# Changes in United States summer temperatures revealed by explainable neural networks

**Zachary M. Labe[1], Nathaniel C. Johnson[2], and Thomas L. Delworth[2]**

[1]Atmospheric and Oceanic Sciences Program, Princeton University, NJ, USA
[2]NOAA/OAR/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

**Key Points:**

- Forced temperature signals have emerged in observations during summer in the United States as detected by an artificial neural network
- Increasing spatial resolution improves neural network skill for predicting the year of a given summer temperature map
- Trends in western United States land surface fields are linked to an early timing of emergence in a state-of-the-art climate model

Corresponding author: Zachary M. Labe, zachary.labe@noaa.gov

## Abstract

To better understand the regional changes in summertime temperatures across the conterminous United States (CONUS), we adopt a recently developed machine learning framework that can be used to reveal the timing of emergence of forced climate signals from the noise of internal climate variability. Specifically, we train an artificial neural network (ANN) on seasonally-averaged temperatures across the CONUS and then task the ANN to output the year associated with an individual map. In order to correctly identify the year, the ANN must therefore learn time-evolving patterns of climate change amidst the noise of internal climate variability. The ANNs are first trained and tested on data from large ensembles and then evaluated using observations from a station-based dataset. To understand how the ANN is making its predictions, we leverage a collection of ad hoc feature attribution methods from explainable artificial intelligence (XAI). We find that anthropogenic signals in seasonal mean minimum temperature have emerged by the early 2000s for the CONUS, which occurred earliest in the Eastern United States. While our observational timing of emergence estimates are not as sensitive to the spatial resolution of the training data, we find a notable improvement in ANN skill using a higher resolution climate model, especially for its early 20th century predictions. Composites of XAI maps reveal that this improvement is linked to temperatures around higher topography. We find that increases in spatial resolution of the ANN training data may yield benefits for machine learning applications in climate science.


## Plain Language Summary

While temperatures around the world continue to warm due to human-caused climate change, some areas have observed smaller temperature trends than others. Understanding this regional variability in the rate of warming is important when assessing future projections. One location that has observed less warming is across the United States during their summer season. To evaluate temperature variability in this region using real-world observations and climate model simulations, we use a statistical method from artificial intelligence called neural networks. The goal of the neural network setup is to learn temperature patterns across the United States and then identify whether climate change effects have exceeded the range of natural variability that has occurred in the past. This is called the timing of emergence, which is the first year that the effect has clearly appeared. We find that the average United States minimum temperature increase has al-

ready emerged in historical records. However, we find no timing of emergence for the average maximum temperature, other than in the Western United States. Another important finding of this study is that by using higher resolution climate model data (i.e., more latitude and longitude points), we find better accuracy in the neural network predictions.

## 1 Introduction

The detection of a forced signal rising above the background of internal climate variability - often referred to as the 'timing of emergence' (ToE) - can be a potentially useful metric for societal and ecological planning to account for changes in weather and climate that exceed the known variability over a particular region (IPCC et al., 2021). Quantifying the ToE for a number of variables has been widely addressed within climate science research over the last few decades (e.g., Giorgi & Bi, 2009; Hawkins & Sutton, 2012; Mahlstein et al., 2012; Hawkins et al., 2014; Fischer & Knutti, 2014; King et al., 2015; Mora et al., 2013; Schlunegger et al., 2020; Hawkins et al., 2020; Satoh et al., 2022), but its precise definition is still quite sensitive to the choice of dataset, future greenhouse emission scenario, baseline reference period, spatial scale, temporal filtering, consideration of internal climate variability, and statistical testing for ToE consistency. The ToE can even be examined through a lens of compound events and combined variables (Mahony & Cannon, 2018; Rader et al., 2022; François & Vrac, 2023). While it is often implied to be associated with anthropogenic climate change, the ToE definition can also be influenced by multidecadal variability in the climate system (Lehner et al., 2017). To partially alleviate this issue, recent ToE work (e.g., Rodgers et al., 2015; Lehner et al., 2017; Schlunegger et al., 2020; Wyser et al., 2021) has examined the use of initial condition large ensembles which better isolate the role of uncertainty due to internal variability within a single global climate model (GCM) system. This is done to avoid conflating different uncertainties, such as when calculating a multi-model mean (Hawkins et al., 2014; Lehner et al., 2020). In this context, the spread of ToE estimates can be compared across individual ensemble members (i.e., range of internal variability) relative to the overall ensemble mean (i.e., radiatively forced signal).

One example of an extratropical region that experiences large variances in simulated decadal temperature trends by these climate model large ensembles is across the contiguous United States (Deser et al., 2012; Milinski et al., 2020; Lehner & Deser, 2023). While this variability is often analyzed for boreal winter months (McKinnon & Deser,

2018; Deser et al., 2016; B. Yu et al., 2020), substantial regional variability in CONUS summertime temperature trends has also been observed (Program, 2018). Of particular interest is a portion of the Southeast to Central CONUS in which little to no warming has been observed (Kunkel et al., 2006), or even has slightly cooled depending on the choice of trend period, duration, or observational dataset (Grotjahn & Huynh, 2018; Partridge et al., 2018). This 'warming hole' has been most prominently found in trends of daytime maximum temperatures (Pan et al., 2004; Mueller et al., 2016). It remains unclear to what extent this persistent CONUS warming hole is driven by an unusual realization of internal climate variability (a. Meehl et al., 2012; Meehl et al., 2015), land-use surface feedbacks (Mueller et al., 2016; Alter et al., 2018), changes in anthropogenic aerosols (Leibensperger et al., 2012; S. Yu et al., 2014; Banerjee et al., 2017), an indirect response to external greenhouse gas forcing (Eischeid et al., submitted), or a combination of several of these factors (Pan et al., 2004; Mascioli et al., 2017). In contrast, the Western United States has observed larger warming trends during the last few decades, which has contributed to the formation of prolonged drought risk and favorable environments for wildfire ignition (Diffenbaugh et al., 2015; Abatzoglou & Williams, 2016; Parks & Abatzoglou, 2020; Williams et al., 2020). Though the ratio of new record high temperatures compared to record lows continues to widen (Meehl et al., 2022), the overall detectability of CONUS temperature signals continues to remain challenging, partially due to the anomalous warmth observed in the Dust Bowl era (Hansen et al., 2001; Peterson et al., 2013; Donat et al., 2016). Given the broad range of consequences associated with future projected warming over the CONUS (Wuebbles et al., 2014; Program, 2018), it remains urgent to better characterize the ToE of summertime mean extreme temperatures in order to better aid in future decision-making on regional health hazards and other impacts that could fall outside of historical climate variability (Mankin et al., 2020; Deser, 2020; Schwarzwald & Lenssen, 2022; Bevacqua et al., 2023).

The aim of this study is to evaluate whether patterns of radiatively-forced temperature change have emerged across the CONUS in summer, despite substantial internal variability and only a marginally positive overall mean temperature trend in the recent observational record. To confront these challenges, we turn to a novel explainable artificial neural network (ANN) approach (Barnes et al., 2018, 2019) that can spatially leverage temperature signals across given geographic maps for identifying the ToE of individual climate variables. The advantage is that we are not limited to traditional signal-

to-noise metrics that only consider point-by-point statistics for disentangling the role of internal variability from anthropogenic climate change, because here the ANNs can instead potentially learn nonlinear relationships in the data. We also apply this framework to different regions of the CONUS, which supports that the Central United States warming hole has contributed to a lack of visible anthropogenic signal in mean observed summertime temperatures. Although this machine learning ToE indicator may not necessarily reflect local conditions, such as those considered for adaptation planning for ecosystem services (Weiskopf et al., 2020) or for rapid changes to biodiversity (Henson et al., 2017; Trisos et al., 2020), the spatial aggregate metric as designed here remains relevant for industries that assess climate risk over larger geographic domains (e.g., Lempert et al., 2004; Mills, 2005; Fischer et al., 2013; Lawrence et al., 2020; Ignjacevic et al., 2021).

Notably, we also reveal that the ANN can distinguish the year of temperature maps during a climate model's historical simulation of the 20th century. This is surprising given the greater influence of internal variability during this period that continues until around 1980 when the forced greenhouse gas signal begins to more clearly emerge. These skillful predictions by the ANN indicate that changing temperature patterns within a climate model may be distinguished from the noise of internal variability well before a statistically significant mean temperature trend is detected. This early 20th century detection skill improves even more when training on maps of higher spatial resolution compared to a lower resolution configuration of the same large ensemble climate model. Using methods from explainable artificial intelligence (XAI), we conclude this study by examining the regional patterns of temperature change linked to this ANN performance and its possible dependence on the size of the training data climate maps.

## 2 Data

### 2.1 Climate Model Large Ensembles

For the main results of this study, we use a collection of 30-member initial condition large ensemble simulations from a fully-coupled global climate model (GCM) called the Seamless System for Prediction and EArth System Research (SPEAR; Delworth et al., 2020). SPEAR is the newest seasonal to multidecadal prediction and projection system from the National Oceanic and Atmospheric Administration (NOAA) Geophysical Fluid Dynamics Laboratory (GFDL). SPEAR uses the same atmospheric model code

(AM4) and land model code (LM4) as in the GFDL CM4 model (Zhao et al., 2018a, 2018b; Held et al., 2019), but employs a different configuration of the MOM6 ocean code (Adcroft et al., 2019) so as to optimize the model for seasonal to decadal predictions and projections. SPEAR incorporates 33 vertical atmospheric levels and can be designed for different atmosphere/land resolution configurations ranging from 0.25° to 1.0°. All model versions include a common ocean grid of approximately 1.0° spacing (though refined to 0.33° around the equator). The SPEAR system has already been successfully used for several studies in evaluating the predictability of temperature variability and heat extremes across North America (e.g., Jia et al., 2022; Yang et al., 2022).

**Table 1.** Summary of the GFDL SPEAR_MED large ensemble simulations used in this study (Delworth et al., 2020). Additional details on SPEAR can be found at `https://www.gfdl.noaa.gov/spear_large_ensembles/`.

| Name | Scenario Forcing | Years | # Members | Horizontal Resolution (Atmosphere / Ocean) |
|---|---|---|---|---|
| SPEAR_MED_SSP119 | Historical to 2014, SSP1-1.9 | 1921-2100 | 30 | nominal 0.5°/nominal 1.0° |
| SPEAR_MED_SSP245 | Historical to 2014, SSP2-4.5 | — | — | — |
| SPEAR_MED_SSP585 | Historical to 2014, SSP5-8.5 | — | — | — |
| SPEAR_MED_NATURAL | Natural Forcing Only | — | — | — |

Due to the availability of more individual ensemble members to train, validate, and test our neural network than in the configuration with the highest atmospheric resolution, we focus on simulations from only the SPEAR_MED (atmosphere/land of 0.5°) and SPEAR_LO (atmosphere/land of 1.0°) configurations. Ensemble members in each are initialized from conditions in an 1850 control simulation that are branched 20 years apart. Both SPEAR_MED and SPEAR_LO are forced with CMIP6 historical forcing through 2014 (Eyring et al., 2016). From 2015 to 2100, they are then forced with future projections from the Shared Socioeconomic Pathway 5-8.5 scenario (SSP5-8.5; Kriegler et al., 2017; Riahi et al., 2017). Recent work has shown that SSP5-8.5 is likely an unrealistical extreme future emission scenario (e.g., Burgess et al., 2020; Hausfather & Peters, 2020; Peters & Hausfather, 2020). Although our study here is mostly focused on the ToE in the recent and historical past, we also compare the sensitivity of our machine learning results to training and testing on SPEAR_MED simulations conducted with more prob-

able SSP scenarios (Pielke et al., 2022), including the low-end SSP1-1.9 and more moderate SSP2-4.5 (O'Neill et al., 2016).

Finally, we compare our SPEAR climate change simulations with a large ensemble experiment starting from the same initial conditions in 1921 as SPEAR_MED, but holding all anthropogenic forcings (i.e., greenhouse gases, anthropogenic aerosols, land use/land change) fixed at 1921 levels. This experiment, referred to as SPEAR_MED_NATURAL is instead prescribed with only natural radiative forcings, such as those due to solar irradiance and volcanoes (historical to 2014, hypothetical thereafter; see Delworth et al. (2022)). By comparing SPEAR_MED and SPEAR_MED_NATURAL, we can extract the role of anthropogenic forcing on changes in summertime temperatures in the climate model and gain insights for understanding how well the neural network performs by training on data without a long-term anthropogenic signal.

To summarize, we consider collections of 30-member large ensembles from either SPEAR_MED or SPEAR_LO for designing our machine learning architecture. These simulations are conducted on two different horizontal atmosphere/land grids, which we will now refer to as either MED (nominal 0.5° grid) or LOW (nominal 1.0° grid) throughout the text. We point this out since a key focus of this work is on comparing the effect of higher spatial resolution on the performance of the neural network. Given the limited availability of other fully-coupled large ensembles with high-resolution atmospheric models, we can only assess the MED grid using simulations conducted by SPEAR_MED or its previous generation version called the Forecast-Oriented Low Ocean Resolution (FLOR) system (Vecchi et al., 2014). FLOR is a fully-coupled global climate model based upon GFDL's CM2.5 (Delworth et al., 2012); its large ensemble includes 30 members with CMIP5 historical forcing from 1921 to 2004 and Representative Concentration Pathway 8.5 (RCP8.5; Riahi et al., 2011; Vuuren et al., 2011) thereafter from 2005 to 2100. FLOR has a land-atmosphere resolution of 0.5° using the AM2.5 and LM3 model components (Milly et al., 2014), but includes a coarser ocean from OM2.1 (Gnanadesikan et al., 2006) at a nominal resolution of 1.0°. Since FLOR does not offer a corresponding LOW version of the large ensemble like SPEAR, we simply bilinearly interpolate its corresponding temperature maps to the LOW grid (denoted as FLOR (LO)) for again attempting to compare the advantange of inputting more (or less) spatial information into our neural network framework.

We also briefly make use of three more GCM large ensembles, which offer a similar horizontal resolution and data availability as SPEAR_LO (i.e., data from at least 1921 to 2100 and a grid size of approximately $1.0° \times 1.0°$). These include the 40-member Community Earth System Model Large Ensemble Project Version 1 (CESM1-LE; Hurrell et al., 2013; Kay et al., 2015) (CMIP5 class; RCP8.5), 100-member CESM2-LE (Danabasoglu et al., 2020; Rodgers et al., 2021) (CMIP6 class; SSP3-7.0), and the 50-member large ensemble using the sixth version of the Model for Interdisciplinary Research on Climate (MIROC6-LE; Tatebe et al., 2019; Shiogama et al., 2023) (CMIP6 class; SSP5-8.5). While other climate model large ensembles are available from the multi-model large ensemble archive that contain at least 30 members (NCAR, 2020; Deser et al., 2020), their horizontal resolution is generally too coarse for our regional deep learning approach, particularly when considering the three smaller geographic areas of the United States. In other words, an insufficient number of grid points nearly reduces the problem to a change-point time series task. This then limits the real utility of the neural network methodology, which here is to exploit any possible (non)linear temperature patterns across the maps in order to identify the emergence of forced climate signals.

For the climate model data, we leverage monthly mean near-surface daily maximum, minimum, and average temperature data (i.e., TMAX, TMIN, TAVG) and then calculate the seasonal mean over June to August (JJA) using only grid points across the conterminous United States. A summary of the large ensemble data can be found in Tables 1 and S1, which include the final horizontal resolution elected to be used as input to the neural network.

### 2.2 Observations

To evaluate the ToE of summertime surface temperatures in the United States, we primarily use the NOAA Monthly U.S. Climate Gridded Dataset (NClimGrid; Vose et al., 2014), which is a station-based gridded product of temperature (and precipitation) across land areas of the conterminous United States since 1895. NClimGrid is based on the interpolation of quality-controlled station data onto 5 km latitude/longitude grids using records from the Global Historical Climatology Network (GHCN; Durre et al., 2010; Menne et al., 2012). The homogenized dataset from NClimGrid also accounts for bias correction of artificial station breaks, such as for changes in weather station locations, instruments, and other temporal inconsistencies (Menne & Williams, 2009). Area-average

NClimGrid temperature errors are larger over CONUS prior to 1990, but generally still within 1°C for both TMAX and TMIN (Vose et al., 2014). Here we focus on the period of 1921 to 2022, which overlaps with the output from the SPEAR climate model simulations.

