Explainable AI for Climate Science

Opening the Black Box to Reveal Planet Earth

Dr. Elizabeth A. Barnes
Professor, Dept. of Atmospheric Science
Colorado State University

This research has been funded, in part, by grants from the NSF, NOAA, DOE, DARPA and with special thanks to David Wallerstein

SciPy, July 11 2024

A challenging future

Humans on the move

Unprecedented numbers of people have and will be impacted and seeking to escape from worsening climate hazards.

Earth system is threatened

Global environmental change has pushed society well outside of humanity's evolutionary experience, and critical ecosystem services are being threatened.

Anticipating what's to come

Predicting human-Earth system futures and change is critical for preparing society as we navigate our turbulent future

AI for climate

AI for climate is not new

But the field's interest and research has exploded in recent years!

Nonlinear Principal Component Analysis by Neural Networks: Theory and Application to the Lorenz System

ADAM H. MONAHAN

Oceanography Unit, Department of Earth and Ocean Sciences, and Crisis Points Group, Peter Wall Institute for Advanced Studies, University of British Columbia, Vancouver, British Columbia, Canada

1998

(Manuscript received 28 October 1998, in final form 7 May 1999)

ABSTRACT

A nonlinear generalization of principal component analysis (PCA), denoted nonlinear principal component analysis (NLPCA), is implemented in a variational framework using a five-layer autoassociative feed-forward neural network. The method is tested on a dataset sampled from the Lorenz attractor, and it is shown that the NLPCA approximations to the attractor in one and two dimensions, explaining 76% and 99.5% of the variance,

New Approach to Calculation of Atmospheric Model Physics: Accurate and I Neural Network Emulation of Longwave Radiation in a Climate Model

VLADIMIR M. KRASNOPOLSKY

NOAA/NCEP/SAIC, Camp Springs, and Earth System Science Interdisciplinary Center, University of Maryland, Colleg College Park, Maryland

MICHAEL S. FOX-RABINOVITZ AND DMITRY V. CHALIKOV

Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

2004

(Manuscript received 26 April 2004, in final form 25 October 2004)

ABSTRACT

A new approach based on a synergetic combination of statistical/machine learning and deterministic modeling within atmospheric models is presented. The approach uses neural networks as a statistical or machine learning technique for an accurate and fast emulation or statistical approximation of model physics parameterizations. It is applied to development of an accurate and fast approximation of an atmospheric longwave radiation parameterization for the NCAR Community Atmospheric Model, which is the most

An Application of Adaptive Logic to Meteorological Prediction

H. R. GLAHN

U. S. Weather Bureau, Washington, D. C.

1964

(Manuscript received 12 May 1964, in revised form 25 July 1964)

ABSTRACT

Two adaptive logic models and a training algorithm for each are described. These models are tested on a meteorological prediction problem with the use of large developmental and test data samples. Discriminant analysis is used for comparison. It is found that ceiling heights at Washington National Airport could be forecast better with the discriminant analysis technique than with the adaptive logic models.

1. Introduction

In recent years there has been an increase in the research effort concerning adaptive logic systems. These systems, including their associated hardware, are also known as learning machines or trainable networks. The desire is to create a system that can be trained to give the correct response when it is presented certain input data or stimuli in somewhat the same way a living organism might adjust to its environment or learn from past experience. It is also hoped that the system would be capable of certain generalizations so that when it is presented a data pattern that it has never encountered before it will still tend to give the correct response.

Adaptive logic models can be used for pattern recognition and, since the forcasting of certain weather elements involves, to some extent, the recognition of patterns, an application of adaptive systems to objective weather forecasting is thereby suggested.

2. Adaptive logic models

One adaptive model that has been described with minor variations by several authors, including Mattson,[1] Sebestyen (1962), Widrow (1962), Hu[2], Duda and Machanik[3] and Hu and Root (1964) is described below.

In Fig. 1 each "S" unit represents one of the $n+1$ numerical inputs to the system. Each "A" unit is an adjustable weight which is altered during the training procedure so that the process of summation, performed in the "Σ" unit, of $A_i S_i$, $i=0, 1, 2, \ldots, n$, and quantization into a binary output will provide the correct dichotomous classification. S_0 is held constant at $+1$ while S_i, $i=1, 2, \ldots, n$ may be different for different training sample points. This complete element has been called by Widrow (1962) an ADALINE for "adaptive linear neuron."

A training program for this model is indicated in Fig. 2. The value D is a positive constant which permits $\sum_{i=0}^{n} A_i S_i$ to have a value of D, rather than zero, immediately after an alteration of the A_i for the sample point which occasioned the alteration. After the output is made to agree with the desired output for a particular sample point, the A_i may again be changed for another sample point. This latter alteration may in turn cause the output to be incorrect for the former sample point. Repeated iterations over the data sample can be made and eventually the correct response may be given for all sample points if such distinction is possible with this model.

