https://doi.org/10.1029/2021ms002464.

Lagerquist, R., A. McGovern, and D. J. Gagne II, 2019: Deep Learning for Spatially Explicit Prediction of Synoptic-Scale Fronts. *Weather Forecast.*, *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.* Technischen Universität Berlin, https://doi.org/10.14279/DEPOSITONCE-7942.

Lee, Y., C. D. Kummerow, and I. Ebert-Uphoff, 2021: Applying machine learning methods to detect convection using Geostationary Operational Environmental Satellite-16 (GOES-16) advanced

baseline imager (ABI) data. *Atmos. Meas. Tech.*, *14*(4), 2699–2716, https://doi.org/10.5194/amt-14-2699-2021.

Li, J., D. W. J. Thompson, E. A. Barnes, and S. Solomon, 2017: Quantifying the Lead Time Required for a Linear Trend to Emerge from Natural Climate Variability. *J. Clim.*, *30*(24), 10179–10191, https://doi.org/10.1175/JCLI-D-16-0280.1.

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

Maher, N., S. Milinski, and R. Ludwig, 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.

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

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2021: Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset. arXiv preprint arXiv:2103.10005.

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

Marvel, K., and C. Bonfils, 2013: Identifying external influences on global precipitation. *Proc. Natl. Acad. Sci. U. S. A.*, *110*(48), 19301–19306, https://doi.org/10.1073/pnas.1314382110.

Montavon, G., A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, 2019: Layer-Wise Relevance Propagation: An Overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, and K.-R. Müller, Eds., Vol. 11700 of, Springer International Publishing, 193–209.

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

North, G. R., and M. J. Stevens, 1998: Detecting Climate Signals in the Surface Temperature Record. *J. Clim.*, *11*(4), 563–577, https://doi.org/10.1175/1520-0442(1998)011<0563:DCSITS>2.0.CO;2.

O'Neill, B. C., and Coauthors, 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., and Z. Hausfather, 2020: The Berkeley Earth land/ocean temperature record. *Earth Syst. Sci. Data*, *12*(4), 3469–3479, https://doi.org/10.5194/essd-12-3469-2020.

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

Sanderson, B. M., K. W. Oleson, W. G. Strand, F. Lehner, and B. C. O'Neill, 2018: A new

ensemble of GCM simulations to assess avoided impacts in a climate mitigation scenario. *Clim. Change*, *146*(3), 303–318, https://doi.org/10.1007/s10584-015-1567-z.

Santer, B. D., and Coauthors, 1996: A search for human influences on the thermal structure of the atmosphere. *Nature*, *382*(6586), 39–46, https://doi.org/10.1038/382039a0.

Santer, B. D., and Coauthors, 2011: Separating signal and noise in atmospheric temperature changes: The importance of timescale. *J. Geophys. Res.*, *116*(D22), https://doi.org/10.1029/2011jd016263.

Santer, B. D., J. C. Fyfe, S. Solomon, J. F. Painter, C. Bonfils, G. Pallotta, and M. D. Zelinka, 2019: Quantifying stochastic uncertainty in detection time of human-caused climate signals. Proc. Natl. Acad. Sci. U. S. A., 116, 19821–19827, https://doi.org/10.1073/pnas.1904586116.

Scaife, A. A., and D. Smith, 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., and I. M. Held, 2001: Discriminants of Twentieth-Century Changes in Earth Surface Temperatures. *J. Clim.*, *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*: Hamburg, Germany,.

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

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

Solomon, A., and M. Newman, 2012: Reconciling disparate twentieth-century Indo-Pacific ocean temperature trends in the instrumental record. *Nat. Clim. Chang.*, *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. *J. Appl. Meteorol. Climatol.*, *26*(10), 1401–1405, https://doi.org/10.1175/1520-0450(1987)026<1401:TFCCAA>2.0.CO;2.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*(1), 1929–1958.

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

Toms, B. A., E. A. Barnes, and I. Ebert-Uphoff, 2020: Physically interpretable neural networks for the geosciences: Applications to earth system variability. *J. Adv. Model. Earth Syst.*, *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*. D.R. Reidmiller, C.W. Avery, D.R. Easterling, K.E. Kunkel, K.L.M. Lewis, T.K. Maycock, and B.C. Stewart, Eds. U.S. Global Change Research Program,.