**Table 2.** Summary of the station-based and atmospheric reanalysis datasets (i.e., observations) used in this study. The observational maps are then regridded to make inferences using the neural networks onto either the same LOW (1.0°; ERA5/NClimGrid/20CRv3) or MED (0.5°; ERA5/NClimGrid) atmospheric resolution as from GFDL SPEAR. See Section 2.1.

| Name | Data | Years | Horizontal Resolution | Reference |
|---|---|---|---|---|
| ERA5 | ECMWF Reanalysis v5 | 1940-2022 | $\sim 0.25°$ | Hersbach et al. (2020) |
| NClimGrid | NOAA Monthly U.S. Climate Gridded Dataset | 1895-2022 | $\sim 1/24°$ | Vose et al. (2014) |
| 20CRv3 | NOAA-CIRES-DOE Twentieth Century Reanalysis v3 | 1836-2015 | $\sim 1.0°$ | Slivinski et al. (2019, 2021) |

For comparing the sensitivity of our observational neural network predictions to the use of NClimGrid, we also briefly evaluate our results with two atmospheric reanalysis datasets: the European Center for Medium-Range Weather Forecasts (ECMWF) fifth generation of atmospheric reanalysis (ERA5) available from 1940 to 2022 and the NOAA/Cooperative Institute for Research in Environmental Sciences/Department of Energy Twentieth Century Reanalysis (20CR) version 3 (20CRv3) available from 1836 to 2015. Using ECMWF's Integrated Forecast System release 41r2 and four-dimensional variational analysis as a data assimilation scheme, ERA5 provides global data at a horizontal resolution of 31 km in near-real time (Hersbach et al., 2020). It is constrained by numerous observations, like land-based weather stations, satellites, radiosondes, and other aircraft records. We focus on near-surface temperature (2-m height) from ERA5 through its entire available temporal period (1940 to 2022). To compare with a longer reanalysis record, we use near-surface temperature (2-m height) from 20CRv3 (Slivinski et al., 2019) during the overlapping period from 1921 to 2015. Unlike ERA5, this 20CR product only assimilates surface pressure observations (Compo et al., 2011), which is completed through four-dimensional incremental analysis updates and an 80-member ensemble Kalman filter approach (Lei & Whitaker, 2016; Slivinski et al., 2019). 20CRv3 uses the coupled atmosphere-land National Centers for Environmental Prediction (NCEP) Global Forecasting System (GFS) version 14.0.1 with boundary conditions from prescribed sea surface temperatures and sea-ice concentration. Overall, 20CRv3 is an improvement

over its predecessor (20CRv2c) for simulating synoptic dynamics and other long-term surface climate fields (Slivinski et al., 2021), though work continues to improve temperature data earlier in the twentieth century due to greater uncertainties (Gillespie et al., 2023).

We use monthly mean temperature output from all observational and reanalysis datasets to calculate the JJA seasonal mean. The neural network used here requires the input data to have the same latitude and longitude dimensions. Therefore, we bilinearly regrid NClimGrid and ERA5 onto the MED (0.5°) and LOW (1.0°) spatial maps which are also used by the climate model large ensemble data. The coarser 20CRv3 dataset is instead only interpolated onto the LOW resolution grid. Importantly, these three observationally based products encompass a wide range of different structural methodologies and uncertainties, which therefore provide ample opportunity to test the robustness of the neural network results on out-of-sample data (Table 2). A comparison of average JJA CONUS temperature anomalies is also presented in Figure S1 for NClimGrid, ERA5, and 20CRv3.

## 3 Methods

### 3.1 Neural Network Framework

We adopt a machine learning ToE method first proposed by Barnes et al. (2019), which uses a neural network to input geographic maps of climate variables and then to output the year associated with each map. While this is quite a simple prediction problem, it has been shown that the neural network must learn to leverage time-evolving patterns of forced climate signals in order to correctly identify the year with a single map (Barnes et al., 2020). This attribution method has since been used in a wide range of climate applications (e.g., Anderson & Stock, 2022), such as for disentangling the role of aerosols and greenhouse gases in single-forcing large ensembles (Labe & Barnes, 2021), quantifying anthropogenic signals in extreme precipitation (Madakumbura et al., 2021), and identifying the ToE of combined variables like precipitation and temperature (Rader et al., 2022). For this work, we take a similar approach, but build upon these previous efforts by focusing on a narrower application. Here we train on high-resolution climate model data and evaluate the ToE on a smaller spatial region during a period with an observed absence of daytime warming (e.g., boreal summer in the Central United States)

(Partridge et al., 2018). While previous machine learning efforts in climate science have usually interpolated data to coarser grids for reasons such as computational limitations, we are particularly interested in considering whether the neural network skill (or the actual ToE) changes by training and evaluating on higher resolution data.

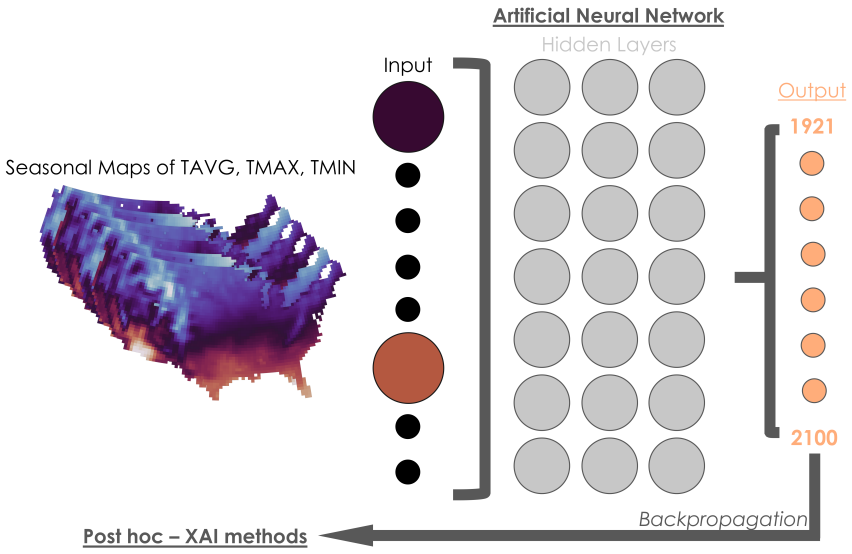


**Figure 1.** Schematic of the artificial neural network (ANN) used to take an input map of average June to August (JJA) temperatures over the contiguous United States (CONUS) and then output the likelihood that the map is from a particular decade. Fuzzy classification (Zadeh, 1965) is used to decode this decadal likelihood to associate each map with a single year (i.e., our final predicted output). The ANN consists of different combinations of hidden layers and nodes depending on the horizontal resolution of the training data. (see Section 3 for the architecture and hyperparameter choices). Explainable artificial intelligence (XAI) attribution methods are then used to reveal the regions that acted to increase or decrease the likelihood of the ANN's predicted year.

For this study, we use an ANN, which is a statistical algorithm that can learn to approximate nonlinear functions from large quantities of data. They have become increasingly popular tools for Earth science prediction problems and in numerical modeling (Boukabara et al., 2021; Chantry et al., 2021; Irrgang et al., 2021). ANNs are fully-connected networks, which in their simplest form are comprised of an input layer, a set number of hidden layers and nodes, and an output layer (i.e., the final prediction). Ev-

ery node in this feed-forward architecture receives information from the previous layer and can be individually computed by weighting the sum of the inputs and an added bias term. The weights and biases are updated iteratively until the training process is finished, such as when the loss function is minimized (i.e., a measure of machine learning model error). Given enough available data and limited overfitting, the ANN can then be used to make skillful predictions on data it did not see during training (i.e., testing data). More thorough introductions to neural networks can be found in e.g., Lecun et al. (2015); Goodfellow et al. (2016); Neapolitan and Jiang (2018). Domain-specific tutorials for machine learning applications, such as in atmospheric science, are also provided by Chase, Harrison, Burke, et al. (2022) and Chase, Harrison, Lackmann, and McGovern (2022).

Figure 1 shows the ANN architecture used for this study. The ANN receives vectorized latitude by longitude maps of JJA temperatures (TMAX, TMIN, or TAVG) with either 10,080 input values (70×144) per sample for the MED resolution maps or 2,520 input values (35×72) for the LOW resolution maps. In addition, we also evaluate ANNs using regional map inputs for the Western, Central, and Eastern CONUS as depicted in Figure 2. We focus on only land areas and therefore mask all other areas by assigning values of zero, which the ANN then learns to ignore. This vector is fed into the ANN hidden layers, and the output is the probability of a particular decade midpoint (a classification problem). These output values are then translated to a particular year (a regression problem) using fuzzy classification (Zadeh, 1965), which is described in detail in Barnes et al. (2020). Briefly, by denoting the central year of a particular decade (e.g., 1995 for 1990-1999), a particular output (e.g., 1994) can be mapped to more than one decade class. Using triangular membership functions (Zadeh, 1965) with a width of one decade, the weighted sum of the decadal class probabilities can finally be mapped to a specific year. For example, the year 1994 has a probability of 0.9 for falling within the decade class midpoint of 1995 and a probability of 0.1 for the decade class midpoint of 1985. We refer to the regression problem of the predicted year throughout the rest of the study for evaluating the ANN skill and ToE calculations.

We select our ANN architecture for different input maps by considering the effect of spatial region and grid resolution. The final ANNs are selected by identifying the lowest median Mean Absolute Error (MAE) on validation data after considering 20 ANNs (randomized combinations of training, testing, and validation data and initialization seeds)

over architectures that range in complexity by the number of hidden layers and nodes (see Figures S6 and S10). In other words, this is related to the number of parameters that the model can use to learn the relevant climate patterns to more accurately predict the year of a map (i.e., variations of deeper or shallower neural networks). For CONUS inputs on the MED grid, we use an ANN with 3 hidden layers of 10 nodes each. For CONUS inputs on the LOW grid, we use an ANN with 3 hidden layers of 20 nodes each. Finally, for the regional CONUS maps on both the MED and LOW grids, we use an ANN with 2 hidden layers of 100 nodes each. Despite selecting different ANN architectures, we find that our results are generally robust across minor changes in hyperparameter options. The rectified linear unit (ReLu; Agarap, 2018) activation function is used for the non-linear transformation in the hidden layers, and a softmax operator is included in the output layer to ensure that the decadal class probabilities of the output vector sum up to one. All ANNs here use the binary cross-entropy loss function, stochastic gradient descent optimizer (Ruder, 2016) with Nesterov momentum set to 0.9 (Nesterov, 1983), a learning rate of 0.01, and a batch size set to 32.

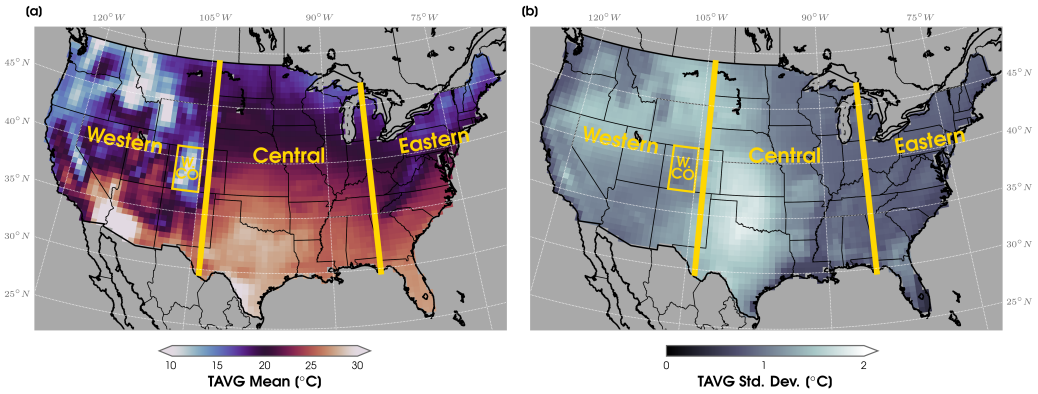


**Figure 2.** (a) Composite of JJA average near-surface temperature (TAVG) from an example of the training data mean used to standardize the input maps, which is calculated here using 24 ensemble members from SPEAR_MED over the period of 1981-2010. (b) As in (a), but for the training data standard deviation. See Section 3.1 for more details. The vertical yellow longitude lines are displayed at 104°W and 85°W to differentiate the three regions of the CONUS considered for this work (i.e., the Western USA, Central USA, and Eastern USA). The thin yellow box outlines the western Colorado (W. CO) region of interest used for this study (37°N-41°N and 108°W-105°W).

Unless otherwise stated, we train on 24 ensemble members, validate on 4 ensemble members, and test on 2 ensemble members. To limit overfitting on the training data, we apply a few different methods. First, we apply early stopping, which ends the ANN training process if there is no improvement in the validation loss for 25 consecutive epochs. The epoch with the best performance is ultimately selected. Next, we apply ridge regularization (set to 0.001) to the weights of the first hidden layer ($L_2$; Friedman, 2012), which acts to limit how sensitive the weights are to outliers in the input data. This also helps to smooth out any spatial autocorrelation that exists in the temperature maps and improve overall interpretability (Sippel et al., 2019, 2020; Barnes et al., 2020). The sensitivity of our ANN results to the choice of the $L_2$ parameter are shown in Figures S6 and S10.

Before inputting the data into the ANN, we standardize all climate model temperature maps by subtracting the training data mean and dividing by the training data standard deviation over the 1981 to 2010 climatological baseline. This is computed separately at every grid point. Note that similar skill is found for training and testing data using other reference periods, such as 1951-1980. An example of the training mean and standard deviation for 24 ensemble members in SPEAR_MED is shown in Figure 2. Due to mean state biases that may exist between the climate model large ensembles and observations, we separately standardize the observations by their own mean and standard deviation over 1981 to 2010 before making ANN inferences. Though, as we discuss later, it is still possible that differences in the amount of mean warming between the climate model simulations and observations could influence the machine learning skill and related ToE results.

In addition to evaluating our ANN and ToE predictions, we consider several ad hoc attribution methods of XAI. Explainability methods have increasingly been shown to aid in building trust and understanding for the decision-making process of neural networks, including for climate science applications (e.g., Sonnewald & Lguensat, 2021; Labe & Barnes, 2022; Shin et al., 2022; Rampal et al., 2022; Diffenbaugh & Barnes, 2023; Mamalakis et al., 2023). Output from XAI attribution methods describe the contribution of every input sample's latitude and longitude grid point (described here as "relevance") to the overall prediction of the ANN. In other words, the XAI algorithms return a relevance heatmap (unitless) for every input year. To evaluate the sensitivity our explainability results across different methods (Mamalakis et al., 2022; Bommer et al., 2023), we consider three dif-

ferent attribution techniques: the layerwise relevance propagation z-rule ($LRP_z$; Bach et al., 2015), LRP epsilon-rule ($LRP_\epsilon$; Bach et al., 2015) and Integrated Gradients (Sundararajan et al., 2017). Though given the similarity of the results across the XAI methods, we only show relevance figures from $LRP_z$ and Integrated Gradients for brevity in the main results. A more detailed overview on an application of LRP to a geoscience problem can be found in Toms et al. (2020). Lastly, we caution that although these XAI techniques are very useful for outlining the important climate patterns learned by the ANN, they do not imply causation, such as for the specific physical drivers.

Our XAI heatmaps are based on composites of the testing ensemble members (or observations), where positive areas of relevance can be interpreted as regions that pushed the ANN toward its predicted year. Negative areas of relevance are subsequently interpreted as vice versa, i.e., locations that tried to push the ANN away from making its yearly prediction. We only consider relevance maps for testing data predictions that are accurate to within 5 years of the actual year.

### 3.2 Calculated Timing of Emergence for Observations

An annotated graphic of our ToE approach is shown in Figure 3. The ANN in this example is trained and tested on SPEAR_MED for maps of mean JJA TAVG. The testing ensemble predictions are shown using green scatter points with the actual year of a TAVG map on the x-axis, and the ANN predicted year is on the y-axis. Predictions for maps of mean JJA TAVG from NClimGrid data are then shown with red markers. The fundamentals of this ToE estimate are based on Mora et al. (2013), which is calculated as the year that a map of temperature first departs the bounds of historical climate variability and continues to do so for all future years (e.g., the red shading in Figure 3). This general approach was also demonstrated using ANNs in Barnes et al. (2019) and Rader et al. (2022). Here we define the ToE as the first predicted year that is greater than the maximum prediction during our historical baseline of 1921-1950 (gray shading in Figure 3). In addition to 1921-1950 being the earliest 30-year period available from SPEAR_MED, it more importantly overlaps with the observed anomalous warmth of the 1930s in the United States (Program, 2018) (Figure 4). Thus, we can directly compare whether an observed JJA temperature forced signal has emerged outside of this historical record that includes the extreme heat of the Dust Bowl era.