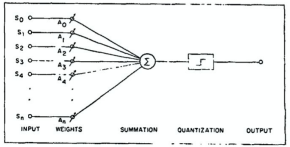
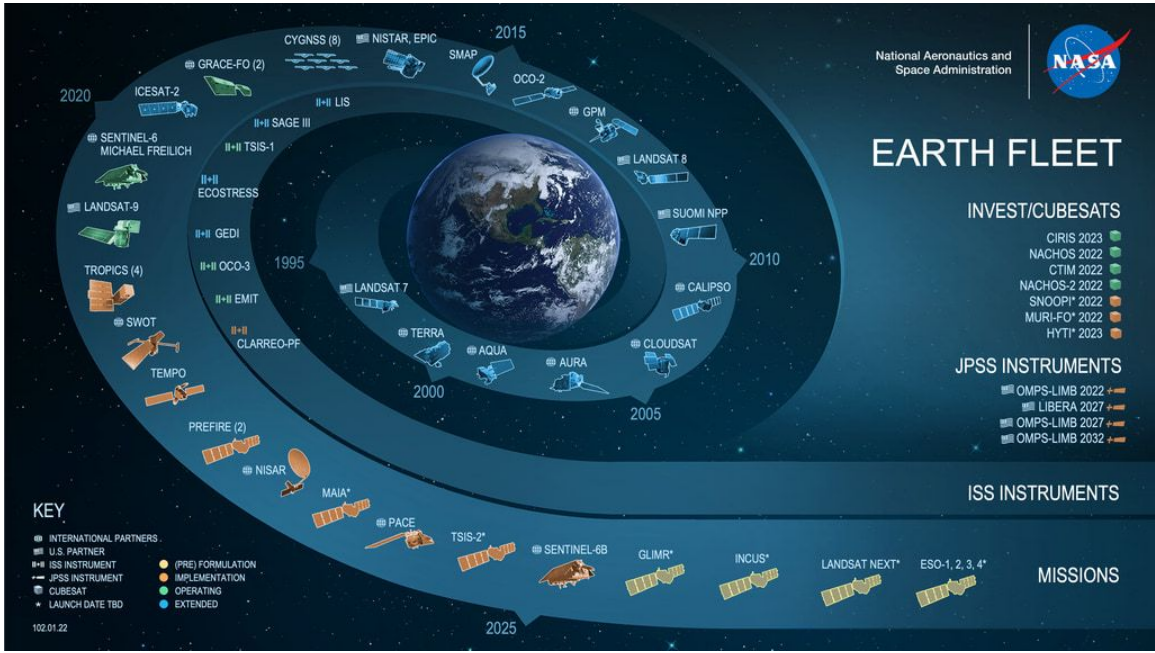


FIG. 1. A simple adaptive logic element.

[1] Mattson, R. L., 1959: The design and analysis of an adaptive system for statistical classification. S.M. Thesis, Massachusetts Institute of Technology, Cambridge, Mass., 60 pp.
[2] Hu, M. J., 1963: A trainable weather-forecasting system. Tech. Rep. No. 6759–1, Contract No. AF33(616)–7726, Systems Theory Laboratory, Stanford University, Stanford, Calif., 19 pp.
[3] Duda, R. O., and J. W. Machanik, 1963: An adaptive prediction technique and its application to weather forecasting. Paper presented at the Western Electronic Show and Convention, San Francisco, Calif.

So much data.

Our field has so much (imperfect) data: ground observations, space-borne observations, reanalyses, climate model output, etc…

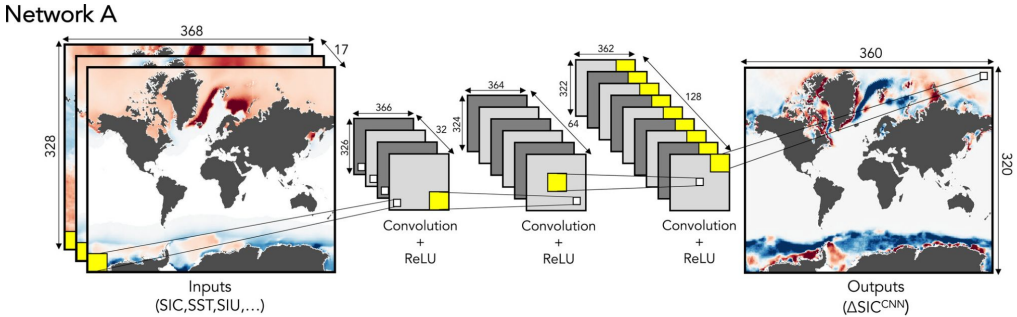


1

ML for post-processing data [e.g. data compression, data analysis]

Predicting the Errors of Forecast Systems

e.g. Chapman et al. (2019), Cahill et al. (in review), Pan et al. (2021), *Gregory et al. (2023)*



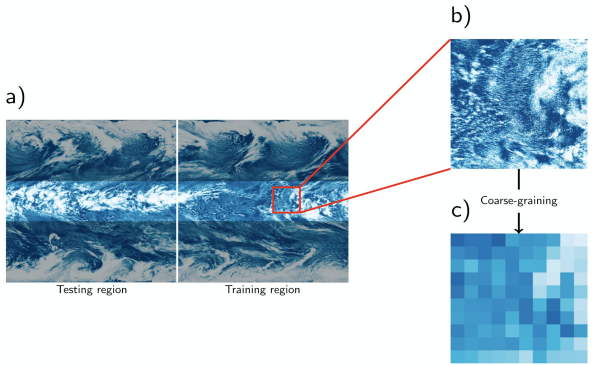
Machine Learning: a tool with many uses.

1 **ML for post-processing data**
[e.g. data compression, data analysis]

2 **ML to improve climate models**
[e.g. parameterizations]

Improved Model Parameterizations

e.g. Rasp et al. (2018; PNAS); Schneider et al. (2017; GRL); O'Gorman and Dwyer (2018); Beucler et al. (2020; PRL); Dagon et al. (2020); *Brenowitz and Bretherton (2018)*



Machine Learning: a tool with many uses.