Weyn, J. A., D. R. Durran, and R. Caruana, 2020: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.*, *12*(9), e2020MS002109, https://doi.org/10.1029/2020ms002109.

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

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

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

Zappa, G., B. J. Hoskins, and T. G. Shepherd, 2015: Improving Climate Change Detection through Optimal Seasonal Averaging: The Case of the North Atlantic Jet and European Precipitation. *J. Clim.*, *28*(16), 6381–6397, https://doi.org/10.1175/JCLI-D-14-00823.1.

Appendices

Appendix A Neural Network Specifications

All units of the neural networks use a rectified linear unit (ReLU) activation function, except for the output layer which uses a soft-max layer to rescale the final outputs of the neural network such that they sum to one. We train the neural networks using the binary cross-entropy loss between the predicted class likelihoods and the correct class membership weights, such that the loss function is minimized when the two are equal. More information on the ReLU activation function, the soft-max layer, and the loss function can be found in sections A1, A2, and A3 of Barnes et al. (2020), respectively.

The neural networks were trained using the Keras Adam optimizer, an adaptive stochastic gradient descent algorithm (Kingma and Ba, 2014). We used a learning rate that started at 0.001 and decayed linearly to 0.0005 over the span of 150 epochs. Although the Adam optimizer is designed to alter the learning rate based on the momentum of training, the decaying learning rate allowed the neural networks to train more quickly with improved performance. Weights and biases were initialized using random, normal values.

As discussed in Section 3.2 and Figure A.3, we applied a ridge penalty (L2 regularization) to the input layer (see Barnes et al., 2020). The ridge penalty was selected such that the time of emergence detected by the neural networks was the earliest. All input vectors used a ridge penalty of 0.1, except for seasonal-mean temperature and precipitation combined input vector, for which the TOE was earlier for a ridge penalty of 0.01 (see Figure A.3).

Appendix B K-means Clustering

Before applying k-means clustering, all LRP maps are converted into binary maps. Every grid point on each LRP map is assigned a one or a zero depending on whether its relevance value is greater than or less than the mean relevance across all maps and grid points. In this way, ones indicate regions of high relevance, and zeros indicate regions of low relevance. K-means clustering is then applied to these binary LRP maps (3200 in total, samples from 32 climate models for 100 neural networks).

Appendix C Additional Observational Datasets

In addition to the observational datasets in Section 2.2, we also test two additional precipitation observations in Figures A.4 and A.5. First, we use the European Center for Medium-Range Weather Forecasts' ERA5 global reanalysis (Hersbach et al., 2020) at 6-hour resolution to construct observational monthly mean precipitation fields from 1980 to the present. Second, we use the Japan Meteorological Agency's Japanese 55-year Reanalysis (JRA55; Kobayashi et al., 2015) mean 3-hour precipitation forecasts to construct observational monthly mean precipitation fields from 1959 to the present.