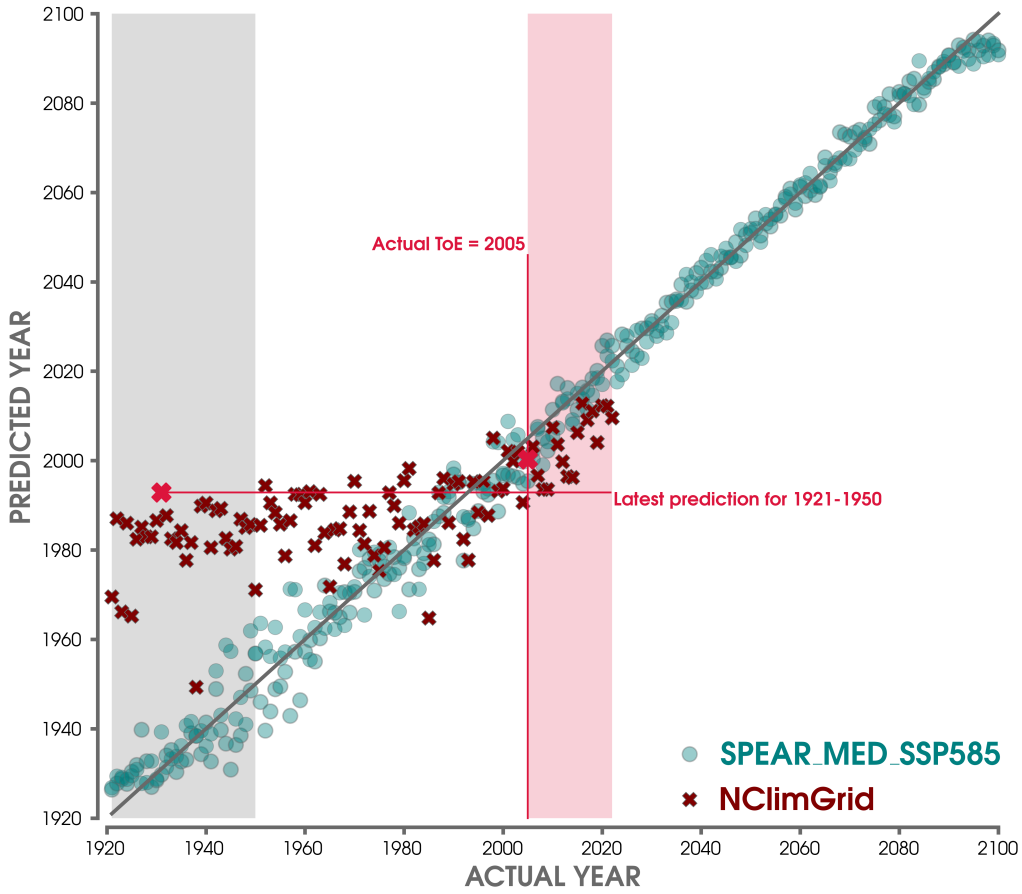


**Figure 3.** Schematic of the output provided by the ANN and the subsequent calculation of the timing of emergence (ToE) based on CONUS maps from observations. Blue scatter points denote ANN predictions based on SPEAR_MED (historical + SSP5-8.5 forcing) testing ensemble members for the inputs of TAVG averaged over JJA. The actual year is shown on the x-axis, and the predicted year on shown on the y-axis. Red markers are used for ANN predictions after inputing maps from NClimGrid. A perfect prediction (1:1 slope) is annotated behind all ANN predictions with a solid gray line. To first calculate the ToE of the NClimGrid maps, the latest predicted year in the 1921 to 1950 climatological period (vertical gray shading) is identified (left, bright red marker). The actual ToE (right, bright red marker) is then the first year when all proceeding predictions (vertical red shading) exceed the year of the 1921-1950 maximum. The observed ToE from this ANN is 2005.

We also compare our results with a more traditional baseline ToE estimate using SPEAR_MED data. For a given ensemble member, the ToE is the first year that the 10-year running JJA mean consistently stays above the 1921 to 1950 climatological period by greater than two standard deviations (Lehner et al., 2017). This variability is again based on the 10-year running mean temperature in 1921-1950. The actual ToE composites from this method are then calculated from the mean across all ensemble members and at every grid point.

# 4 Results

## 4.1 Changes in United States Summertime Temperatures

Before estimating the ToE by the neural network framework, we start by assessing changes in temperature in observations and as simulated by SPEAR_MED. Figure 4 shows the time-mean JJA temperature anomalies averaged over the CONUS for TMAX, TMIN, and TAVG from 1921 to 2022. We find that observations from NClimGrid lie outside the ensemble spread of SPEAR_MED in all three temperatures metrics during the Dust Bowl of the mid-1930s. This is especially prominent for JJA TMAX, which reaches values of more than 1°C greater than the warmest ensemble member from SPEAR_MED. The peak TAVG during this early 20th century period was reached in 1936 (Cowan et al., 2017, 2020), but it is now statistically tied with 2021 as the hottest summer on record (within 0.01°C) over the CONUS (Thompson et al., 2022) (Figure 4c). The large climate response following the eruption of Mount Pinatubo (Parker et al., 1996), however, is well captured by SPEAR_MED. 1992 subsequently remains the coldest mean summer TMAX in the NClimGrid observational record (since at least 1921) (Figure 4a). In more recent years, temperatures from NClimGrid have remained consistently below the ensemble mean and therefore exhibit less net warming than the forced response in SPEAR_MED, especially for TMAX. We also compare NClimGrid with ERA5 and 20CRv3 reanalysis products for TAVG anomalies in Figure S1. While NClimGrid and ERA5 agree well in capturing the interannual variability and long-term trends, we find a larger discrepancy in TAVG prior to 1975 where 20CRv3 shows larger warm anomalies. These discontinuities have been pointed out in previous studies using older model generations of 20CR (Ferguson & Villarini, 2014), which were found to be largest in the mid-20th century for the Central United States (Ferguson & Villarini, 2012).

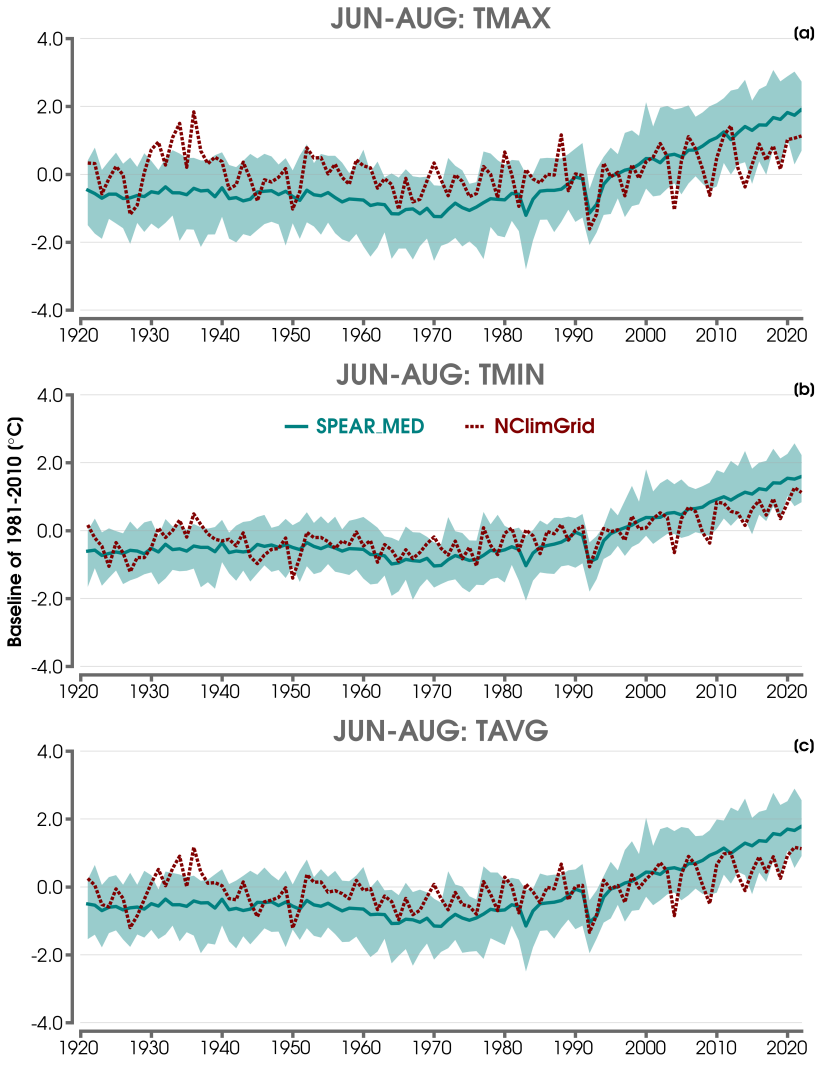


**Figure 4.** (a) Time series of mean JJA maximum temperature (TMAX) anomalies over the CONUS from 1921 to 2022 for the ensemble mean of SPEAR_MED (dark green line) compared to observations from NClimGrid (dashed red line). The spread across SPEAR_MED ensemble members is shown with the light green shading. Anomalies are computed for each dataset with respect to their own 1981-2010 climatological mean. (b) As in (a), but for the mean JJA minimum temperature (TMIN). (c) As in (a), but for the mean JJA average temperature (TAVG)

Figure S2 shows the time series of TAVG anomalies broken down by the three regions of interest, including the Western, Central, and Eastern United States. This more clearly distinguishes the anomalous heatwaves of the mid-1930s across the Central United States, which again fall outside the ensemble spread of SPEAR_MED. All three regional domains experience substantial interannual temperature variability in the NClimGrid record and reveal less long-term warming compared to the SPEAR_MED ensemble mean over the 1990 to 2022 period.

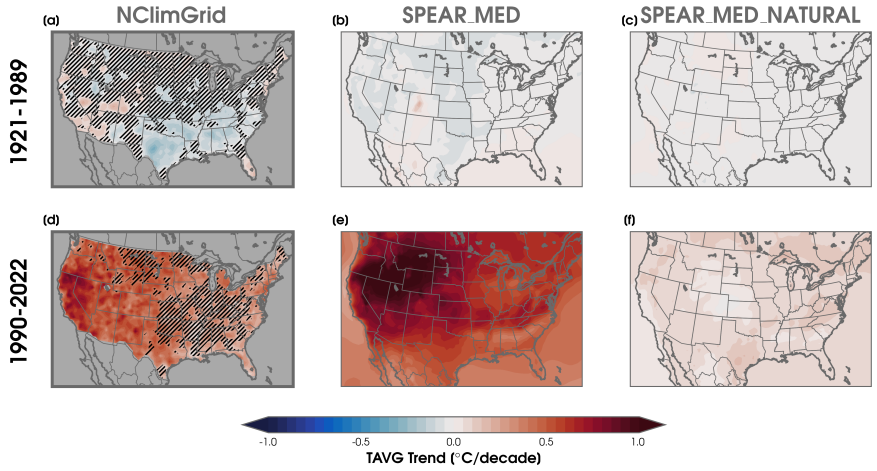


**Figure 5.** Linear least squares trends of average JJA TAVG from 1921 to 1989 (a, b, c) and 1990 to 2022 (d, e, f) for NClimGrid (a, d), the ensemble mean from SPEAR_MED (b, e), and the ensemble mean from SPEAR_MED_NATURAL (c, f). For maps of NClimGrid, black hatch marks indicate TAVG trends that are not statistically significant following a Mann-Kendall test (Bevan & Kendall, 1971; Mann, 1945) for the 95% confidence level.

Figure 5 provides spatial maps of the TAVG trends for 1921 to 1989 and 1990 to 2022. Statistically significant cooling trends in NClimGrid are found over the southern United States (Figure 5a), which are close to the warming hole region for this period (Mascioli et al., 2017). The observational trends are also compared to SPEAR_MED and its parallel natural forcing-only simulation (SPEAR_MED_NATURAL) in Figure 5b-c, which does not simulate any long-term TAVG trends in their ensemble means. However, greater warming is found in observations for the more recent past (1990-2022; Figure 5d), which is largest over the Southwestern United States. The warming hole spatial pattern is again found over the Central United States (Figure 5d). In comparison to observations, SPEAR_MED

simulates greater ensemble mean warming over the entire CONUS and shows the largest TAVG trends over the northern Rocky Mountains (Figure 5e). Without anthropogenic forcings, such as greenhouse gases or aerosols, Figure 5f reveals little to no warming simulated over CONUS as found in the ensemble mean of SPEAR_MED_NATURAL. Lastly, comparing over the entire 1921 to 2022 historical record, we find that the ensemble mean of SPEAR_MED simulates slightly greater warming trends over CONUS (Figure S3b) and does not show evidence of the warming hole as in Figure S3a. Though it remains unclear whether this is simply due to internal variability in the observational record, which would not be captured in a composite of the ensemble mean trends of a large ensemble (Eischeid et al., submitted).

Focusing on JJA mean maximum and minimum temperatures, Figure S4 shows larger recent trends in TMAX than TMIN across the Northern and Western United States in the observed record. Consistent with previous findings, the warming hole is also more prominent in TMAX. This is reflected by an area of insignificant cooling across the Southeastern United States (Figure S4a). Although SPEAR_MED again simulates greater positive trends in TMAX (Figure S4b) and TMIN (Figure S4d) for the CONUS (exceeding 1°C per decade at its local maximum), there are similarities in the spatial pattern of mean warming compared to NClimGrid. This includes a relative maximum in warming over the Western United States and relative minimum over the Southeastern United States (Figure S4d,b).

Lastly, we show in Figure S5 the JJA mean CONUS TAVG for NClimGrid compared to a collection of 30-member large ensembles of SPEAR_MED, but using different radiative forcing scenarios from 2015 to 2100. This includes future projections from SSP1-1.9, SSP2-4.5, and SSP5-8.5, which are compared to the natural-only forcing experiment of SPEAR_MED_NATURAL from 1921 to 2100. The forced response in the SPEAR_MED historical simulations only begins to clearly rise outside the variability in the natural forcing simulation between 1990 and 2000. This occurs a decade later when comparing NClimGrid to SPEAR_MED_NATURAL. While there is a large range in projected ensemble mean JJA TAVG change across the climate change scenarios to 2100, the uncertainty due to internal variability alone is almost 2°C across the ensemble spreads. Notably, we also find that the ensemble mean TAVG begins to cool by 2040 for the aggressive climate mitigation scenario of SSP1-1.9. This continues through the end of the 21st century for SSP1-1.9. We also point out that there are only negligible differences

across future forcing scenarios in the ensemble means of JJA TAVG until between 2030 and 2040, but by after 2080, there is very little overlap in their ensemble spreads due to the greater effects of projection scenario uncertainty.

### 4.2 Predictions by Neural Networks

We now turn to the machine learning results in Figure 6, which shows the skill of the ANN for predicting the year of CONUS maps of TAVG, TMAX, or TMIN. Note that Figure 6a is from the same ANN as the one displayed in the annotated schematic in Figure 3. Again, we focus our results on only the testing ensemble members from SPEAR_MED, which are data that the ANN has never seen before. The testing data predictions (blue scatter points) closely follow the 1:1 line (or perfect prediction) in all three ANNs, which suggests that the ANN is able to distinguish individual JJA temperature maps despite the background noise of internal climate variability. The robustness of these results to different ANN architectures and ridge regularization parameters are shown in Figure S6 for TAVG, which are each assessed for 20 ANN iterations that used different combinations of training, validation, and testing ensemble members and random initialization seeds. The median MAE score from this distribution of ANNs is displayed in Figure S6h for the architecture used to produce Figure 6 (see Section 3.1). This helps to ensure that our high skill is not simply due to the chance that our ANN performed well on only one subset of testing data or overfit on the training ensemble members.