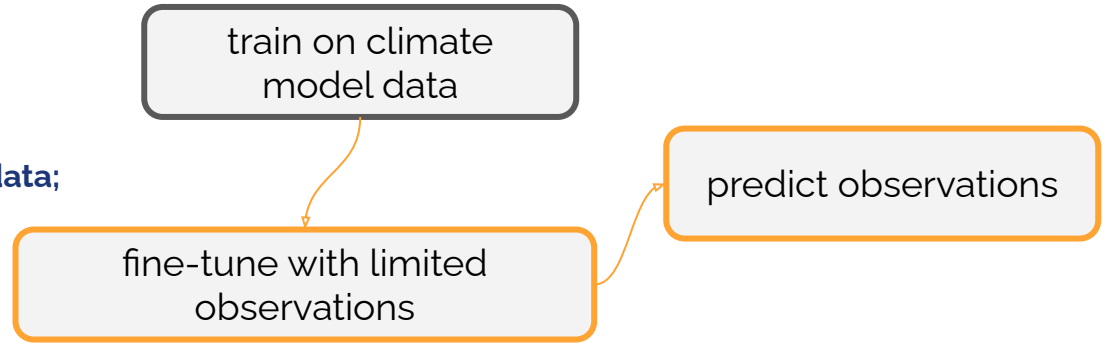
1. ML for post-processing data
[e.g. data compression, data analysis]

2. ML to improve climate models
[e.g. parameterizations]

3. Combine disparate datastreams to explore complex systems

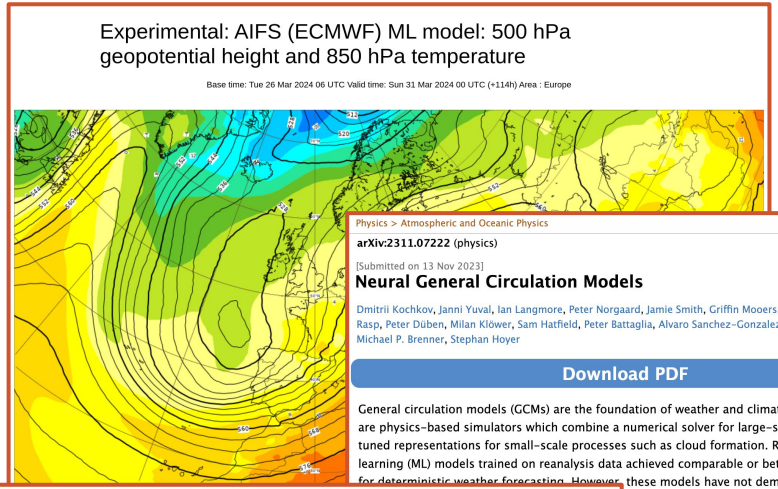
Machine Learning: a tool with many uses.

1. **ML for post-processing data [e.g. data compression, data analysis]**
2. **ML to improve climate models [e.g. parameterizations]**
3. **Combine disparate datastreams to explore complex systems**
4. **Merging observations and model data; [e.g. transfer learning]**



Machine Learning: a tool with many uses.

1. **ML for post-processing data [e.g. data compression, data analysis]**
2. **ML to improve climate models [e.g. parameterizations]**
3. **Combine disparate datastreams to explore complex systems**
4. **Merging observations and model data. [e.g. transfer learning]**
5. **Deep-learning weather + climate emulators**



Physics > Atmospheric and Oceanic Physics

arXiv:2311.07222 (physics)

[Submitted on 13 Nov 2023]

Neural General Circulation Models

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, James Lottes, Stephan Rasp, Peter Düben, Milan Klöwer, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, Stephan Hoyer

Download PDF

General circulation models (GCMs) are the foundation of weather and climate prediction. GCMs are physics-based simulators which combine a numerical solver for large-scale dynamics with tuned representations for small-scale processes such as cloud formation. Recently, machine learning (ML) models trained on reanalysis data achieved comparable or better skill than GCMs for deterministic weather forecasting. However, these models have not demonstrated improved
for long-term weather and climate
combines a differentiable solver for
d show that it can generate forecasts of
imate on par with the best ML and physics-
ML models for 1-10 day forecasts, and with the
recasts ensemble prediction for 1-15 day
ure, NeuralGCM can accurately track climate
ultiple decades, and climate forecasts with 140

Physics > Atmospheric and Oceanic Physics

arXiv:2310.02074 (physics)

[Submitted on 3 Oct 2023]

ACE: A fast, skillful learned global atmospheric model for climate prediction

Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K. Clark, Brian Henn, James Duncan, Noah D. Brenowitz, Karthik Kashinath, Michael S. Pritchard, Boris Bonev, Matthew E. Peters, Chris

Download PDF

Existing ML-based atmospheric models are not suitable for climate predict
long-term stability and physical consistency. We present ACE (AI2 Climate
parameter, autoregressive machine learning emulator of an existing compr
resolution global atmospheric model. The formulation of ACE allows evalua
such as the conservation of mass and moisture. The emulator is stable for
conserves column moisture without explicit constraints and faithfully repro
model's climate, outperforming a challenging baseline on over 80% of track
requires nearly 100x less wall clock time and is 100x more energy efficient
model using typically available resources.

Physics > Atmospheric and Oceanic Physics

[Submitted on 22 Mar 2024]

An ensemble of data-driven weather prediction models for operational sub-seasonal forecasting

Jonathan A. Weyn, Divya Kumar, Jeremy Berman, Najeeb Kazmi, Sylwester Klocek, Pete Luferenko, Kit Thambiratnam

We present an operations-ready multi-model ensemble weather forecasting system which uses hybrid data-driven weather prediction models coupled with the European Centre for Medium-range Weather Forecasts (ECMWF) ocean model to predict global weather at 1-degree resolution for 4 weeks of lead time. For predictions of 2-meter temperature, our ensemble on average outperforms the raw ECMWF extended-range ensemble by 4-17%, depending on the lead time. However, after applying statistical bias corrections, the ECMWF ensemble is about 3% better at 4 weeks. For other surface parameters, our ensemble is also within a few percentage points of ECMWF's ensemble. We demonstrate that it is possible to achieve near-state-of-the-art subseasonal-to-seasonal forecasts using a multi-model ensembling approach with data-driven weather prediction models.