Appendix D Supplementary Figures

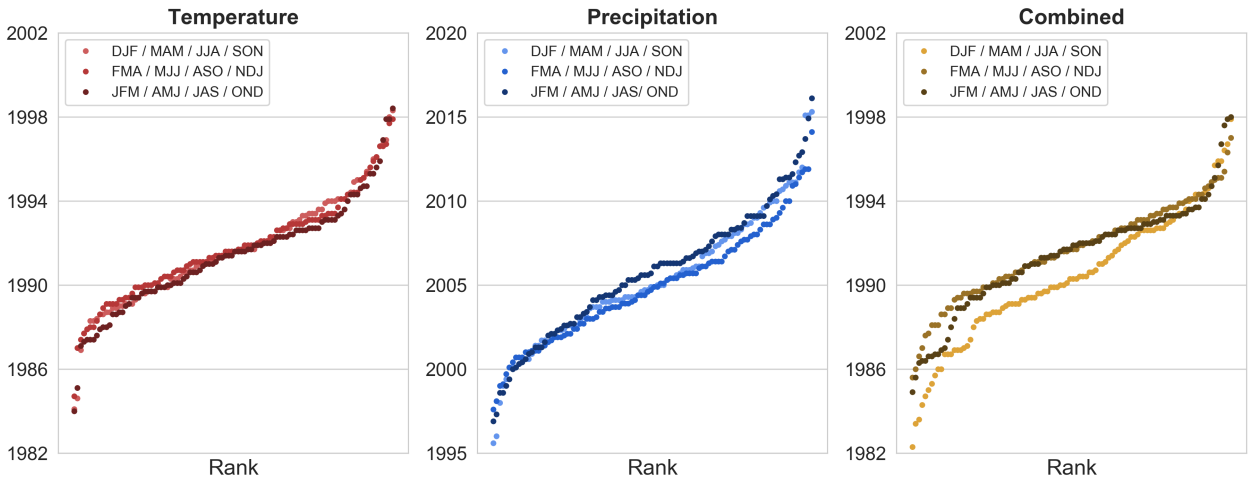

Figure A.1: TOE detected by the neural networks given different definitions of season. As in Figure 4.3, but for each possible three-month combination of seasons. All three definitions lead to similar TOE when neural networks are trained on global maps of temperature or precipitation. When temperature and precipitation are combined, meteorological seasons lead to the earliest detection of forced change.

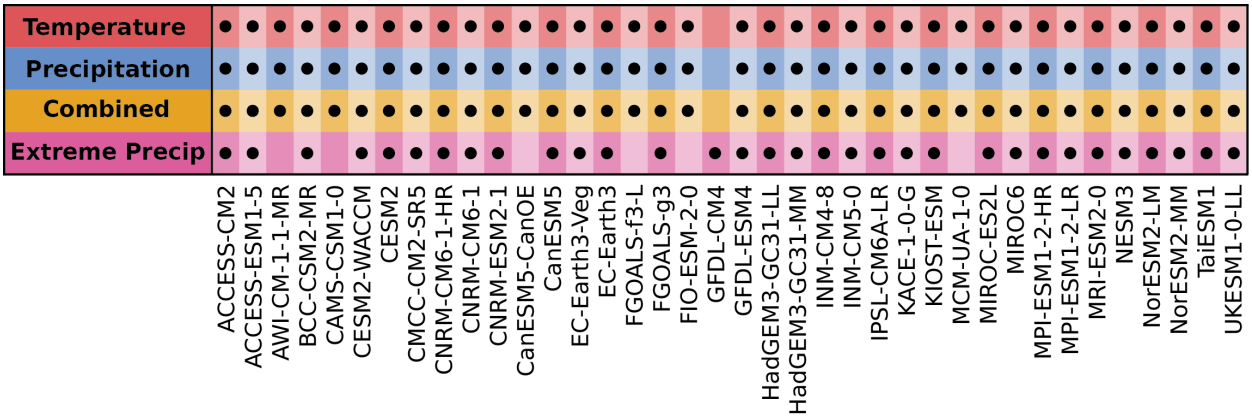

Figure A.2: Climate models used for each input variable. Temperature, precipitation, and temperature and precipitation combined used the same 37 CMIP6 climate models. Extreme precipitation fields came from 32 climate models for which daily precipitation fields were available.