The results for the observational predictions are also shown in Figure 6. To restate from earlier, these predictions are obtained by inputting maps of JJA temperature into the ANN after it has already been trained and tested on the climate model large ensemble data. However, unlike the predictions for SPEAR_MED, we do not find that the ANN can correctly predict the year during most of the 20th century for TAVG, TMAX, or TMIN. As described in Rader et al. (2022), since the ANN is not confident in predicting the year of a given temperature map, it tends to predict around the same year in the middle of the entire time series (i.e., to reduce its potential error penalty in the loss function). However, especially for TAVG and TMIN, we find that the ANN observational predictions begin to lie on the 1:1 line after around 1995. One measure that can be used to reveal whether the ANN has identified patterns of forced change is by evaluating the order of the predicted years (Labe & Barnes, 2021). Thus, this suggests that the ANN is beginning to identify common patterns of forced climate change in NClimGrid that were learned

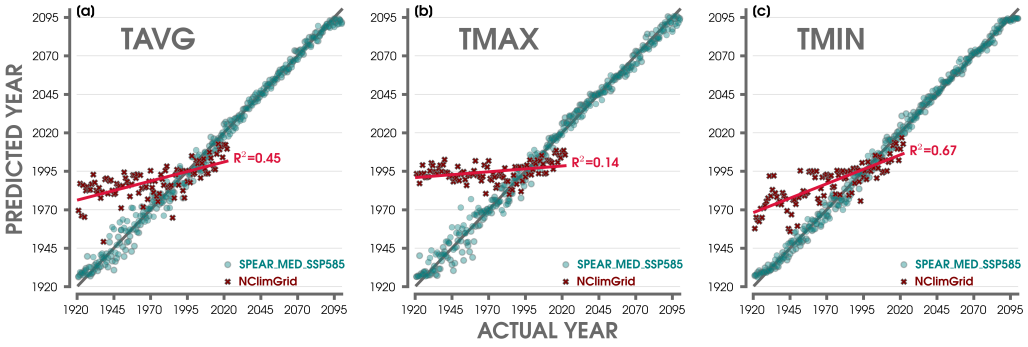


**Figure 6.** (a) Predictions of SPEAR_MED testing ensemble members by the ANN for input maps of TAVG averaged over JJA. The actual year is denoted on the x-axis and the predicted year on the y-axis. Red markers are shown for ANN predictions after inputing maps from NClimGrid. A red line is displayed showing the linear least squares regression through the NClimGrid predictions along with its corresponding $R^2$ value. A perfect prediction (1:1 slope) is annotated behind all ANN predictions with a solid gray line. (b) As in (a), but for an ANN trained and tested on maps of TMAX. (c) As in (a), but for an ANN trained and tested on maps of TMIN.

from the SPEAR_MED temperature maps in more recent years. As discussed later, we relate this point to the ToE of observed temperature change.

Since the SSP5-8.5 radiative forcing may be an unrealistic future climate scenario (Peters & Hausfather, 2020), we examine our results using ANNs trained on TAVG maps from the same historical forcing in SPEAR_MED, but then following either the SSP1-1.9 pathway (Figure S7a) or SSP2-4.5 pathway (Figure S7b). Overall, we find very similar skill for the testing ensemble member predictions across the SSP scenarios relative to SSP5-8.5 (Figure S7c), which is used throughout the remainder of the study. The predictions for inputs of TAVG from NClimGrid are also strikingly similar. We do point out that there is some higher testing data error toward the end of the 21st century, especially for SSP1-1.9 (Figure S7a), which suggests that the forced patterns of change may become less prominent after climate mitigation efforts (Figure S5). This implies more evidence for highlighting that the ANNs are learning to extract time-evolving climate signals, including from within a single ensemble member's realization of internal climate variability. Both of these detection outcomes are not as easily addressed by traditional signal-to-noise time-mean statistics.

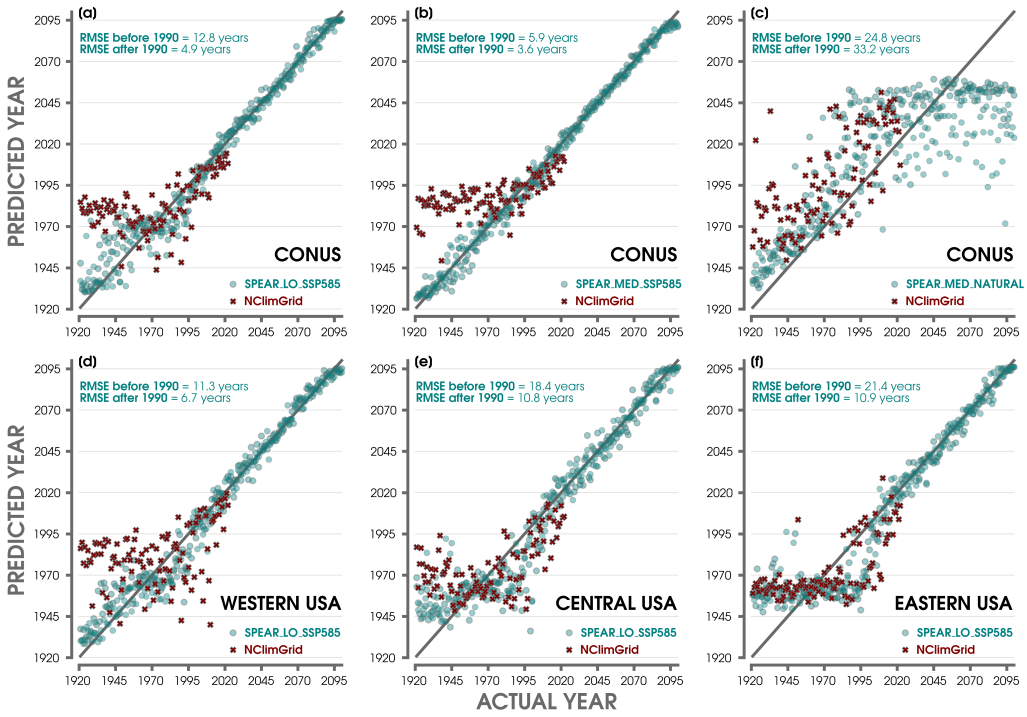


**Figure 7.** (a) Predictions of SPEAR_LO testing ensemble members by the ANN for input CONUS maps of TAVG averaged over JJA. The actual year is denoted on the x-axis and the predicted year on the y-axis. Red markers are shown for ANN predictions after inputing maps from NClimGrid. A perfect prediction (1:1 slope) is annotated behind all ANN predictions with a solid gray line. The Root Mean Squared Error (RMSE) for the SPEAR_LO testing ensemble members is included for predictions over the actual years of before and after the year 1990. (b) As in (a), but for SPEAR_MED. (c) As in (a), but for SPEAR_MED_NATURAL. (d) As in (a) but for SPEAR_LO predictions based on input maps of only the Western USA. (e) As in (a), but for SPEAR_LO predictions based on input maps of only the Central USA. (f) As in (a) but for SPEAR_LO predictions based on input maps of only the Eastern USA.

Another possibility is that our ANN inferences made on the observational maps are sensitive to the choice of data product. We evaluate this prospect in Figure S8 by inputting maps from either ERA5 reanalysis (Figure S8a) or 20CRv3 reanalysis (Figure S8b). The results from ERA5 are very similar to NClimGrid and once again indicate that only by about 2000 is the ANN able to identify the order of the years of the CONUS maps with a high degree of accuracy. On the other hand, we find greater divergence in predicting the year for maps of 20CRv3. This could be due to the lower resolution of the training and testing data (i.e., LOW grid; Section 2) and/or a result of the discontinuity in the TAVG anomalies prior to 1980 (Figure S1). As a result of the greater uncertainties in the early 20CR data and lack of available data for the Dust Bowl era in ERA5, we focus on NClimGrid for the remainder of this study.

To identify the spatial regions that are important for the ANN testing skill for SPEAR_MED, we evaluate composites of the XAI relevance maps in Figure S9 for the $\text{LRP}_z$, $\text{LRP}_\epsilon$, and Integrated Gradients methods. These composites are assessed over the entire time series from 1921 to 2100. Again, positive areas can be interpreted as regions that were more relevant for the overall ANN yearly predictions. Although there are some small differences outlined between the LRP methods and Integrated Gradients, all three composites reveal that the northern Rocky Mountains, such as in western Montana, are an important indicator region. Other relevant temperature regions include areas in the Eastern United States, especially in southern Florida and on the leeward side of the Appalachian Mountains. There is also a notable gradient along this topographic boundary, with negative areas of relevance (i.e., locations that pushed the ANN to predict another year/decade) stretching from the Ohio Valley to western New York State.

Although the ANNs are clearly able to learn a climate signal that distinguishes one climate model temperature map year from another, this does not necessarily imply it is related to anthropogenic forcing. We therefore explore this possibility in Figure 7b, which shows the predictions for an ANN trained on SPEAR_MED compared to the simulation with only natural forcing (Figure 7c). This reveals that the ANN is no longer able to make an accurate prediction of the year when trained on maps from SPEAR_MED_NATURAL. Similarly, there is a much larger temporal spread in predictions after inputting NClimGrid data into this trained ANN (Figure 7c) compared to the SPEAR_MED network. That is, the ANN is likely using the response to external forcings, such as those prescribed in SPEAR_MED, to more skillfully predict the year of summertime temperature maps even

when temperature trends are weaker prior to 1990 (Figure 2). Having said that, there is slightly smaller spread in the earlier yearly testing predictions of SPEAR_MED_NATURAL, which could be due to the ANN detecting a minor influence of solar or volcanic forcings. We did briefly explore training on an simulation of SPEAR with anthropogenic aerosols held fixed to 1921 levels (not shown), but found similar yearly map predictions as those from using SPEAR_MED, which implies a limited role for anthropogenic aerosols on our ANN ToE results.

As noted by the results when evaluating 20CRv3, a last possibility is that training on the high spatial resolution of SPEAR_MED is having an important role in the skill of the testing ensemble members. Put in another way, the ANN could be more likely to weight spatial information, such as temperatures around topographical gradients, for identifying the relevant climate indicators. We can compare this effect by training on data with the LOW grid from the SPEAR_LO configuration, which is demonstrated in Figure 7a. In addition to higher Root Mean Squared Error (RMSE) for SPEAR_LO predictions before and after 1990, there is also greater spread in prediction years after inputting NClimGrid TAVG maps. Similar to earlier, the robustness of the ANN skill for training and testing on SPEAR_LO across different architectures is shown in Figure S10. Thus, the overall effect of grid size is explored more in Section 4.2.2.

#### *4.2.1 Regional Variations in Timing of Emergence*

So far, we have demonstrated that an ANN can distinguish the year of a given map of summertime temperatures across the contiguous United States after training on a high-resolution climate model large ensemble (SPEAR_MED). Consistent with recent work (e.g., Barnes et al., 2020; Labe & Barnes, 2021; Rader et al., 2022), the ANNs here are learning time-evolving temperature patterns associated with external forcing to differentiate each individual year and in the correct sequential order. Moreover, the ANNs can make skillful predictions on the order of temperature map years from out-of-sample observations, but only in the last decade or two.

To associate the period when the ANN predictions for observations begin to fall along the 1:1 prediction line, we compute the observed ToE following the methods in Section 3.2 and outlined in Figure 3. In short, we find the maximum predicted year during the 1921-1950 reference period and then identify the ToE as the point where the forward-

looking ANN predictions no longer fall below this historical maximum. This is first calculated using NClimGrid maps of CONUS that are seasonally averaged for JJA. To ensure the robustness of our observed ToE estimates, we conduct 100 ANNs that are trained on SPEAR_MED and use the same architecture as previously outlined. The uncertainty spread in these ToE predictions, as displayed in Figure 8a for TAVG, TMAX, and TMIN, can be attributed to differences in ensemble members used for training and validation and through the choice of 100 different random initialization seeds. Figure 8b displays the Spearman's rank correlation calculated between the actual and predict years across the NClimGrid inputs as a measure of skill for the ANN to correctly identify the order of the years (Labe & Barnes, 2021). Albeit, by construction, earlier ToEs will also correspond to higher correlation coefficients.

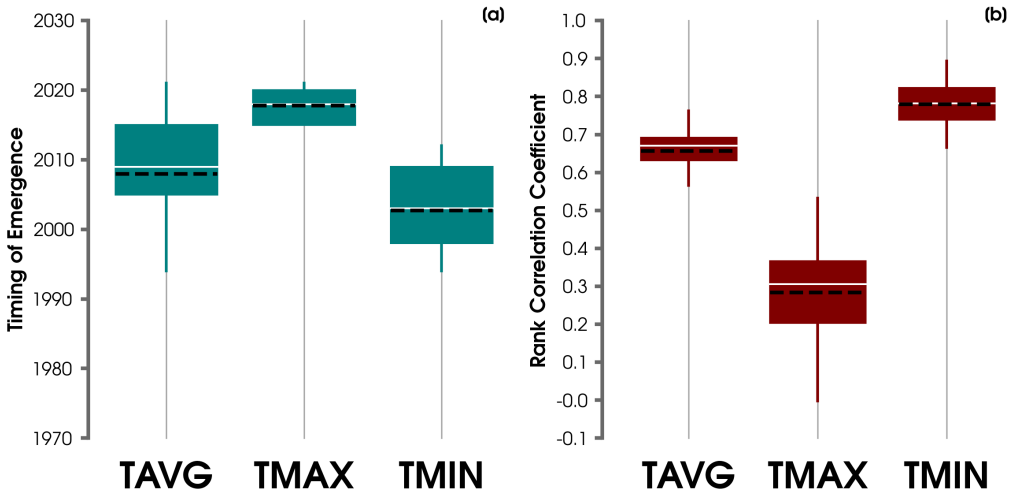


**Figure 8.** (a) Distribution of ToE predictions for inputs of CONUS maps using NClimGrid after training ANNs on SPEAR_MED data of TAVG, TMAX, or TMIN. The median ToE is shown with a thin white horizontal line. The mean ToE is shown with a dashed black line. Each distribution of ToE years is constructed from 100 ANN iterations (use of random initialization seeds and different combinations of training, testing, and validation ensemble members). (b) As in (a), but for the distribution of Spearman's Rank correlation coefficients between all the actual years and predicted ANN years for NClimGrid data.

The earliest median ToE is found for maps of TMIN, which is calculated to be 2003 (R=0.78). On the other hand, the latest median ToE occurs in 2018 for TMAX (R=0.31). An important caveat, however, is that these ToE estimates could be biased early. This

is mainly an issue for late ToE predictions, like for TMAX, where by construction there are few future years to compare with against the historical 1921-1950 maximum. We also cannot rule out temporary reductions in temperature for a single JJA future year as a result of internal climate variability (Maher et al., 2020) or from the influence of an event like an explosive volcanic eruption (Sear et al., 1987).

We now investigate the ToE for the three selected regions across the United States by separately training on mean JJA maps of TAVG, TMAX, and TMIN from SPEAR_MED, but only over each smaller domain (outlined in Figure 2). Figure 9 shows the yearly predictions for these testing ensemble members and for regional inputs from NClimGrid. If a forced signal has emerged in the observational record according to our definition, then the ToE is annotated per each region. Note that if the estimated ToE is within five years of present day (2022), then cautionary asterisks are included next to the ToE year, given the greater uncertainty that future predictions over the next several years will remain above the base period maximum. Across all three regions of CONUS, the earliest ToE occurs for TMIN, especially for the Eastern United States at an estimate of 1998 (R=0.84) (Figure 9f). We also find that the results for TMIN closely follow the 1:1 line, especially after 1950 for the Central and Eastern United States (Figures 8e-f). This suggests that these regional climate signals learned by the ANN after training on SPEAR_MED are generalizable to NClimGrid. Most regions have not observed the emergence of a signal in JJA TMAX (Figure 9a-c), and the earliest possible estimate here is for the Western United States (ToE=2014, R=0.28) (Figure 9a). These results are consistent with recent station-based studies finding greater warming rates in TMIN than TMAX (Meehl et al., 2009; Abatzoglou & Barbero, 2014; Meehl et al., 2016; Program, 2018).