Subjects: **Atmospheric and Oceanic Physics (physics.ao-ph)**; Machine Learning (cs.LG)
Cite as: arXiv:2403.15598 **[physics.ao-ph]**
(or arXiv:2403.15598v1 **[physics.ao-ph]** for this version)

Machine Learning: a tool with many uses.

1. ML for post-processing data
[e.g. data compression, data analysis]

2. ML to improve climate models
[e.g. parameterizations]

3. Combine disparate datastreams to explore complex systems

4. Merging observations and model data.
[e.g. transfer learning]

5. Deep-learning weather + climate emulators

6. Climate change communication

Mila

Machine Learning: a tool with many uses.

ILLUSTRATED BY HEARTWOOD VISUALS

1. ML for post-processing data
[e.g. data compression, data analysis]

2. ML to improve climate models
[e.g. parameterizations]

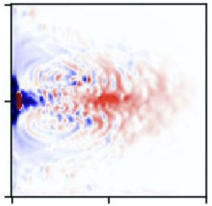
3. Combine disparate datastreams to explore complex systems

4. Merging observations and model data.
[e.g. transfer learning]

5. Deep-learning weather + climate emulators

6. Climate change communication

7. Learn new things!



$$\hat{\mathbf{S}}_{\mathbf{u}}^{BT} \approx \kappa_{BT} \overline{\nabla} \cdot \begin{pmatrix} \zeta^2 - \zeta D & \zeta \tilde{D} \\ \zeta \tilde{D} & \zeta^2 + \zeta D \end{pmatrix}$$

Equation Discovery
e.g. *Zanna & Bolton (2020)*

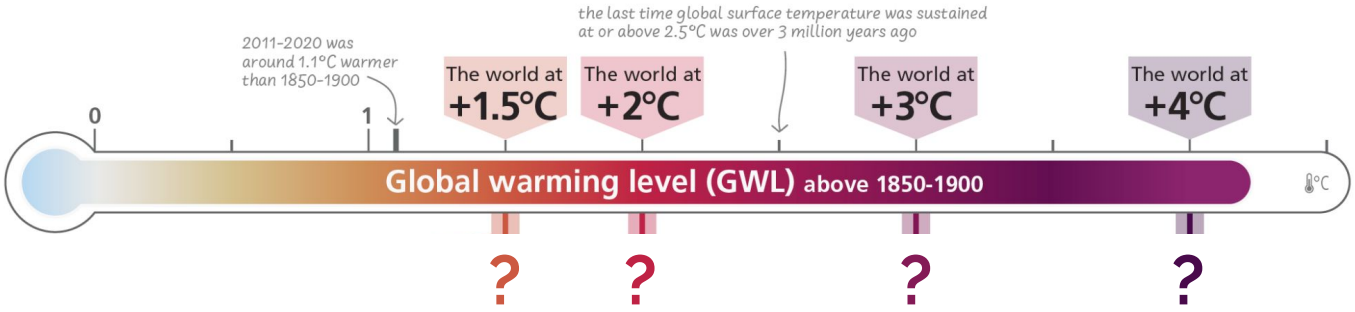
Machine Learning: a tool with many uses.

OUR GOAL:
To develop and implement AI tools to leverage imperfect climate models in support of earth system prediction across time and space.

Climate models provide inaccurate, but invaluable "parallel universes" to mine for information

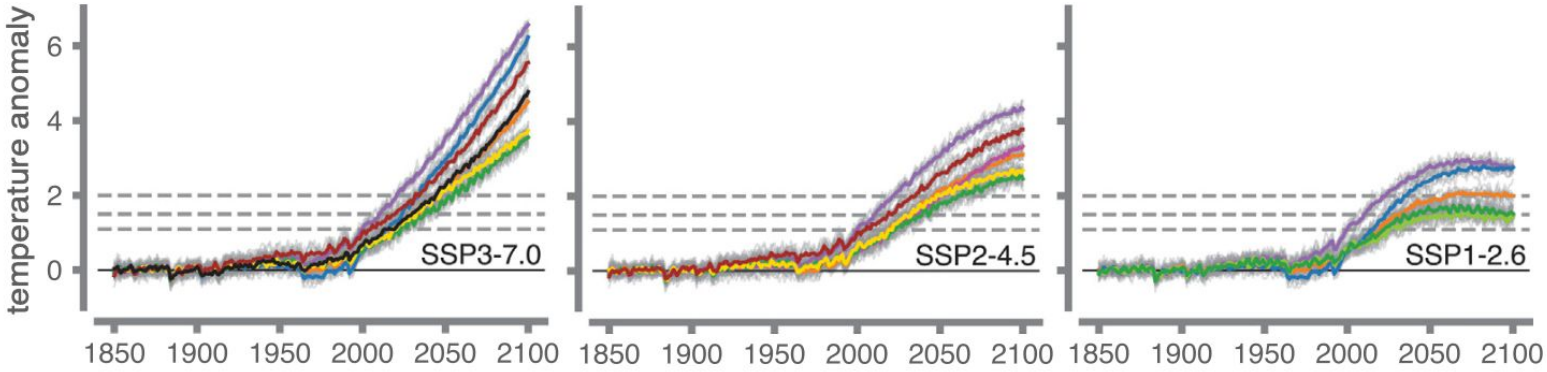
https://futurism.com/parallel-universes-many-worlds-theory