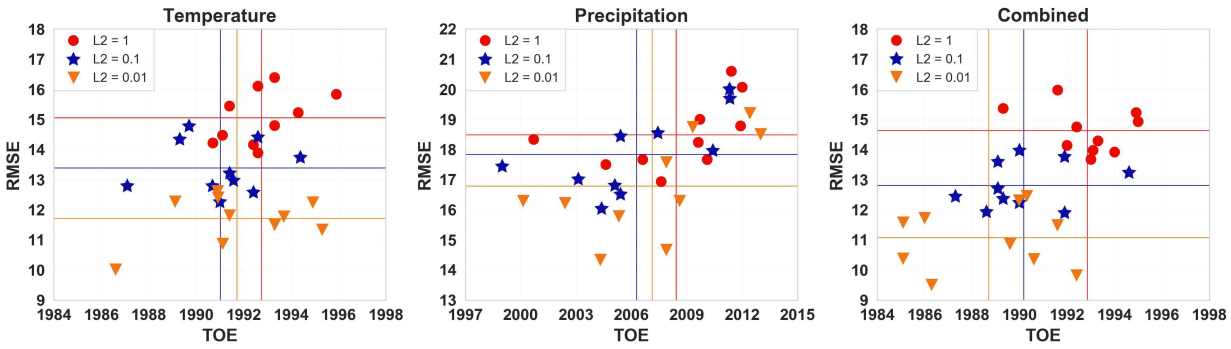

Figure A.3: TOE and RMSE for various ridge penalties. The sensitivity of RMSE and TOE to the ridge (L2) penalty used for 10 neural networks trained on seasonal-mean maps of (a) temperature, (b) precipitation, and (c) temperature and precipitation combined. Each plot shows the RMSE and TOE for neural networks trained with a ridge penalty of 1, 0.1, and 0.01 (denoted by red circles, blue stars, and orange triangles, respectively). The mean RMSE and TOE for all 10 neural networks are indicated by the horizontal and vertical lines. Each neural network for a given variable/ridge penalty differs only in which climate models were part of the training and testing sets. While a ridge penalty of 0.01 leads to the smallest mean RMSE in all cases, using a higher ridge penalty of 0.1 leads to earlier detection of forced change for temperature and precipitation input vectors. As a result, we choose to use the ridge penalties corresponding to an earlier TOE.