In addition to the observational results, we find that the testing predictions from SPEAR_MED closely follow the 1:1 line, especially after 1990. The lowest testing RMSE is found for the Western United States, and generally the worst ANN ensemble member skill is found for TMAX and TAVG in the Eastern United States (Figures 8c and 8i). At the same time, the overall skill in the early to mid-20th century continues to be surprising, especially given the lack of ensemble mean warming for all three regions (Figure S2). Therefore, in order to provide a baseline with a more traditional linear method of calculating the ToE at each grid point, we compare the ANN predictions with the estimated ToE as shown in Figure S11 following Lehner et al. (2017) (see Section 3.2). For most regions of CONUS, the more conventional ToE in SPEAR_MED for TAVG, TMAX,

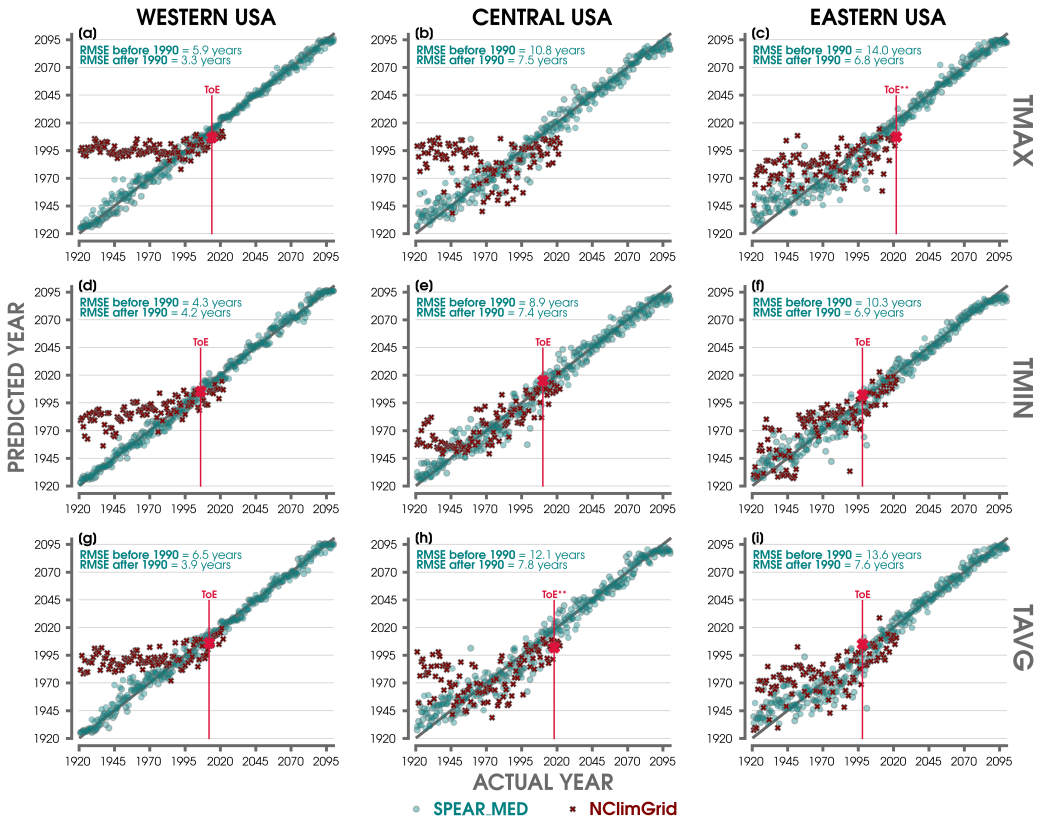


**Figure 9.** ANN predictions of SPEAR_MED testing ensemble members of TMAX maps for the Western USA (a), Central USA (b), and Eastern USA (c). The actual year is denoted on the x-axis and the predicted year on the y-axis. Red markers are shown for ANN predictions after inputing maps from NClimGrid. A perfect prediction (1:1 slope) is annotated behind all ANN predictions with a solid gray line. The RMSE for the SPEAR_MED testing ensemble members is included for predictions over the actual years of before and after the year 1990. If the observed ToE occurs for the NClimGrid predictions, then it is denoted for each region with a bright red marker and vertical line. If five or less years of predictions exist after this calculated ToE, then it is annotated with two added asterisks. (d-f) As in (a-c), but for input maps of TMIN. (g-i) As in (a-c), but for input maps of TAVG.

643 and TMIN occurs in the 1990s or 2000s, though there is some evidence of an earlier ToE
644 across the higher elevations of the Western United States (Figure S11).

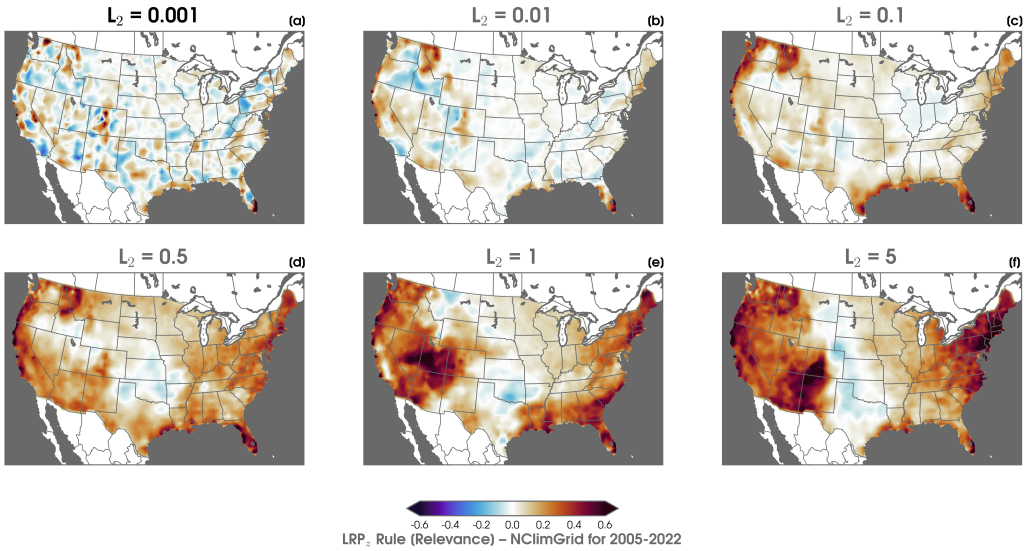


**Figure 10.** (a-f) Composite of $LRP_z$ heatmaps from predictions based on NClimGrid, which are averaged over 2005 to 2022 for ANNs trained on SPEAR_MED data using different $L_2$ regularization values (0.001, 0.01, 0.1, 0.5, 1, 5). The $L_2$ value used for the main results of the paper is labeled in bold font. Positive relevance indicates regions that pushed the ANN to make its predicted year. Negative relevance suggests areas that tried to push the ANN away from its predicted year.

645 Finally, we evaluate the XAI conclusions for NClimGrid (after training on SPEAR_MED)
646 by compositing those heatmaps over 2005 to 2022 in Figure 10a. This temporal range
647 corresponds more closely to the period when a temperature signal has emerged (Figure
648 8). As discussed in Section 3.1, the inclusion of ridge regularization can be a useful pa-
649 rameter to limit the amount of overfitting. Likewise, it is also useful for interpreting the
650 explainability results for how an ANN is making its predictions, as it is analogous to spa-
651 tial smoothing for removing spurious outliers (Sippel et al., 2019; Barnes et al., 2020).
652 These NClimgrid relevance maps are shown in Figure 10 for ANNs trained on SPEAR_MED
653 but using different regularization parameters. For lower ridge parameters, we find greater
654 noise when interpreting the XAI maps, but overall higher positive relevance areas over
655 portions of the Western United States in the vicinity of topography (Figure 10a-c). As
656 the ridge parameter increases (i.e., penalizing larger weights to spread out the impor-

tance more evenly to reduce noisy collinearity), the spatial patterns of the XAI maps are smoothed, but note that the error of the ANN testing data skill begins to subsequently grow too (Figure S6). For the NClimGrid relevance maps in Figure 10d-f, we find that ANN is leveraging temperature patterns in the Eastern and Western United States (positive relevance) to make its yearly predictions. Notably, this spatial pattern of relevance somewhat resembles the summertime warming hole (Mascioli et al., 2017).

#### *4.2.2 Influence of Horizontal Resolution*

As this stage, it is evident that the ANNs are leveraging different spatial features to predict the year depending on the availability of fine detail on a given map grid, especially at temperature gradients around high and low geographic elevations. To evaluate this finding more closely, we now compare our regional results in Figure 7 with ANNs trained and tested on coarser regional maps from SPEAR_LO. Figure 7d-f shows these regional predictions of JJA TAVG and overall indicates poorer testing ensemble member skill for the Western, Central, and Eastern United States. In fact, the ANN is unable to find any time-evolving signals in SPEAR_LO for the Eastern United States until the late 1990s. There is also greater spread in the yearly predictions after inputting LOW grid size maps from NClimGrid. As such, this result is consistent with Section 4.2 that found lower skill in predicting the year of CONUS maps after training on SPEAR_LO compared to SPEAR_MED. To better compare the prediction skill between the two different spatial resolutions of SPEAR_MED and SPEAR_LO, we show in Figure 11 the distribution of MAE scores across 20 different ANNs with the same architecture, but for different combinations of training, testing, validation data and random initialization seeds. We also compare the MAE of ANNs trained on the FLOR climate model large ensemble (MED grid) compared to ANNs trained on maps of CONUS from FLOR, but interpolated onto the LOW grid spacing (FLOR (LO); see Section 2.1). For both of these ANN experiments, we again find lower MAE scores for the ANNs trained on the maps with higher spatial resolution (MED grid). This performance is also contrasted to the experiment without any anthropogenic forcing (not shown). The results of SPEAR_MED_NATURAL show a median MAE score over 1921 to 2100 for its best performing ANN architecture that is about 22 years, more than four times the MAE of the ensemble with anthropogenic forcing (lowest MAE on one ANN iteration/seed is still more than 18 years).

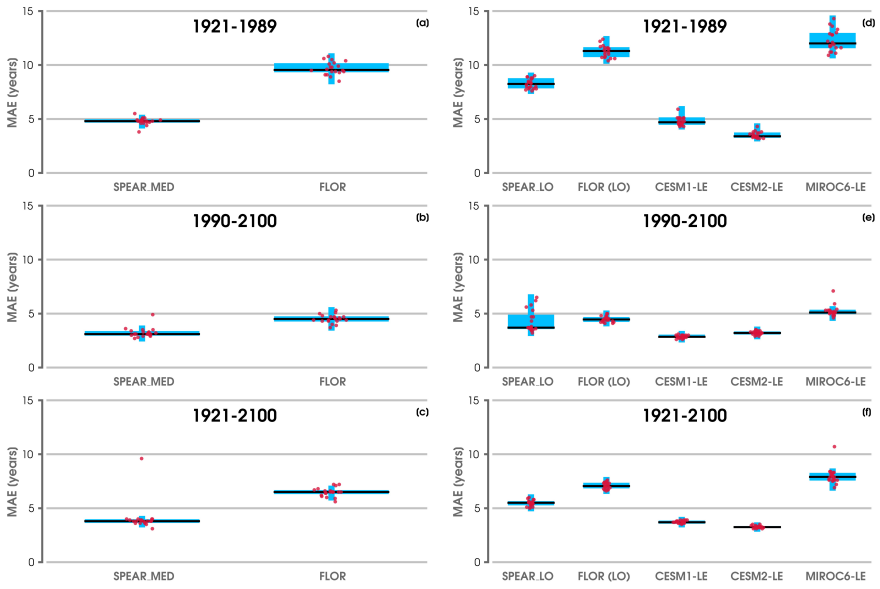


**Figure 11.** (a) Distribution of MAE scores for validation predictions over 1921-1989 based on inputs of CONUS maps for the overall ANN architecture with the lowest median MAE (e.g., Figure S6) after training neural networks on individual climate model large ensembles with the MED resolution (see Section 3.1) (SPEAR_MED or FLOR). Each distribution of scores (red points) is constructed from 20 ANN iterations (different combinations of training, testing, and validation ensemble members and random initialization seeds). The median score is shown with a thin black horizontal line. (b) As in (a), but for MAE scores calculated over 1990-2100. (c) As in (a), but for MAE scores calculated over 1921-2100. (d-f) As in (a-c), but for ANNs trained on individual large ensembles with the LOW resolution (SPEAR_LO, FLOR (LOW), CESM1-LE, CESM2-LE, MIROC6-LE).

Figure 11 also compares the skill using SPEAR and FLOR to other climate model large ensembles, but this is only possible for their coarser resolution (LOW grid). Overall, the largest error in predicting the year of summertime TAVG maps is found for MIROC6-LE. Notably, and despite the coarser resolution, the lowest error is found for CESM2-LE across all climate model large ensembles for both before and after 1990. This is examined more closely in Figure S12 for a single ANN using MIROC6-LE, CESM1-LE, and CESM2-LE along with the predictions after inputting observations from NClimGrid after training each network. Taking into consideration that there are more ensemble members for these three climate models (compared to 30 total members in SPEAR and FLOR), one possibility for the better skill in CESM2-LE is greater availability of training data. Therefore, we conduct three more ANNs that are shown in Figure S12d-f, but use the same number of training and testing data ensemble members as done with SPEAR. Close results are found, and for that reason it is unlikely that the differences in skill are due to more training data ensemble members. Similar to the testing data results, there is greater spread in the NClimGrid predictions after training on MIROC6-LE (Figure S12a,d). Meanwhile, the NClimGrid predictions after training on CESM1-LE and CESM2-LE agree broadly well with those from the SPEAR ANNs, but a TAVG signal again only emerges by the late 1990s (i.e., predictions closer to the 1:1 line).

### 4.3 Early 20th Century Temperature Signals in SPEAR

A remaining question still across all of the results is how the ANN is able to distinguish summertime temperature maps for the climate model large ensemble data prior to the late 20th century. Stated another way, what signals existed prior to 1990 when the forced ensemble mean warming trend hasn't started to clearly emerge yet (e.g., Figure 4)? One advantage of using feature attribution XAI methods is that a relevance map is obtained for each input sample. Accordingly, it's then possible to take XAI composites over different temporal periods to better understand the time evolution of the most relevant temperature signals. Figure 12 shows the relevance maps using the $\text{LRP}_z$ and Integrated Gradients methods for composites of SPEAR_MED testing data before and after 1990. For the 1921 to 1989 period, both XAI methods reveal a hotspot over western Colorado and generally muted relevance elsewhere across the United States. A similar relevance pattern is also found for the XAI results based on inputs of TMAX (Fig-

ure S13). Apart from that, the relevance maps for 1990 to 2100 are more similar to those described in Figure S9.

We compare the effect of regularization parameter and ANN architecture choices on ANN skill for predicting the year of temperature maps from only 1921 to 1989. This is displayed in Figures S14 and S15 for ANNs trained on SPEAR_MED and SPEAR_LO, respectively. As previously discussed in Section 4.2.2, we again find better skill for the higher resolution spatial grid and when using a lower ridge parameter. Relevance maps for ANNs using different regularization parameters are composited for SPEAR_MED in Figure S16, which reveals this effect of the ANN using smaller regional information, like over the Rocky Mountains, for predicting the year of JJA temperature maps. With larger ridge parameters comes worse ANN skill (Figure S6) but improved interpretability for the XAI maps (Figure S16d-f). Correspondingly, it is then likely that the ANN is learning temperature indicators from finer spatial information, especially across western Colorado, during this early 20th century period.

In addition to testing data from SPEAR, the ANNs were also able to predict the year of CONUS maps from other climate model large ensembles (Figure 11 and Figure S12). We again turn to XAI methods in Figure S17, but alternatively for evaluating the temperature signals leveraged by the ANN after inputs of CESM2-LE. The spatial patterns of positive relevance in the early 20th century are different for CESM2-LE over 1921-1989 (Figure S17a-b) compared to SPEAR_MED (Figure 12a-b). Instead, there are positive areas of relevance derived from CESM2-LE over the Southeastern United States and again across portions of the Rocky Mountains but located north of the previous Colorado hotspot.

Although there are no long-term temperature trends in the time-mean of SPEAR_MED for the Western United States prior to around 1990 (Figure S2a), it is still possible that there are spatial patterns of temperature change. Figure S18 shows the linear TAVG trends for NClimGrid, SPEAR_MED, and SPEAR_MED_NATURAL from 1921 to 1950. While observations reveal a few areas of cooling across the West, most of the CONUS does not have any statistically significant temperature trends. In contrast, we find a patch of warming over Western Colorado in SPEAR_MED, which closely aligns with the previously identified XAI relevance hotspot (Figure 12a-b). Conspicuously, this warming is absent for TAVG trends in SPEAR_MED_NATURAL (Figure S18c), which indicates it could be

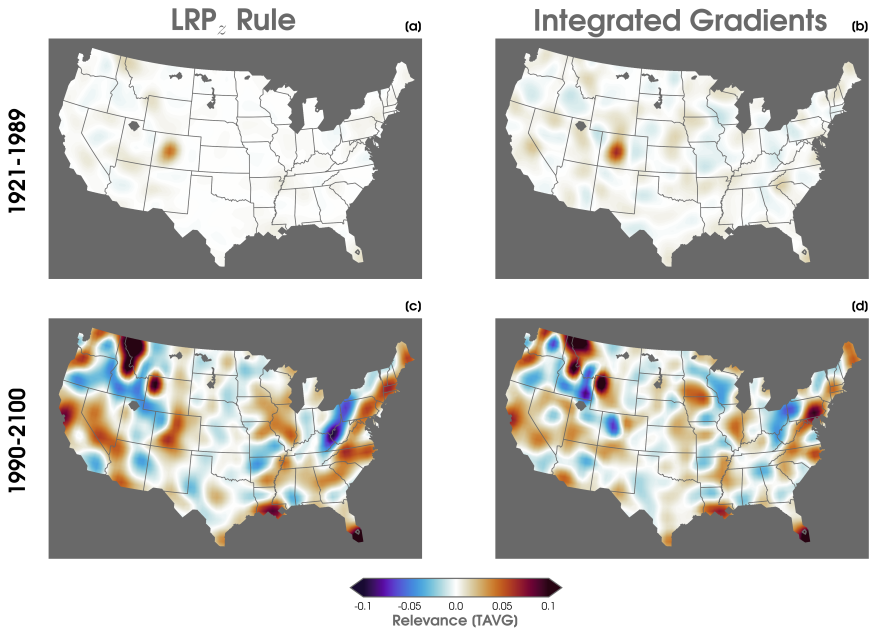


**Figure 12.** (a) Relevance heatmaps of TAVG using the layerwise relevance propagation z-rule (LRP$_z$) and (b) Integrated Gradients method for testing ensemble members from SPEAR_MED composited over 1921 to 1989. The composited heatmaps are smoothed using a Gaussian filter to improve interpretability. Positive relevance indicates regions that pushed the ANN to make its predicted year. Negative relevance suggests areas that tried to push the ANN away from its predicted year. (c-d) As in (a-b), but for composites over 1990-2100.

caused by anthropogenic forcing in SPEAR. We compare trends for TMAX and TMIN in Figure S19 and also find mainly similar results for the small warming trend over western Colorado in the ensemble mean of SPEAR_MED.