Time Remaining Until Critical Warming Thresholds are Reached



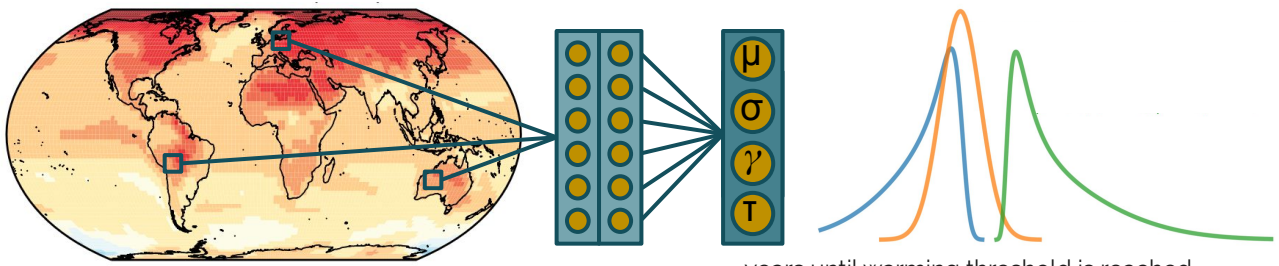
adapted from IPCC AR6 (2023); Summary for Policymakers

Time Remaining Until Critical Warming Thresholds are Reached



Diffenbaugh & Barnes (2023)

Trained on annual maps from 10 realizations from across multiple climate models

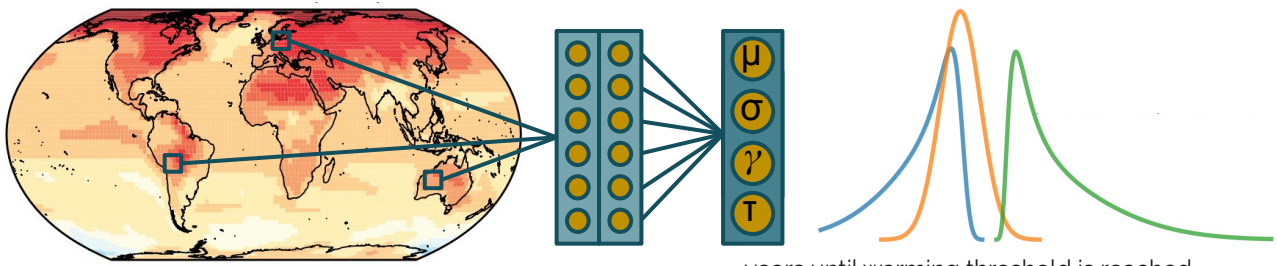


years until warming threshold is reached
+ uncertainties

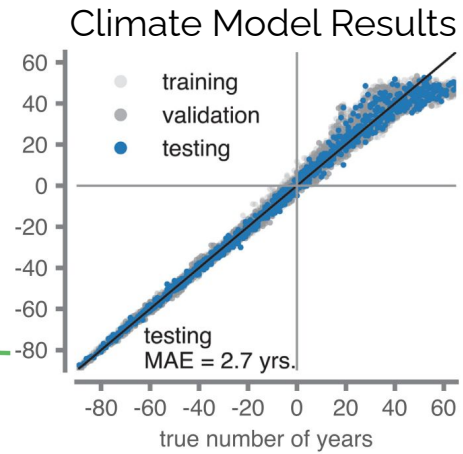
Train neural network to ingest a single annual temperature map and predict the number of years until a warming threshold is reached

Diffenbaugh & Barnes (2023)

Trained on annual maps from 10 realizations from across multiple climate models

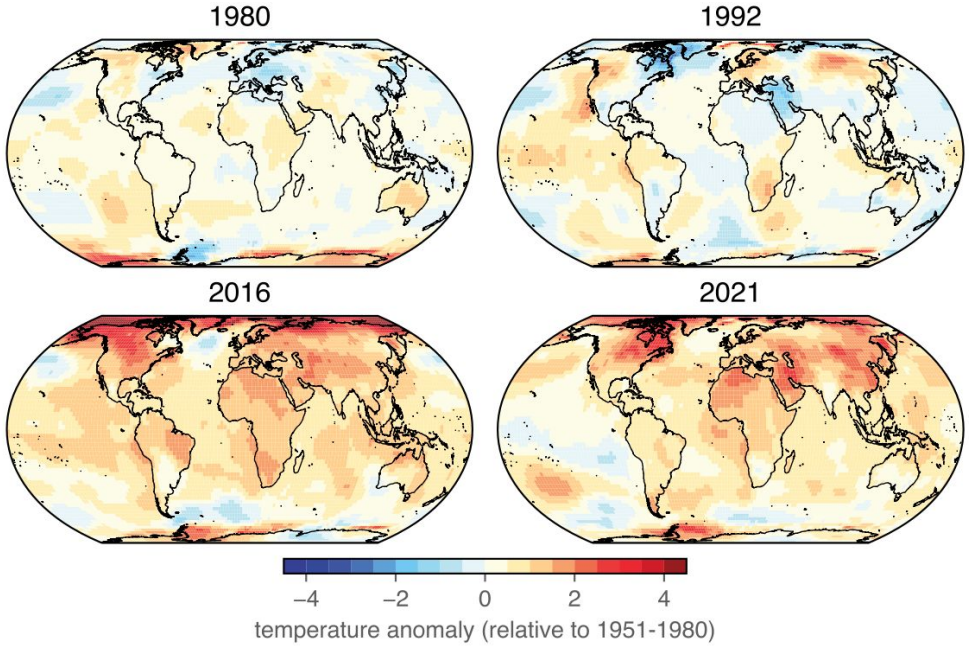


years until warming threshold is reached
+ uncertainties



Train neural network to ingest a single annual temperature map and predict the number of years until a warming threshold is reached