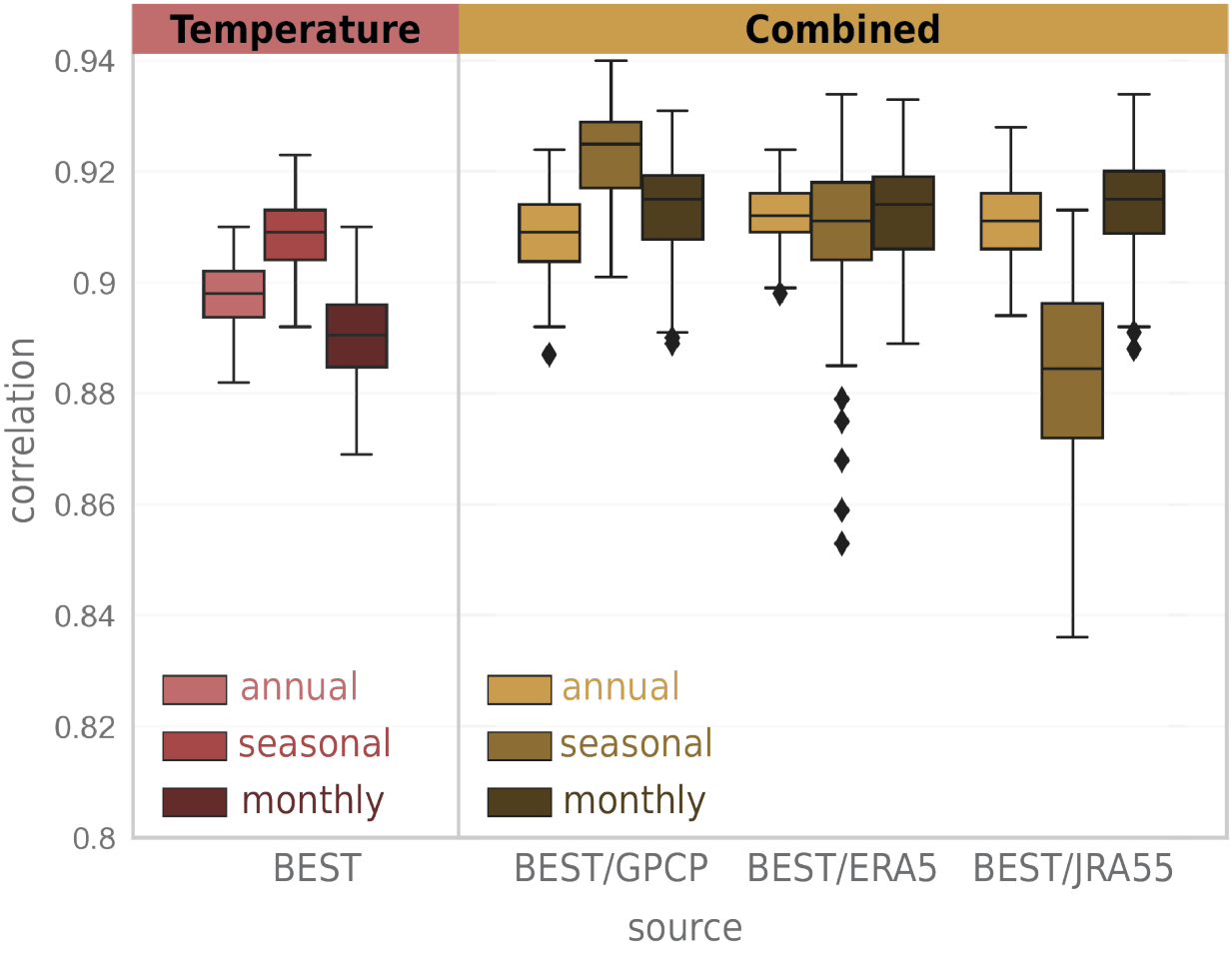

Figure A.4: Sensitivity of observational correlations to the source of precipitation observations: temperature and precipitation combined. Pearson correlations of the actual years with the years predicted by 100 trained neural networks given observations of temperature and precipitation. Correlations 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 minumum/maximum value, whichever is less extreme. The observational correlations for seasonal-mean combined neural networks are sensitive to the dataset of choice, as observational correlations are higher for GPCP than ERA5 or JRA55. This is not the case for the annual-mean and monthly-mean combined neural networks, which have approximately the same correlations regardless of the source of the observations. This is because the seasonal-mean combined neural networks rely on precipitation to predict the year, while the annual-mean and monthly-mean combined neural networks do not, as shown in Figure 4.3.