To understand the possible physical drivers for this local warming in SPEAR_MED, we calculate averages over a small box in western Colorado (outlined in Figure 2). We find that this warming in SPEAR is strongly correlated with decreasing evaporation rates (Figure S20a and Figure S21a), greater surface runoff, and even higher land surface temperature warming (not shown). Despite this evidence, we do not find any corresponding changes to precipitation or antecedent wintertime snowfall that could completely explain this effect (not shown). It is necessary to consider, however, that this small warming trend is still insignificant compared to the large spread of internal climate variability as simulated by the individual ensemble members over the region (Figure S20b-d). These trends are not found for the ensemble mean of SPEAR_MED_NATURAL (Figure S21b), as it is important to recall that land use and land change fields are also set to 1921 levels in SPEAR_MED_NATURAL. We hence hypothesize that this small warming signal over western Colorado could be related to the land surface forcing in SPEAR_MED, such as through the prescription of interactive vegetation influencing the surface energy budget, but this is outside the scope of this study, and more work would be needed to answer this question.

Finally, one last point we wish to make is that despite the ANN using this temperature signal over western Colorado to help predict the year of a given map, it is not the only reason for the high 1921-1989 skill (Figure 10). This is reflected in the results from Figure 9, which shows that the ANN can still reasonably predict the year even for regions of CONUS that do not include western Colorado.

## 5 Discussion and Conclusions

In this study, we used a machine learning approach to identify the ToE of summertime mean temperature extremes across the CONUS. There are several differences in this methodology compared to more traditional signal-to-noise metrics used in earlier ToE works. One advantage is that the ANN needs to learn time-evolving patterns of forced change to make an accurate prediction, instead of only comparing time-mean metrics over different epochs. In fact, we show that after training, the ANN is able to resolve these

forced signals even within a single ensemble member, i.e., one realization of internal climate variability. Changes in these temperature signals can then be visualized by applying XAI methods from one year to the next. Further, the ANN has the ability to learn and identify (non)linear relationships across the entire spatial map to associate with the ToE. These temperature patterns could differ from metrics calculated at the point-by-point level or aggregate over larger domains, which all could subsequently impact ToE estimates. By design, if observational predictions fall close to the 1:1 line, this suggests that climate change patterns in the training data are generalizable to the real-world. This is found to be the case in our ANN framework for the last two recent decades across different large ensemble climate models used for training.

We calculate the ANN-derived ToE by comparing to an early 20th century baseline period, which encompasses the record heatwaves of the Dust Bowl-era. Nonetheless, we find the emergence of a forced signal as early as the late 1990s for the observed TMIN in the Eastern United States. More broadly, we also find the emergence of summertime TAVG and TMIN across the entire conterminous United States. In other words, the ANN can still distinguish a climate signal in JJA temperature maps during recent years, despite the overall observed mean not exceeding the record warmth of the mid 1930s. It is also possible that if the size of the spatial domain of the temperature maps were increased, such as to consider all of North America, that the ToE may be identified even earlier than found in this work (e.g., Barnes et al., 2019; Sippel et al., 2020; Labe et al., 2023).

The ANNs are also able to make accurate predictions of the year for a given temperature map using the climate model large ensemble data, and they are skillful before the temperature response to greenhouse gas forcing overwhelms later in the 20th century. This suggests that the ANN is still able to leverage patterns of temperature indicators for distinguishing these temperature maps. We find that this is related to the higher spatial resolution of the training data particularly in the vicinity of complex topography. We perceive that this is not simply related to more available data samples, as we do not find any skill improvement when training on additional individual ensemble members. Rather it is more likely this is related to the ANN learning information from the high-resolution grid and thus the ability for a climate model to represent finer temperature structures and gradients.

To consider this point further, these results indicate that there is considerable potential to use the machine learning methods described in this study to detect the emergence of climate signals much earlier than with conventional methods, but the ability to realize this potential is likely hindered by climate model error or biases. For example, this could be analogous to the difference between potential predictability and forecast skill in climate prediction studies. If you calculated the ToE in the same way described here but using individual ensemble members as truth, you could define a "potential ToE" that would be much earlier than the actual ToE (again, analogous to predictability). The difference between the actual and potential ToE likely would reflect errors in the simulation of certain physical processes that the climate model deems important for distinguishing forced changes from internal variability. The ability to realize that potential ToE depends, however, on whether those processes are realistic and if the gap between simulated and observed changes can be narrowed. In this work, the gap also seems to be related to GCM resolution, as the earliest potential ToE generally occurs in the higher-resolution simulations. Similarly, we are also limited by the ability of most climate models to accurately simulate the temperature variability realized over the 20th century across the CONUS (i.e., the warming hole spatial pattern) (Eischeid et al., submitted).

Moving forward, our findings have several potential broader implications for future work related to using machine learning methods on climate science applications. While many XAI applications have regridded to coarser inputs because of lower computational cost, this may come at the expensive of better machine learning model performance that potentially could be achieved if using high-resolution data. For example, it could be interesting to further explore this effect for applications of machine learning in subseasonal to decadal prediction (Merryfield et al., 2020; Meehl et al., 2021), where neural networks may be able to derive more information from features such as simulated mesoscale eddy activity. But this remains an active area of research even for the climate model development community (Hewitt et al., 2017; Scaife et al., 2019). It could also be interesting to leverage these XAI tools for diagnosing biases in GCMs, such as the utility briefly explored here for identifying an unexpected temperature response in western Colorado which may be related to the land surface forcing fields.

However, having a greater number of input samples (e.g., higher resolution input map) can also raise the risk of statistical overfitting. It also tends to result in lower in-

terpretability for understanding the machine learning models, even after applying XAI attribution methods (Samek et al., 2019; Barnes et al., 2020; Toms et al., 2020). There are thus tradeoffs to balance in all of these machine learning design choices. Despite our observational ToE estimates, which are found to be robust across a range of ANN experiments and training data sets, we propose that it would be helpful for more work to investigate the sensitivity of machine learning model skill to variations in input data across a variety of climate science applications.

## Open Research

Atmospheric reanalysis data are openly available for ERA5 (`https://doi.org/10.24381/cds.6860a573`), which is supported by the Copernicus Climate Change Service (C3S; Thépaut et al., 2018) Climate Data Store (CDS). Twentieth Century Reanalysis Project version 3 (20CRv3) data are provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA (`https://psl.noaa.gov/data/gridded/data.20thC_ReanV3.html`). Monthly U.S. Climate Gridded Dataset (NClimGrid; version 1) data are provided by NOAA/NCEI (`https://doi.org/10.7289/V5SX6B56`). Climate model large ensembles used in this study are available from the Climate Data Gateway at NCAR (`https://www.earthsystemgrid.org/`), Earth System Grid Federation (`https://esgf-node.llnl.gov/search/cmip6/`), Facility for Weather and Climate Assessments (FACTS; Murray et al., 2020) (`https://www.psl.noaa.gov/repository/factsdocs`), and GFDL's data portal (`https://nomads.gfdl.noaa.gov/`). References for all data sets are provided throughout the main text.

Data preprocessing was completed using NCO v5.0.1 (Zender, 2008), CDO v1.9.10 (Schulzweida, 2019), and NCL v6.6.2 (NCAR, 2019). Computer code for the exploratory data analysis and machine learning architecture is available at `https://github.com/zmlabe/TOE_TMIN-TMAX`. *(Reviewers, please note that this GitHub URL will transition to an archived-DOI repository at Zenodo if this paper is considered for publication).* Python v3.9.13 (Rossum & Drake, 2009) packages used for the main results of this work include Numpy v1.22.4 (Harris et al., 2020), SciPy v1.8.1 (Virtanen et al., 2020), Scikit-learn v1.1.1 (Pedregosa et al., 2011), TensorFlow v2.7.0 (Abadi et al., 2016), and iNNvestigate v2.0.2 (Alber et al., 2019). Matplotlib v3.5.2 (Hunter, 2007) and Basemap v1.3.6 (*Basemap*, 2022) were used for plotting figures. Colormaps were provided by CMasher v1.6.3 (van der Velden, 2020), cmocean v2.0 (Thyng et al., 2016), Palettable's cubehelix v3.3.0 (Green, 2011), and Scientific v7.0.0 (Crameri, 2018; Crameri et al., 2020).

## Conflict of Interest

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


## Acknowledgments

We thank Drs. Liwei Jia and Yanda Zhang for helpful reviews on a previous draft of this manuscript. We also thank Dr. Bor-Ting Jong for other helpful discussions on land surface feedbacks. Zachary Labe acknowledges support under award NA18OAR4320123 from the National Oceanic and Atmospheric Administration (NOAA), U.S. Department of Commerce. This project is also supported by the NOAA Climate Program Office. Computational resources for this study was through base funding of the Geophysical Fluid Dynamics Laboratory (GFDL) provided by NOAA. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of NOAA, or the U.S. Department of Commerce.


## References


Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Zheng, X. (2016). Tensorflow: A system for large-scale machine learning..

Abatzoglou, J. T., & Barbero, R. (2014). Observed and projected changes in absolute temperature records across the contiguous united states. *Geophysical Research Letters*, *41*. doi: 10.1002/2014GL061441

Abatzoglou, J. T., & Williams, A. P. (2016). Impact of anthropogenic climate change on wildfire across western us forests. *Proceedings of the National Academy of Sciences of the United States of America*, *113*. doi: 10.1073/pnas.1607171113

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., ... Zhang, R. (2019, 10). The gfdl global ocean and sea ice model om4.0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, *11*, 3167-3211. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019MS001726` doi: 10.1029/2019MS001726

Agarap, A. F. (2018, 3). Deep learning using rectified linear units (relu). *arXiv*. Retrieved from `http://arxiv.org/abs/1803.08375`

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

*Learning Research*, *20*.

Alter, R. E., Douglas, H. C., Winter, J. M., & Eltahir, E. A. (2018). Twentieth century regional climate change during the summer in the central united states attributed to agricultural intensification. *Geophysical Research Letters*, *45*. doi: 10.1002/2017GL075604

a. Meehl, G., Arblaster, J. M., & Branstator, G. (2012, 9). Mechanisms contributing to the warming hole and the consequent u.s. east–west differential of heat extremes. *Journal of Climate*, *25*, 6394-6408. Retrieved from http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-11-00655.1 doi: 10.1175/JCLI-D-11-00655.1

Anderson, C., & Stock, J. (2022, 12). An interpretable model of climate change using correlative learning.. Retrieved from https://arxiv.org/abs/2212.04478v1

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

Banerjee, A., Polvani, L. M., & Fyfe, J. C. (2017). The united states "warming hole": Quantifying the forced aerosol response given large internal variability. *Geophysical Research Letters*, *44*. doi: 10.1002/2016GL071567

Barnes, E. A., Anderson, C., & Ebert-Uphoff, I. (2018, 11). An ai approach to determining the time of emergence of climate change. In (p. 19-22). NCAR. Retrieved from https://opensky.ucar.edu/islandora/object/technotes:571 doi: http://dx.doi.org/10.5065/D6BZ64XQ

Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019, 11). Viewing forced climate patterns through an ai lens. *Geophysical Research Letters*, *46*, 13389-13398. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2019GL084944 doi: 10.1029/2019GL084944

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020, 9). Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, *12*. Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2020MS002195 doi: 10.1029/2020MS002195

*Basemap.* (2022, 10). Retrieved from https://github.com/matplotlib/basemap

Bevacqua, E., Suarez-Gutierrez, L., Jézéquel, A., Lehner, F., Vrac, M., Yiou, P., & Zscheischler, J. (2023, 4). Advancing research on compound weather and climate events via large ensemble model simulations. *Nature Communications 2023 14:1*, *14*, 1-16. Retrieved from https://www.nature.com/articles/s41467-023-37847-5 doi: 10.1038/s41467-023-37847-5

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

Bommer, P., Kretschmer, M., Hedström, A., Bareeva, D., & Höhne, M. M. C. (2023, 3). Finding the right xai method – a guide for the evaluation and ranking of explainable ai methods in climate science. *arXiv*. Retrieved from https://arxiv.org/abs/2303.00652v1

Boukabara, S.-A., Krasnopolsky, V., Penny, S. G., Stewart, J. Q., McGovern, A., Hall, D., ... Hoffman, R. N. (2021, 5). Outlook for exploiting artificial intelligence in the earth and environmental sciences. *Bulletin of the American Meteorological Society*, *102*, E1016-E1032. Retrieved from https://journals.ametsoc.org/view/journals/bams/102/5/BAMS-D-20-0031.1.xml doi: 10.1175/BAMS-D-20-0031.1

Burgess, M. G., Ritchie, J., Shapland, J., & Pielke, R. (2020, 12). Ipcc baseline scenarios have over-projected co2 emissions and economic growth. *Environmental Research Letters*, *16*, 014016. Retrieved from https://iopscience.iop.org/article/10.1088/1748-9326/abcdd2 doi: 10.1088/1748-9326/ABCDD2

Chantry, M., Christensen, H., Dueben, P., & Palmer, T. (2021, 4). Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft ai. *Philosophical Transactions of the Royal Society A*, *379*. Retrieved from https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0083 doi: 10.1098/RSTA.2020.0083

Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., & McGovern, A. (2022, 8). A machine learning tutorial for operational meteorology. part i: Traditional machine learning. *Weather and Forecasting*, *37*, 1509-1529. Retrieved from https://journals.ametsoc.org/view/journals/wefo/37/8/WAF-D-22-0070.1.xml doi: 10.1175/WAF-D-22-0070.1

Chase, R. J., Harrison, D. R., Lackmann, G., & McGovern, A. (2022, 10). A machine learning tutorial for operational meteorology, part ii: Neural networks and deep learning. *arXiv*. Retrieved from https://arxiv.org/abs/2211.00147v1 doi: 10.48550/arxiv.2211.00147

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., ... Worley, S. J. (2011, 1). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, *137*, 1-28. doi: 10.1002/qj.776

Cowan, T., Hegerl, G. C., Colfescu, I., Bollasina, M., Purich, A., & Boschat, G. (2017). Factors contributing to record-breaking heat waves over the great plains during the 1930s dust bowl. *Journal of Climate*, *30*. doi: 10.1175/JCLI-D-16-0436.1

Cowan, T., Undorf, S., Hegerl, G. C., Harrington, L. J., & Otto, F. E. (2020, 5). Present-day greenhouse gases could cause more frequent and longer dust bowl heatwaves. *Nature Climate Change*, *10*, 505-510. Retrieved from https://www.nature.com/articles/s41558-020-0771-7 doi: 10.1038/s41558-020-0771-7

Crameri, F. (2018, 1). Scientific colour maps. *Zenodo*. Retrieved from https://zenodo.org/record/4153113 doi: 10.5281/ZENODO.4153113

Crameri, F., Shephard, G. E., & Heron, P. J. (2020, 12). The misuse of colour in science communication. *Nature Communications*, *11*, 1-10. Retrieved from https://doi.org/10.1038/s41467-020-19160-7 doi: 10.1038/s41467-020-19160-7

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., ... Strand, W. G. (2020, 2). The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001916. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019MS001916 doi: 10.1029/2019MS001916

Delworth, T. L., Cooke, W. F., Adcroft, A., Bushuk, M., Chen, J.-H., Dunne, K. A., ... Zhao, M. (2020, 3). Spear: The next generation gfdl modeling system for seasonal to multidecadal prediction and projection. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001895. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019MS001895

doi: 10.1029/2019MS001895

Delworth, T. L., Cooke, W. F., Naik, V., Paynter, D., & Zhang, L. (2022, 8). A weakened amoc may prolong greenhouse gas–induced mediterranean drying even with significant and rapid climate change mitigation. *Proceedings of the National Academy of Sciences of the United States of America*, *119*, e2116655119. Retrieved from `https://www.pnas.org/doi/abs/10.1073/pnas.2116655119` doi: 10.1073/PNAS.2116655119/SUPPL_FILE/PNAS.2116655119.SAPP.PDF

Delworth, T. L., Rosati, A., Anderson, W., Adcroft, A. J., Balaji, V., Benson, R., ... Zhang, R. (2012, 4). Simulated climate and climate change in the gfdl cm2.5 high-resolution coupled climate model. *Journal of Climate*, *25*, 2755-2781. Retrieved from `https://journals.ametsoc.org/view/journals/clim/25/8/jcli-d-11-00316.1.xml` doi: 10.1175/JCLI-D-11-00316.1