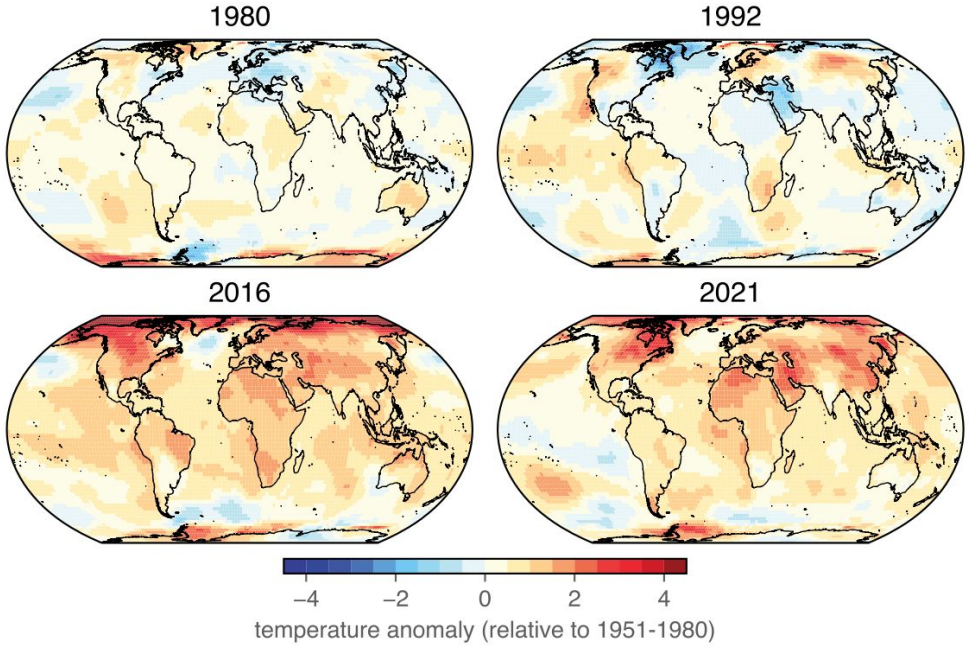
Diffenbaugh & Barnes (2023)

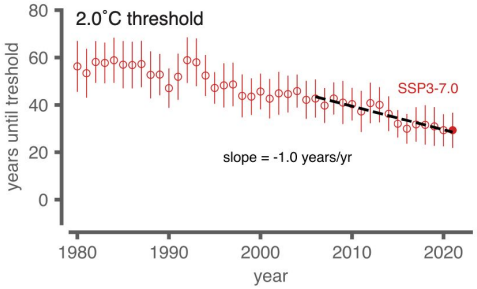


Observations
Berkeley Earth Surface Temperature

Use the trained AI model to predict thresholds based on maps of the observed climate

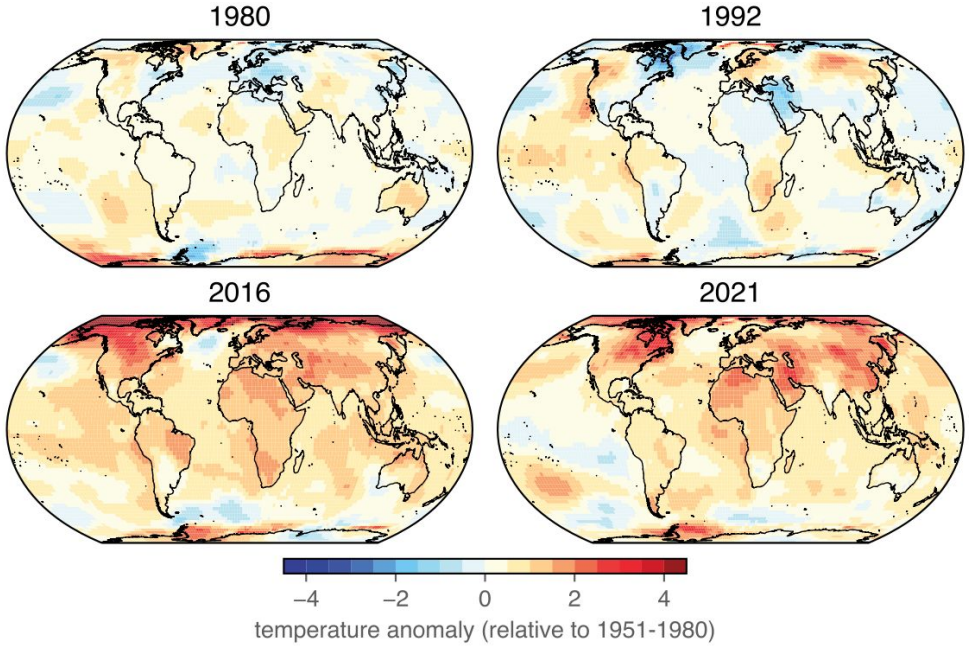
Diffenbaugh & Barnes (2023)

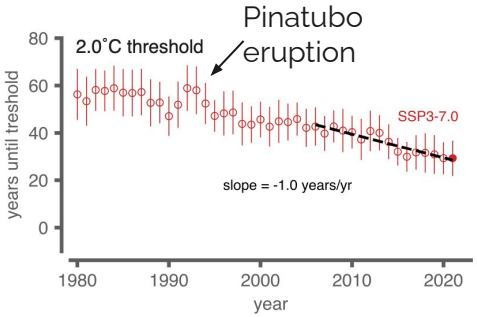




Use the trained AI model to predict thresholds based on maps of the observed climate