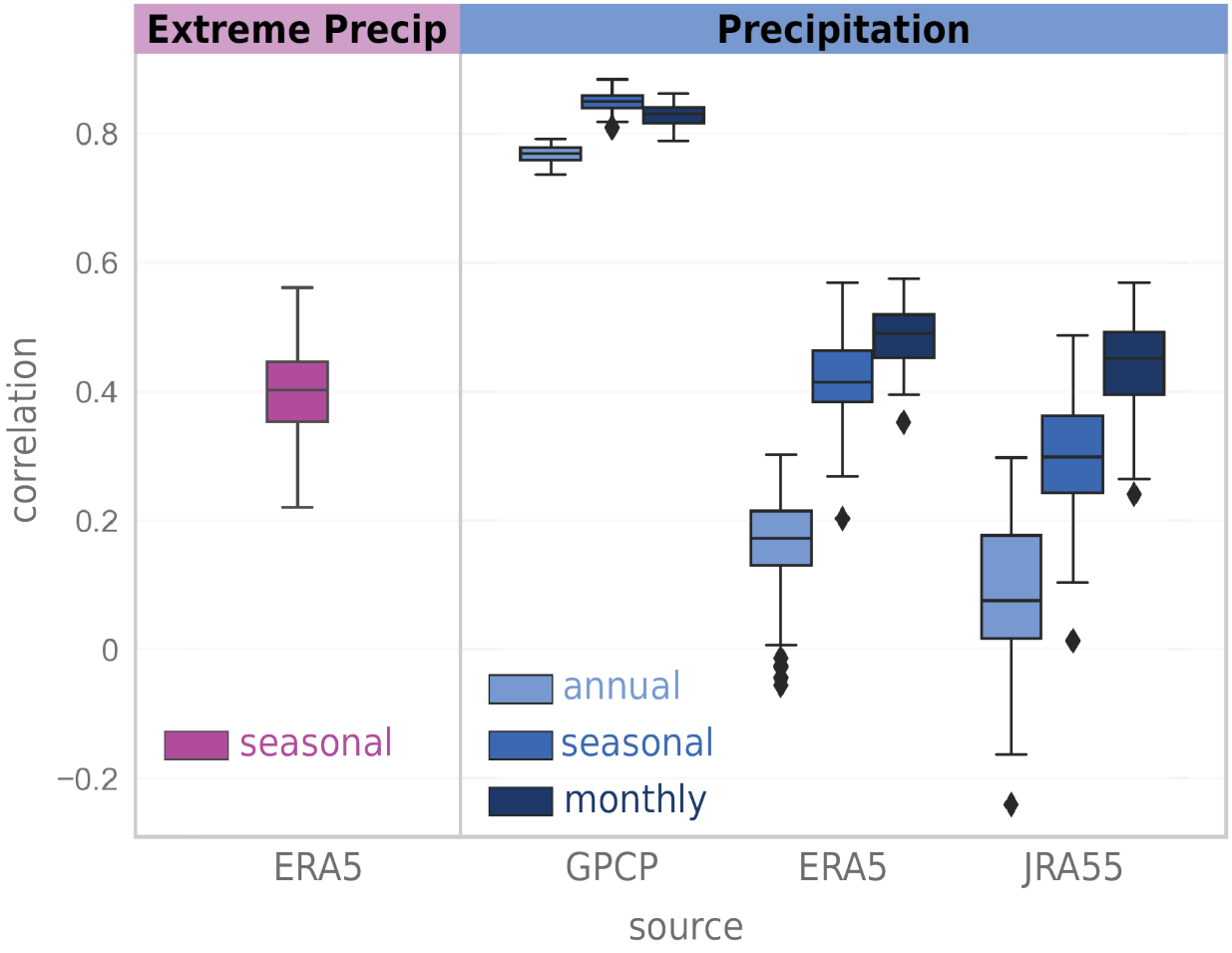

Figure A.5: Sensitivity of observational correlations to the source of precipitation observations: precipitation only. Pearson correlations of the actual years with the years predicted by 100 trained neural networks given observations of precipitation. Correlations were computed for all years beginning 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 minumum/maximum value, whichever is less extreme. The observational correlations are sensitive to the source of precipitation data. Correlations are highest for GPCP, followed by ERA5 and JRA55. The observational correlations for ERA5 seasonal-mean extreme precipitation are similar to those for ERA5 seasonal-mean precipitation.