Deser, C. (2020, 11). Certain uncertainty: The role of internal climate variability in projections of regional climate change and risk management. *Earth's Future*. Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2020EF001854` doi: 10.1029/2020EF001854

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., ... Ting, M. (2020, 3). Insights from earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, 1-10. Retrieved from `http://www.nature.com/articles/s41558-020-0731-2` doi: 10.1038/s41558-020-0731-2

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012, 2). Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, *38*, 527-546. Retrieved from `http://link.springer.com/10.1007/s00382-010-0977-x` doi: 10.1007/s00382-010-0977-x

Deser, C., Terray, L., & Phillips, A. S. (2016). Forced and internal components of winter air temperature trends over north america during the past 50 years: Mechanisms and implications. *Journal of Climate*, *29*. doi: 10.1175/JCLI-D-15-0304.1

Diffenbaugh, N. S., & Barnes, E. A. (2023, 2). Data-driven predictions of the time remaining until critical global warming thresholds are reached. *Proceedings of the National Academy of Sciences of the United States of Amer-*

*ica*, *120*, e2207183120. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2207183120 doi: 10.1073/PNAS.2207183120/SUPPL_FILE/PNAS.2207183120.SAPP.PDF

Diffenbaugh, N. S., Swain, D. L., Touma, D., & Lubchenco, J. (2015). Anthropogenic warming has increased drought risk in california. *Proceedings of the National Academy of Sciences of the United States of America*, *112*. doi: 10.1073/pnas.1422385112

Donat, M. G., King, A. D., Overpeck, J. T., Alexander, L. V., Durre, I., & Karoly, D. J. (2016, 1). Extraordinary heat during the 1930s us dust bowl and associated large-scale conditions. *Climate Dynamics*, *46*, 413-426. Retrieved from https://link.springer.com/article/10.1007/s00382-015-2590-5 doi: 10.1007/S00382-015-2590-5/FIGURES/7

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., & Vose, R. S. (2010, 8). Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, *49*, 1615-1633. Retrieved from https://journals.ametsoc.org/view/journals/apme/49/8/2010jamc2375.1.xml doi: 10.1175/2010JAMC2375.1

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016, 5). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, *9*, 1937-1958. doi: 10.5194/gmd-9-1937-2016

Ferguson, C. R., & Villarini, G. (2012). Detecting inhomogeneities in the twentieth century reanalysis over the central united states. *Journal of Geophysical Research Atmospheres*, *117*. doi: 10.1029/2011JD016988

Ferguson, C. R., & Villarini, G. (2014, 11). An evaluation of the statistical homogeneity of the twentieth century reanalysis. *Climate Dynamics*, *42*, 2841-2866. Retrieved from https://link.springer.com/article/10.1007/s00382-013-1996-1 doi: 10.1007/S00382-013-1996-1/FIGURES/16

Fischer, E. M., Beyerle, U., & Knutti, R. (2013, 11). Robust spatially aggregated projections of climate extremes. *Nature Climate Change*, *3*, 1033-1038. Retrieved from https://www.nature.com/articles/nclimate2051 doi: 10.1038/nclimate2051

Fischer, E. M., & Knutti, R. (2014, 1). Detection of spatially aggregated changes in temperature and precipitation extremes. *Geophysical Research Letters*, *41*, 547-554. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2013GL058499 doi: 10.1002/2013GL058499

François, B., & Vrac, M. (2023, 1). Time of emergence of compound events: contribution of univariate and dependence properties. *Natural Hazards and Earth System Sciences*, *23*, 21-44. doi: 10.5194/NHESS-23-21-2023

Friedman, J. H. (2012, 7). Fast sparse regression and classification. *International Journal of Forecasting*, *28*, 722-738. doi: 10.1016/j.ijforecast.2012.05.001

Gillespie, I., Haimberger, L., Compo, G. P., & Thorne, P. W. (2023, 2). Assessing homogeneity of land surface air temperature observations using sparse-input reanalyses. *International Journal of Climatology*, *43*, 736-760. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.7822 doi: 10.1002/JOC.7822

Giorgi, F., & Bi, X. (2009, 3). Time of emergence (toe) of ghg-forced precipitation change hot-spots. *Geophysical Research Letters*, *36*. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2009GL037593 doi: 10.1029/2009GL037593

Gnanadesikan, A., Dixon, K. W., Griffies, S. M., Balaji, V., Barreiro, M., Beesley, J. A., ... Dunne, J. P. (2006, 3). Gfdl's cm2 global coupled climate models. part ii: The baseline ocean simulation. *Journal of Climate*, *19*, 675-697. Retrieved from https://journals.ametsoc.org/view/journals/clim/19/5/jcli3630.1.xml doi: 10.1175/JCLI3630.1

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

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

Grotjahn, R., & Huynh, J. (2018). Contiguous us summer maximum temperature and heat stress trends in cru and noaa climate division data plus comparisons to reanalyses. *Scientific Reports*, *8*. doi: 10.1038/s41598-018-29286-w

Hansen, J., Ruedy, R., Sato, M., Imhoff, M., Lawrence, W., Easterling, D., ... Karl, T. (2001). A closer look at united states and global surface temperature change. *Journal of Geophysical Research Atmospheres*, *106*. doi: 10.1029/2001JD000354

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, 9). Array programming with numpy. *Nature*, *585*, 357. Retrieved from https://doi.org/10.1038/s41586-020-2649-2 doi: 10.1038/s41586-020-2649-2

Hausfather, Z., & Peters, G. P. (2020). *Rcp8.5 is a problematic scenario for near-term emissions* (Vol. 117). doi: 10.1073/pnas.2017124117

Hawkins, E., Anderson, B., Diffenbaugh, N., Mahlstein, I., Betts, R., Hegerl, G., ... Vecchi, G. (2014, 7). Uncertainties in the timing of unprecedented climates. *Nature 2014 511:7507*, *511*, E3-E5. Retrieved from https://www.nature.com/articles/nature13523 doi: 10.1038/nature13523

Hawkins, E., Frame, D., Harrington, L., Joshi, M., King, A., Rojas, M., & Sutton, R. (2020, 3). Observed emergence of the climate change signal: From the familiar to the unknown. *Geophysical Research Letters*, *47*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2019GL086259 doi: 10.1029/2019GL086259

Hawkins, E., & Sutton, R. (2012). Time of emergence of climate signals. *Geophysical Research Letters*. doi: 10.1029/2011GL050087

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., ... Zadeh, N. (2019, 11). Structure and performance of gfdl's cm4.0 climate model. *Journal of Advances in Modeling Earth Systems*, *11*, 3691-3727. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019MS001829 doi: 10.1029/2019MS001829

Henson, S. A., Beaulieu, C., Ilyina, T., John, J. G., Long, M., Séférian, R., ... Sarmiento, J. L. (2017, 3). Rapid emergence of climate change in environmental drivers of marine ecosystems. *Nature Communications 2017 8:1*, *8*, 1-9. Retrieved from https://www.nature.com/articles/ncomms14682 doi: 10.1038/ncomms14682

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020, 5). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803 doi: 10.1002/qj.3803

Hewitt, H. T., Bell, M. J., Chassignet, E. P., Czaja, A., Ferreira, D., Griffies, S. M., ... Roberts, M. J. (2017). *Will high-resolution global ocean models bene-*

*fit coupled predictions on short-range to climate timescales?* (Vol. 120). doi: 10.1016/j.ocemod.2017.11.002

Hunter, J. D. (2007, 5). Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, *9*, 99-104. doi: 10.1109/MCSE.2007.55

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

Ignjacevic, P., Estrada, F., & Botzen, W. J. (2021, 7). Time of emergence of economic impacts of climate change. *Environmental Research Letters*, *16*, 074039. Retrieved from `https://iopscience.iop.org/article/10.1088/1748-9326/ac0d7a/meta` doi: 10.1088/1748-9326/AC0D7A

IPCC, Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., ... B., Z. (2021). *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change.*

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

Jia, L., Delworth, T. L., Kapnick, S., Yang, X., JOHNSON, N. C., Cooke, W., ... Tseng, K. C. (2022, 7). Skillful seasonal prediction of north american summertime heat extremes. *Journal of Climate*, *35*, 4331-4345. Retrieved from `https://journals.ametsoc.org/view/journals/clim/35/13/JCLI-D-21-0364.1.xml` doi: 10.1175/JCLI-D-21-0364.1

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., ... Vertenstein, M. (2015, 8). The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*, 1333-1349. Retrieved from `http://journals.ametsoc.org/doi/10.1175/BAMS-D-13-00255.1` doi: 10.1175/BAMS-D-13-00255.1

King, A. D., Donat, M. G., Fischer, E. M., Hawkins, E., Alexander, L. V., Karoly, D. J., ... Perkins, S. E. (2015, 9). The timing of anthropogenic emergence in simulated climate extremes. *Environmental Research Letters*, *10*, 094015. Retrieved from `https://iopscience.iop.org/article/10.1088/1748-9326/10/9/094015/meta` doi: 10.1088/1748-9326/10/9/094015

Kriegler, E., Bauer, N., Popp, A., Humpenöder, F., Leimbach, M., Strefler, J., ... Edenhofer, O. (2017, 1). Fossil-fueled development (ssp5): An energy and resource intensive scenario for the 21st century. *Global Environmental Change*, *42*, 297-315. doi: 10.1016/J.GLOENVCHA.2016.05.015

Kunkel, K. E., Liang, X. Z., Zhu, J., & Lin, Y. (2006). Can cgcms simulate the twentieth-century "warning hole" in the central united states? *Journal of Climate*, *19*. doi: 10.1175/JCLI3848.1

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*, e2021MS002464. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021MS002464` doi: 10.1029/2021MS002464

Labe, Z. M., & Barnes, E. A. (2022, 7). Comparison of climate model large ensembles with observations in the arctic using simple neural networks. *Earth and Space Science*, *9*, e2022EA002348. Retrieved from `https://doi.org/10.1029/2022EA002348` doi: 10.1029/2022EA002348

Labe, Z. M., Barnes, E. A., & Hurrell, J. W. (2023, 3). Identifying the regional emergence of climate patterns in the arise-sai-1.5 simulations. *Environmental Research Letters*, *18*, 1-12. Retrieved from `https://iopscience.iop.org/article/10.1088/1748-9326/acc81a` doi: 10.1088/1748-9326/ACC81A

Lawrence, J., Blackett, P., & Cradock-Henry, N. A. (2020, 1). Cascading climate change impacts and implications. *Climate Risk Management*, *29*, 100234. doi: 10.1016/J.CRM.2020.100234

Lecun, Y., Bengio, Y., & Hinton, G. (2015, 5). *Deep learning* (Vol. 521). Nature Publishing Group. Retrieved from `https://www.nature.com/articles/nature14539` doi: 10.1038/nature14539

Lehner, F., & Deser, C. (2023, 4). Origin, importance, and predictive limits of internal climate variability. *Environmental Research: Climate*. Re-

trieved from https://iopscience.iop.org/article/10.1088/2752-5295/accf30https://iopscience.iop.org/article/10.1088/2752-5295/accf30/meta doi: 10.1088/2752-5295/ACCF30

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E., Brunner, L., ... Hawkins, E. (2020). Partitioning climate projection uncertainty with multiple large ensembles and cmip5/6. *Earth System Dynamics Discussions*, 1-28. doi: 10.5194/esd-2019-93

Lehner, F., Deser, C., & Terray, L. (2017). Toward a new estimate of "time of emergence" of anthropogenic warming: Insights from dynamical adjustment and a large initial-condition model ensemble. *Journal of Climate*, *30*. doi: 10.1175/JCLI-D-16-0792.1

Lei, L., & Whitaker, J. S. (2016, 7). A four-dimensional incremental analysis update for the ensemble kalman filter. *Monthly Weather Review*, *144*, 2605-2621. Retrieved from https://journals.ametsoc.org/view/journals/mwre/144/7/mwr-d-15-0246.1.xml doi: 10.1175/MWR-D-15-0246.1

Leibensperger, E. M., Mickley, L. J., Jacob, D. J., Chen, W. T., Seinfeld, J. H., Nenes, A., ... Rind, D. (2012). Climatic effects of 1950-2050 changes in us anthropogenic aerosols-part 2: Climate response. *Atmospheric Chemistry and Physics*, *12*. doi: 10.5194/acp-12-3349-2012

Lempert, R., Nakicenovic, N., Sarewitz, D., & Schlesinger, M. (2004, 7). Characterizing climate-change uncertainties for decision-makers. an editorial essay. *Climatic Change*, *65*, 1-9. Retrieved from https://link.springer.com/article/10.1023/B:CLIM.0000037561.75281.b3 doi: 10.1023/B:CLIM.0000037561.75281.B3/METRICS

Madakumbura, G. D., Thackeray, C. W., Norris, J., Goldenson, N., & Hall, A. (2021, 7). Anthropogenic influence on extreme precipitation over global land areas seen in multiple observational datasets. *Nature Communications 2021 12:1*, *12*, 1-9. Retrieved from https://www.nature.com/articles/s41467-021-24262-x doi: 10.1038/s41467-021-24262-x

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

Mahlstein, I., Hegerl, G., & Solomon, S. (2012). Emerging local warming signals in observational data. *Geophysical Research Letters*, *39*. doi: 10.1029/2012GL053952

Mahony, C. R., & Cannon, A. J. (2018, 2). Wetter summers can intensify departures from natural variability in a warming climate. *Nature Communications*, *9*, 1-9. Retrieved from https://www.nature.com/articles/s41467-018-03132-z doi: 10.1038/s41467-018-03132-z

Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2023, 1). Carefully choose the baseline: Lessons learned from applying xai attribution methods for regression tasks in geoscience. *Artificial Intelligence for the Earth Systems*, *2*. Retrieved from https://journals.ametsoc.org/view/journals/aies/2/1/AIES-D-22-0058.1.xml doi: 10.1175/AIES-D-22-0058.1

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2022, 6). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science*, *1*, e8. Retrieved from https://www.cambridge.org/core/journals/environmental-data-science/article/neural-network-attribution-methods-for-problems-in-geoscience-a-novel-synthetic-benchmark-dataset/DDA562FC7B9A2B307105828619208 60E doi: 10.1017/EDS.2022.7

Mankin, J. S., Lehner, F., Coats, S., & McKinnon, K. A. (2020, 10). The value of initial condition large ensembles to robust adaptation decision-making. *Earth's Future*, *8*. Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2020EF001610 doi: 10.1029/2020EF001610

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*. doi: 10.2307/1907187

Mascioli, N. R., Previdi, M., Fiore, A. M., & Ting, M. (2017). Timing and seasonality of the united states 'warming hole'. *Environmental Research Letters*, *12*. doi: 10.1088/1748-9326/aa5ef4

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

Meehl, G. A., Arblaster, J. M., & Chung, C. T. (2015). Disappearance of the southeast u.s. "warming hole" with the late 1990s transition of the in-

terdecadal pacific oscillation. *Geophysical Research Letters*, *42*. doi: 10.1002/2015GL064586

Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F., ... Xie, S.-P. (2021, 4). Initialized earth system prediction from subseasonal to decadal timescales. *Nature Reviews Earth and Environment*, *2*, 340-357. Retrieved from https://www.nature.com/articles/s43017-021-00155-x doi: 10.1038/s43017-021-00155-x

Meehl, G. A., Tebaldi, C., & Adams-Smith, D. (2016). Us daily temperature records past, present, and future. *Proceedings of the National Academy of Sciences of the United States of America*, *113*. doi: 10.1073/pnas.1606117113

Meehl, G. A., Tebaldi, C., Walton, G., Easterling, D., & McDaniel, L. (2009). Relative increase of record high maximum temperatures compared to record low minimum temperatures in the u.s. *Geophysical Research Letters*, *36*. doi: 10.1029/2009GL040736

Meehl, G. A., Teng, H., Rosenbloom, N., Hu, A., Tebaldi, C., & Walton, G. (2022). How the great plains dust bowl drought spread heat extremes around the northern hemisphere. *Scientific Reports*, *12*. doi: 10.1038/s41598-022-22262-5

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012, 7). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, *29*, 897-910. Retrieved from https://journals.ametsoc.org/view/journals/atot/29/7/jtech-d-11-00103_1.xml doi: 10.1175/JTECH-D-11-00103.1

Menne, M. J., & Williams, C. N. (2009, 4). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, *22*, 1700-1717. Retrieved from https://journals.ametsoc.org/view/journals/clim/22/7/2008jcli2263.1.xml doi: 10.1175/2008JCLI2263.1