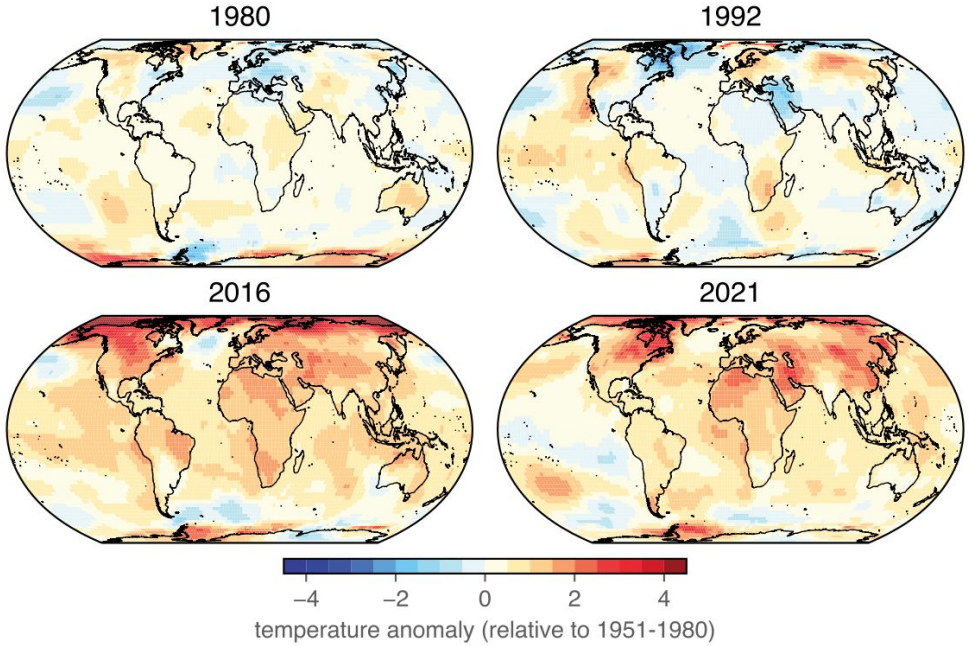
Diffenbaugh & Barnes (2023)

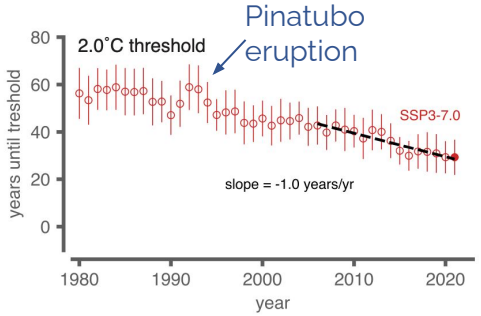


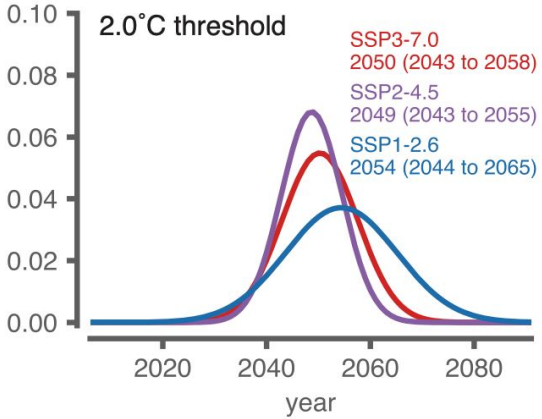


Use the trained AI model to predict thresholds based on maps of the observed climate

Diffenbaugh & Barnes (2023)