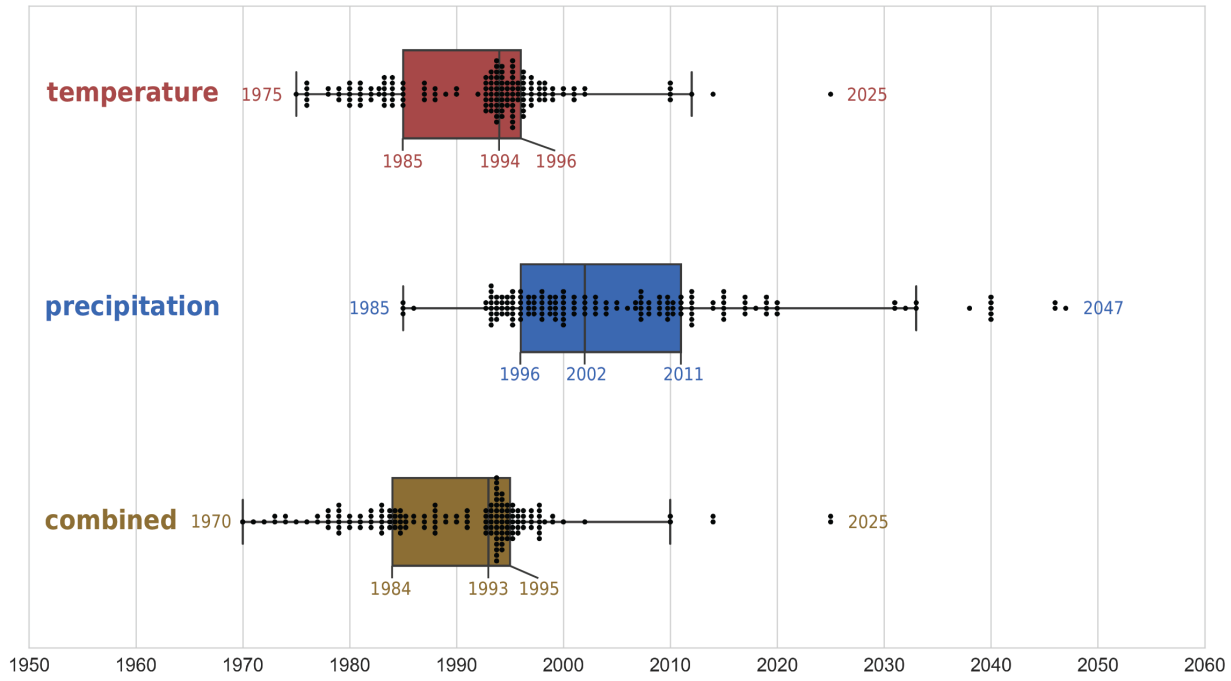

Figure A.6: Time of emergence for seasonal-mean fields. TOE was calculated for each climate model in the testing sets of 100 trained neural networks. Each dot represents five (rounded up) occurrences of the associated TOE year (i.e. one dot represents 1-5 occurrences, two dots represent 6-10 occurrences, and so on). For added clarity, box plots indicate the first, second, and third quartiles of the TOEs for each model, and whiskers denote 1.5 times the interquartile range, or the minimum/maximum point, whichever is less extreme.