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A. S., ... Yeager, S. (2020, 6). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society*, *101*, E869-E896. Retrieved from https://journals.ametsoc.org/view/journals/bams/101/6/bamsD190037.xml doi: 10.1175/BAMS-D-19-0037.1

Milinski, S., Maher, N., & Olonscheck, D. (2020, 10). How large does a large ensemble need to be? *Earth System Dynamics*, *11*, 885-901. Retrieved from

https://esd.copernicus.org/articles/11/885/2020/ doi: 10.5194/esd-11-885-2020

Mills, E. (2005, 8). Insurance in a climate of change. *Science*, *309*, 1040-1044. Retrieved from https://www.science.org/doi/10.1126/science.1112121 doi: 10.1126/SCIENCE.1112121/ASSET/F852A595-6EC5-4DC5-A9C8-D41482F9C058/ASSETS/GRAPHIC/309_1040_F4.JPEG

Milly, P. C., Malyshev, S. L., Shevliakova, E., Dunne, K. A., Findell, K. L., Gleeson, T., ... Swenson, S. (2014, 10). An enhanced model of land water and energy for global hydrologic and earth-system studies. *Journal of Hydrometeorology*, *15*, 1739-1761. Retrieved from https://journals.ametsoc.org/view/journals/hydr/15/5/jhm-d-13-0162_1.xml doi: 10.1175/JHM-D-13-0162.1

Mora, C., Frazier, A. G., Longman, R. J., Dacks, R. S., Walton, M. M., Tong, E. J., ... Giambelluca, T. W. (2013, 10). The projected timing of climate departure from recent variability. *Nature*, *502*, 183-187. Retrieved from https://www.nature.com/articles/nature12540 doi: 10.1038/nature12540

Mueller, N. D., Butler, E. E., Mckinnon, K. A., Rhines, A., Tingley, M., Holbrook, N. M., & Huybers, P. (2016). Cooling of us midwest summer temperature extremes from cropland intensification. *Nature Climate Change*, *6*. doi: 10.1038/nclimate2825

Murray, D., Hoell, A., Hoerling, M., Perlwitz, J., Quan, X. W., Allured, D., ... Webb, R. S. (2020, 7). Facility for weather and climate assessments (facts): A community resource for assessing weather and climate variability. *Bulletin of the American Meteorological Society*, *101*, E1214-E1224. Retrieved from https://journals.ametsoc.org/view/journals/bams/101/7/bamsD190224.xml doi: 10.1175/BAMS-D-19-0224.1

NCAR. (2019). *The ncar command language (version 6.6.2).* Retrieved from http://dx.doi.org/10.5065/D6WD3XH5 doi: http://dx.doi.org/10.5065/D6WD3XH5

NCAR. (2020). *Us clivar multi-model le archive.* Retrieved from https://www.cesm.ucar.edu/projects/community-projects/MMLEA/

Neapolitan, R. E., & Jiang, X. (2018, 11). *Neural networks and deep learning.* Chapman and Hall/CRC. doi: 10.1201/b22400-15

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence o(1/k2). *Doklady AN USSR*, *269*.

O'Neill, B. C., Tebaldi, C., Vuuren, D. P. V., Eyring, V., Friedlingstein, P., Hurtt, G., ... Sanderson, B. M. (2016, 9). The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Development*, *9*, 3461-3482. doi: 10.5194/GMD-9-3461-2016

Pan, Z. T., Arritt, R. W., Takle, E. S., Gutowski, W. J., Anderson, C. J., & Segal, M. (2004, 9). Altered hydrologic feedback in a warming climate introduces a "warming hole". *Geophysical Research Letters*, *31*, L17109. doi: 10.1029/2004GL020528

Parker, D. E., Wilson, H., Jones, P. D., Christy, J. R., & Folland, C. K. (1996). The impact of mount pinatubo on world-wide temperatures. *International Journal of Climatology*, *16*. doi: 10.1002/(SICI)1097-0088(199605)16:5⟨487::AID-JOC39⟩3.0.CO;2-J

Parks, S. A., & Abatzoglou, J. T. (2020, 11). Warmer and drier fire seasons contribute to increases in area burned at high severity in western us forests from 1985 to 2017. *Geophysical Research Letters*, *47*, e2020GL089858. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020GL089858` doi: 10.1029/2020GL089858

Partridge, T. F., Winter, J. M., Osterberg, E. C., Hyndman, D. W., Kendall, A. D., & Magilligan, F. J. (2018, 2). Spatially distinct seasonal patterns and forcings of the u.s. warming hole. *Geophysical Research Letters*, *45*, 2055-2063. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2017GL076463` doi: 10.1002/2017GL076463

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

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

Peterson, T. C., Heim, R. R., Hirsch, R., Kaiser, D. P., Brooks, H., Diffenbaugh, N. S., ... Wuebbles, D. (2013). Monitoring and understanding changes in heat waves, cold waves, floods, and droughts in the united states: State of knowledge. *Bulletin of the American Meteorological Society*, *94*. doi:

10.1175/BAMS-D-12-00066.1

Pielke, R., Burgess, M. G., & Ritchie, J. (2022, 2). Plausible 2005–2050 emissions scenarios project between 2c and 3c of warming by 2100. *Environmental Research Letters*, *17*, 024027. Retrieved from https://iopscience.iop.org/article/10.1088/1748-9326/ac4ebf/meta doi: 10.1088/1748-9326/AC4EBF

Program, U. G. C. R. (2018). *Climate science special report: Fourth national climate assessment, volume i* (Vol. 1). Retrieved from https://science2017.globalchange.gov/downloads/CSSR2017_FullReport.pdf%0Ascience2017.globalchange.gov doi: 10.7930/J0J964J6

Rader, J. K., Barnes, E. A., Ebert-Uphoff, I., & Anderson, C. (2022, 7). Detection of forced change within combined climate fields using explainable neural networks. *Journal of Advances in Modeling Earth Systems*, *14*, e2021MS002941. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1029/2021MS002941 doi: 10.1029/2021MS002941

Rampal, N., Gibson, P. B., Sood, A., Stuart, S., Fauchereau, N. C., Brandolino, C., ... Meyers, T. (2022). High-resolution downscaling with interpretable deep learning: Rainfall extremes over new zealand. *Weather and Climate Extremes*, *38*. doi: 10.1016/j.wace.2022.100525

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., ... Rafaj, P. (2011, 11). Rcp 8.5-a scenario of comparatively high greenhouse gas emissions. *Climatic Change*, *109*, 33-57. Retrieved from https://link.springer.com/article/10.1007/s10584-011-0149-y doi: 10.1007/S10584-011-0149-Y/FIGURES/12

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., ... Tavoni, M. (2017, 1). The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, *42*, 153-168. doi: 10.1016/J.GLOENVCHA.2016.05.009

Rodgers, K. B., Lee, S.-S., Rosenbloom, N., Timmermann, A., Danabasoglu, G., Deser, C., ... Yeager, S. G. (2021, 12). Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, *12*, 1393-1411. Retrieved from https://esd.copernicus.org/articles/12/1393/2021/ doi:

10.5194/ESD-12-1393-2021

Rodgers, K. B., Lin, J., & Frölicher, T. L. (2015, 6). Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an earth system model. *Biogeosciences*, *12*, 3301-3320. doi: 10.5194/BG-12-3301-2015

Rossum, G. V., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Ruder, S. (2016, 9). An overview of gradient descent optimization algorithms. *arXiv*. Retrieved from `http://arxiv.org/abs/1609.04747`

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Muller, K.-R. (2019). Explainable ai: Interpreting, explaining and visualizing deep learning. *Lecture Notes in Computer Science (LNCS)*, *11700*.

Satoh, Y., Yoshimura, K., Pokhrel, Y., Kim, H., Shiogama, H., Yokohata, T., ... Oki, T. (2022, 6). The timing of unprecedented hydrological drought under climate change. *Nature Communications 2022 13:1*, *13*, 1-11. Retrieved from `https://www.nature.com/articles/s41467-022-30729-2` doi: 10.1038/s41467-022-30729-2

Scaife, A. A., Camp, J., Comer, R., Davis, P., Dunstone, N., Gordon, M., ... Vidale, P. L. (2019). Does increased atmospheric resolution improve seasonal climate predictions? *Atmospheric Science Letters*, *20*. doi: 10.1002/asl.922

Schlunegger, S., Rodgers, K. B., Sarmiento, J. L., Ilyina, T., Dunne, J. P., Takano, Y., ... Lehner, F. (2020, 8). Time of emergence and large ensemble intercomparison for ocean biogeochemical trends. *Global Biogeochemical Cycles*, *34*, e2019GB006453. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019GB006453` doi: 10.1029/2019GB006453

Schulzweida, U. (2019, 2). Cdo user guide. *Zenodo*. Retrieved from `https://zenodo.org/record/2558193` doi: 10.5281/ZENODO.2558193

Schwarzwald, K., & Lenssen, N. (2022). The importance of internal climate variability in climate impact projections. *Proceedings of the National Academy of Sciences of the United States of America*, *119*. doi: 10.1073/pnas.2208095119

Sear, C. B., Kelly, P. M., Jones, P. D., & Goodess, C. M. (1987). Global surface-temperature responses to major volcanic eruptions. *Nature*, *330*. doi: 10.1038/330365a0

Shin, N.-Y., Ham, Y.-G., Kim, J.-H., Cho, M., & Kug, J.-S. (2022). Application of deep learning to understanding enso dynamics. *Artificial Intelligence for the*

*Earth Systems*, *1*. doi: 10.1175/aies-d-21-0011.1

Shiogama, H., Tatebe, H., Hayashi, M., Abe, M., Arai, M., Koyama, H., ... Watanabe, M. (2023). Miroc6 large ensemble (miroc6-le): experimental design and initial analyses. *Earth Syst. Dynam. Discuss.*. Retrieved from https://doi.org/10.5194/esd-2023-12 doi: 10.5194/esd-2023-12

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

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

Slivinski, L. C., Compo, G. P., Sardeshmukh, P. D., Whitaker, J. S., McColl, C., Allan, R. J., ... Wyszynski, P. (2021, 2). An evaluation of the performance of the twentieth century reanalysis version 3. *Journal of Climate*, *34*, 1417-1438. Retrieved from https://journals.ametsoc.org/view/journals/clim/34/4/JCLI-D-20-0505.1.xml doi: 10.1175/JCLI-D-20-0505.1

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., ... Wyszyński, P. (2019, 10). Towards a more reliable historical reanalysis: Improvements for version 3 of the twentieth century reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*, 2876-2908. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3598 doi: 10.1002/qj.3598

Sonnewald, M., & Lguensat, R. (2021, 8). Revealing the impact of global heating on north atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, *13*, e2021MS002496. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021MS002496 doi: 10.1029/2021MS002496

Sundararajan, M., Taly, A., & Yan, Q. (2017, 3). Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, *7*, 5109-5118. Retrieved from https://arxiv.org/abs/1703.01365v2 doi: 10.48550/arxiv.1703.01365

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., ... Kimoto, M. (2019, 7). Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in miroc6. *Geoscientific Model Development*, *12*, 2727-2765. doi: 10.5194/GMD-12-2727-2019

Thompson, V., Kennedy-Asser, A. T., Vosper, E., Lo, Y. T. E., Huntingford, C., Andrews, O., ... Mitchell, D. (2022). The 2021 western north america heat wave among the most extreme events ever recorded globally. *Science Advances*, *8*. doi: 10.1126/sciadv.abm6860

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

Thépaut, J. N., Pinty, B., Dee, D., & Engelen, R. (2018). The copernicus programme and its climate change service. In (Vol. 2018-July). doi: 10.1109/IGARSS.2018.8518067

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020, 9). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*. Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2019MS002002` doi: 10.1029/2019MS002002

Trisos, C. H., Merow, C., & Pigot, A. L. (2020, 4). The projected timing of abrupt ecological disruption from climate change. *Nature 2020 580:7804*, *580*, 496-501. Retrieved from `https://www.nature.com/articles/s41586-020-2189-9` doi: 10.1038/s41586-020-2189-9

van der Velden, E. (2020, 2). Cmasher: Scientific colormaps for making accessible, informative and 'cmashing' plots. *Journal of Open Source Software*, *5*, 2004. Retrieved from `https://joss.theoj.org/papers/10.21105/joss.02004` doi: 10.21105/JOSS.02004

Vecchi, G. A., Delworth, T., Gudgel, R., Kapnick, S., Rosati, A., Wittenberg, A. T., ... Zhang, S. (2014, 11). On the seasonal forecasting of regional tropical cyclone activity. *Journal of Climate*, *27*, 7994-8016. Retrieved from `https://journals.ametsoc.org/view/journals/clim/27/21/`

jcli-d-14-00158.1.xml doi: 10.1175/JCLI-D-14-00158.1

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

Vose, R. S., Applequist, S., Squires, M., Durre, I., Menne, C. J., Williams, C. N., ... Arndt, D. (2014, 5). Improved historical temperature and precipitation time series for u.s. climate divisions. *Journal of Applied Meteorology and Climatology*, *53*, 1232-1251. Retrieved from https://journals.ametsoc.org/view/journals/apme/53/5/jamc-d-13-0248.1.xml doi: 10.1175/JAMC-D-13-0248.1

Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., ... Rose, S. K. (2011, 8). The representative concentration pathways: an overview. *Climatic Change*, *109*, 5-31. Retrieved from http://link.springer.com/10.1007/s10584-011-0148-z doi: 10.1007/s10584-011-0148-z

Weiskopf, S. R., Rubenstein, M. A., Crozier, L. G., Gaichas, S., Griffis, R., Halofsky, J. E., ... Whyte, K. P. (2020, 9). Climate change effects on biodiversity, ecosystems, ecosystem services, and natural resource management in the united states. *Science of The Total Environment*, *733*, 137782. doi: 10.1016/J.SCITOTENV.2020.137782

Williams, A. P., Cook, E. R., Smerdon, J. E., Cook, B. I., Abatzoglou, J. T., Bolles, K., ... Livneh, B. (2020). Large contribution from anthropogenic warming to an emerging north american megadrought. *Science*, *368*. doi: 10.1126/science.aaz9600

Wuebbles, D., Meehl, G., Hayhoe, K., Karl, T. R., Kunkel, K., Santer, B., ... Sun, L. (2014). Cmip5 climate model analyses: Climate extremes in the united states. *Bulletin of the American Meteorological Society*, *95*. doi: 10.1175/BAMS-D-12-00172.1

Wyser, K., Koenigk, T., Fladrich, U., Fuentes-Franco, R., Karami, M. P., & Kruschke, T. (2021, 7). The smhi large ensemble (smhi-lens) with ec-earth3.3.1. *Geoscientific Model Development*, *14*, 4781-4796. doi: 10.5194/GMD-14-4781-2021

Yang, X., Delworth, T. L., Jia, L., Johnson, N. C., Lu, F., & McHugh, C. (2022, 9). On the seasonal prediction and predictability of winter surface temperature swing index over north america. *Frontiers in Climate*, *4*, 167. doi: 10.3389/FCLIM.2022.972119/BIBTEX

Yu, B., Li, G., Chen, S., & Lin, H. (2020, 8). The role of internal variability in climate change projections of north american surface air temperature and temperature extremes in canesm2 large ensemble simulations. *Climate Dynamics*, *55*, 869-885. Retrieved from `https://link.springer.com/article/10.1007/s00382-020-05296-1` doi: 10.1007/S00382-020-05296-1/FIGURES/15

Yu, S., Alapaty, K., Mathur, R., Pleim, J., Zhang, Y., Nolte, C., ... Nagashima, T. (2014). Attribution of the united states "warming hole": Aerosol indirect effect and precipitable water vapor. *Scientific Reports*, *4*. doi: 10.1038/srep06929

Zadeh, L. A. (1965, 6). Fuzzy sets. *Information and Control*, *8*, 338-353. doi: 10.1016/S0019-9958(65)90241-X

Zender, C. S. (2008). Analysis of self-describing gridded geoscience data with netcdf operators (nco). *Environmental Modelling and Software*, *23*. doi: 10.1016/j.envsoft.2008.03.004

Zhao, M., Golaz, J. C., Held, I. M., Guo, H., Balaji, V., Benson, R., ... Xiang, B. (2018a, 3). The gfdl global atmosphere and land model am4.0/lm4.0: 1. simulation characteristics with prescribed ssts. *Journal of Advances in Modeling Earth Systems*, *10*, 691-734. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2017MS001208` doi: 10.1002/2017MS001208

Zhao, M., Golaz, J. C., Held, I. M., Guo, H., Balaji, V., Benson, R., ... Xiang, B. (2018b, 3). The gfdl global atmosphere and land model am4.0/lm4.0: 2. model description, sensitivity studies, and tuning strategies. *Journal of Advances in Modeling Earth Systems*, *10*, 735-769. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2017MS001209` doi: 10.1002/2017MS001209