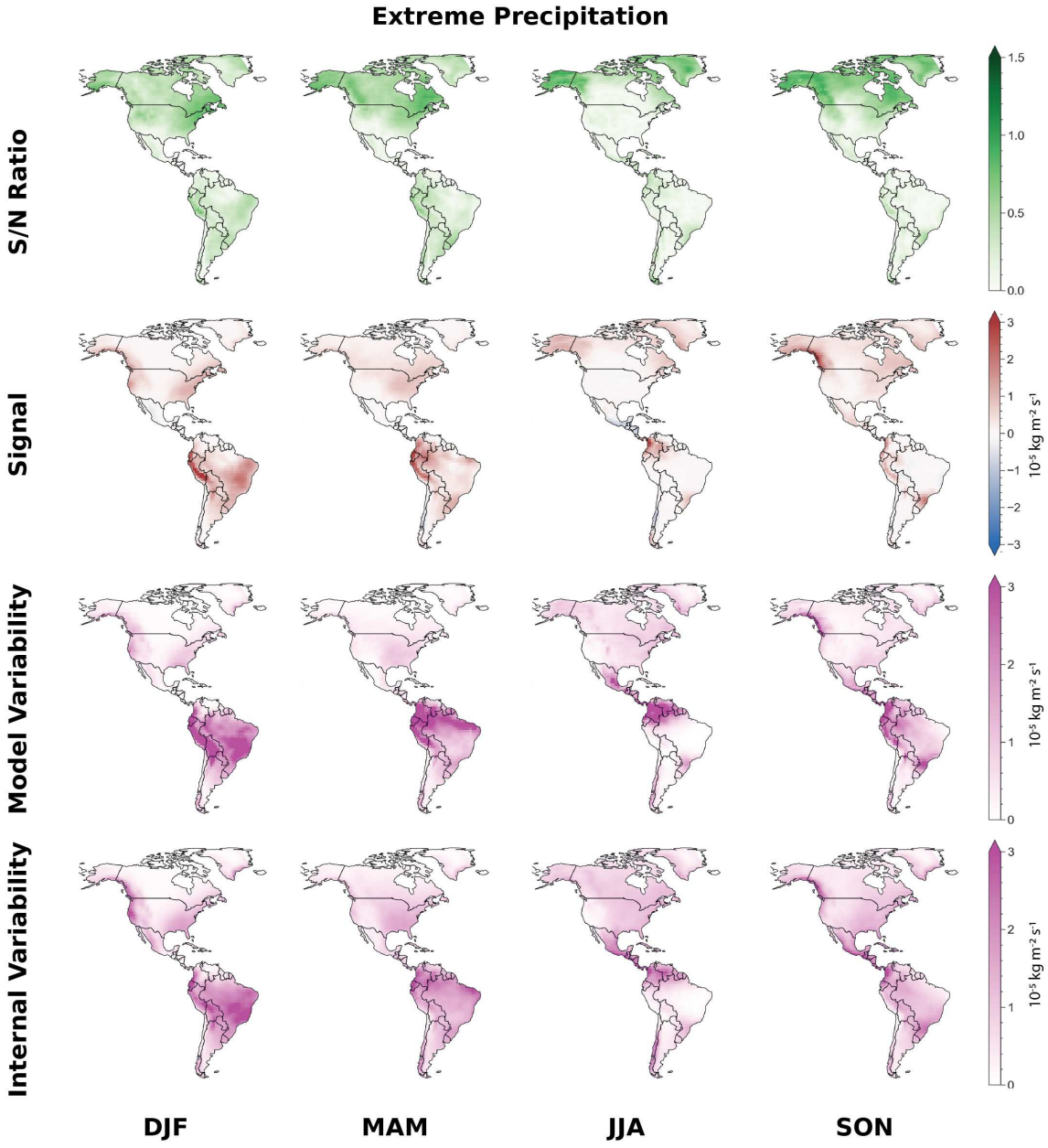

Figure A.7: Signal and noise for extreme precipitation over the Americas. Plots of S/N ratio, and its components (signal, climate model variability, and internal variability) for extreme precipitation in each season over North and South America. The signal is most clear over the northern-most latitudes. The S/N ratio is below 1.5 in all seasons indicating that there is considerable noise relative to the signal of change.

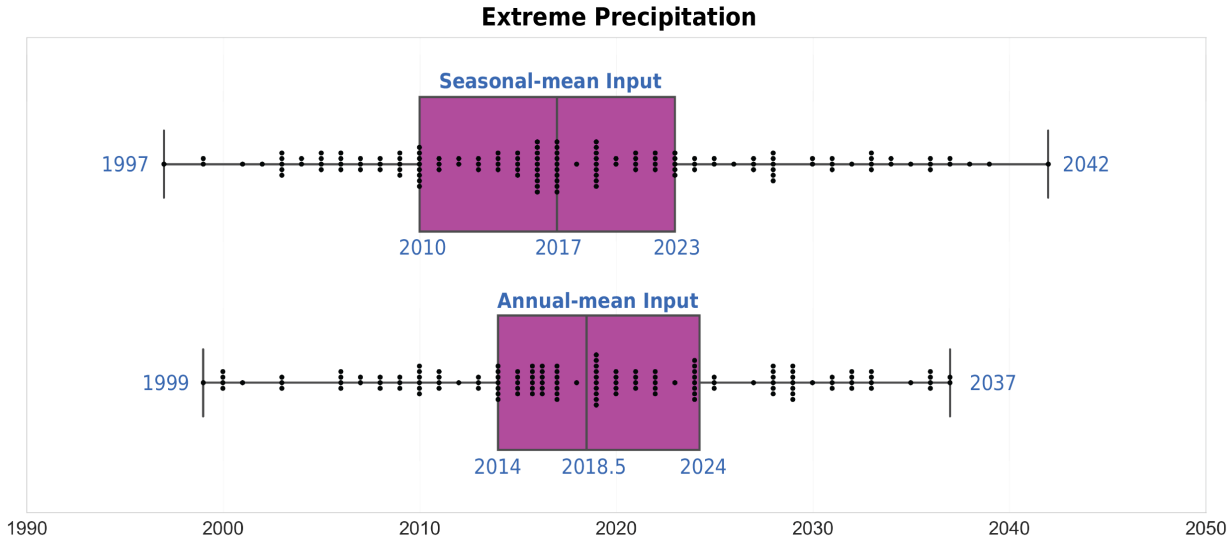

Figure A.8: Time of emergence for extreme precipitation over the Americas. TOE was calculated for each climate model in the testing sets of 100 trained neural networks. Each dot represents five (rounded up) occurrences of the associated TOE year (i.e. one dot represents 1-5 occurrences, two dots represent 6-10 occurrences, and so on). For added clarity, box plots indicate the first, second, and third quartiles of the TOEs for each model, and whiskers denote 1.5 times the interquartile range, or the minimum/maximum point, whichever is less extreme.

Appendix E Supplementary References

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

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, *146*(730), 1999–2049, https://doi.org/10.1002/qj.3803.

Kingma, D. P., and J. Ba, 2014: Adam: A Method for Stochastic Optimization. *arXiv* [cs.LG],.

Kobayashi, S., and Coauthors, 2015: The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *Journal of the Meteorological Society of Japan*, *93*(1), 5–48, https://doi.org/10.2151/jmsj.2015-001.