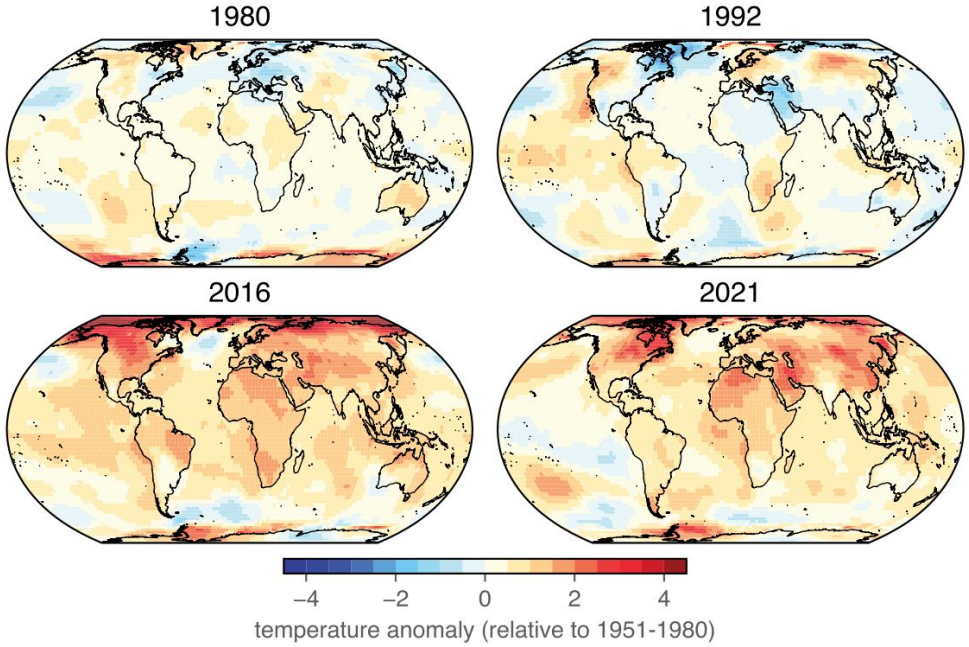


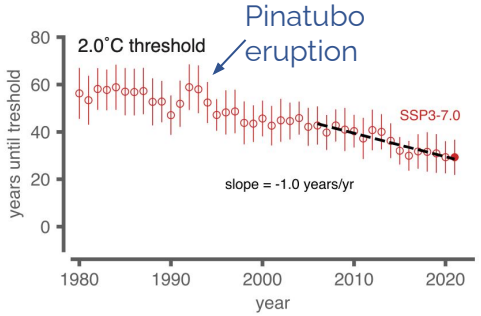


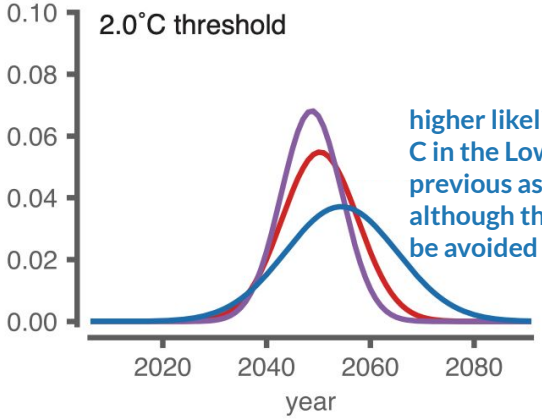


Use the trained AI model to predict thresholds based on maps of the observed climate

Diffenbaugh & Barnes (2023)



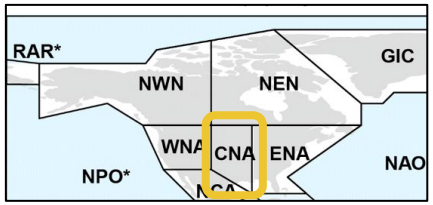




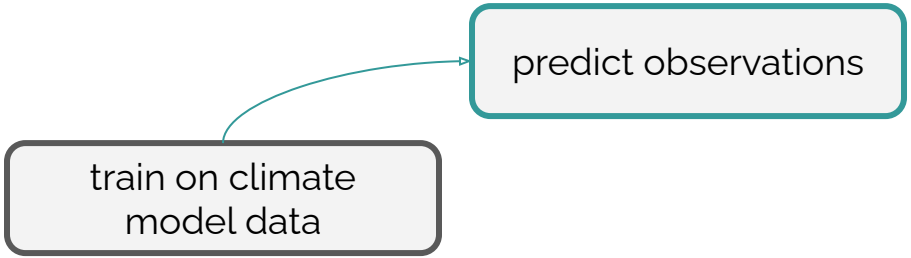
higher likelihood of reaching 2° C in the Low scenario than some previous assessments — although the possibility it could be avoided is not ruled out.

Use the trained AI model to predict thresholds based on maps of the observed climate

Diffenbaugh & Barnes (2023)

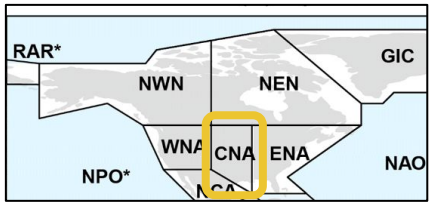


Transfer Learning

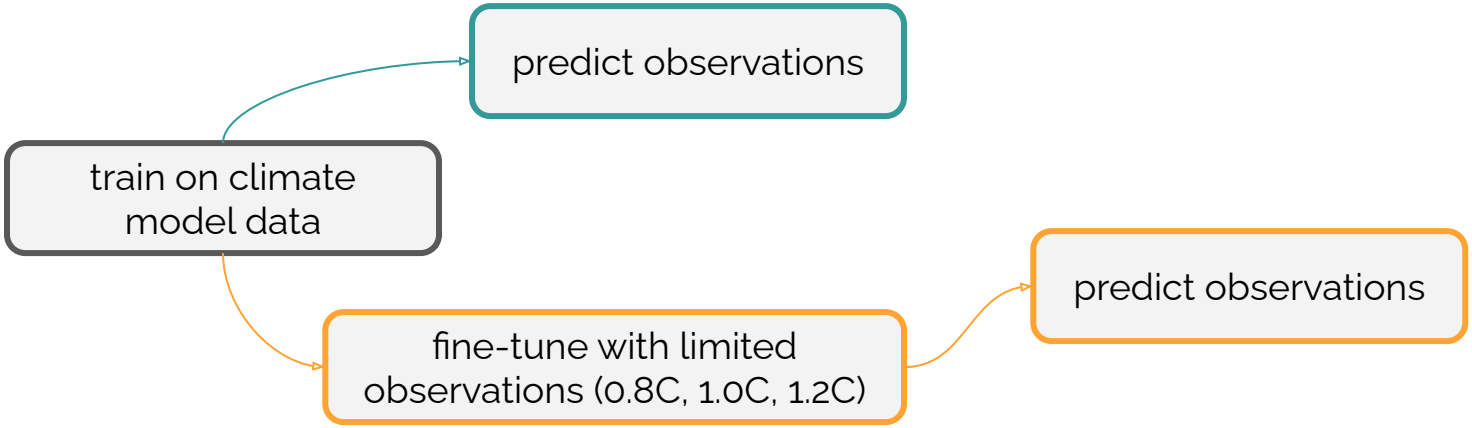


Regional transfer learning provides new insights

Barnes, Diffenbaugh & Seneviratne (2024)

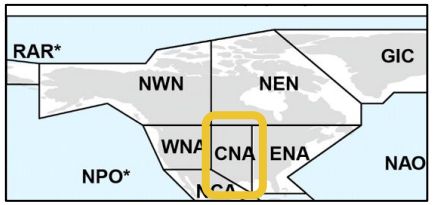


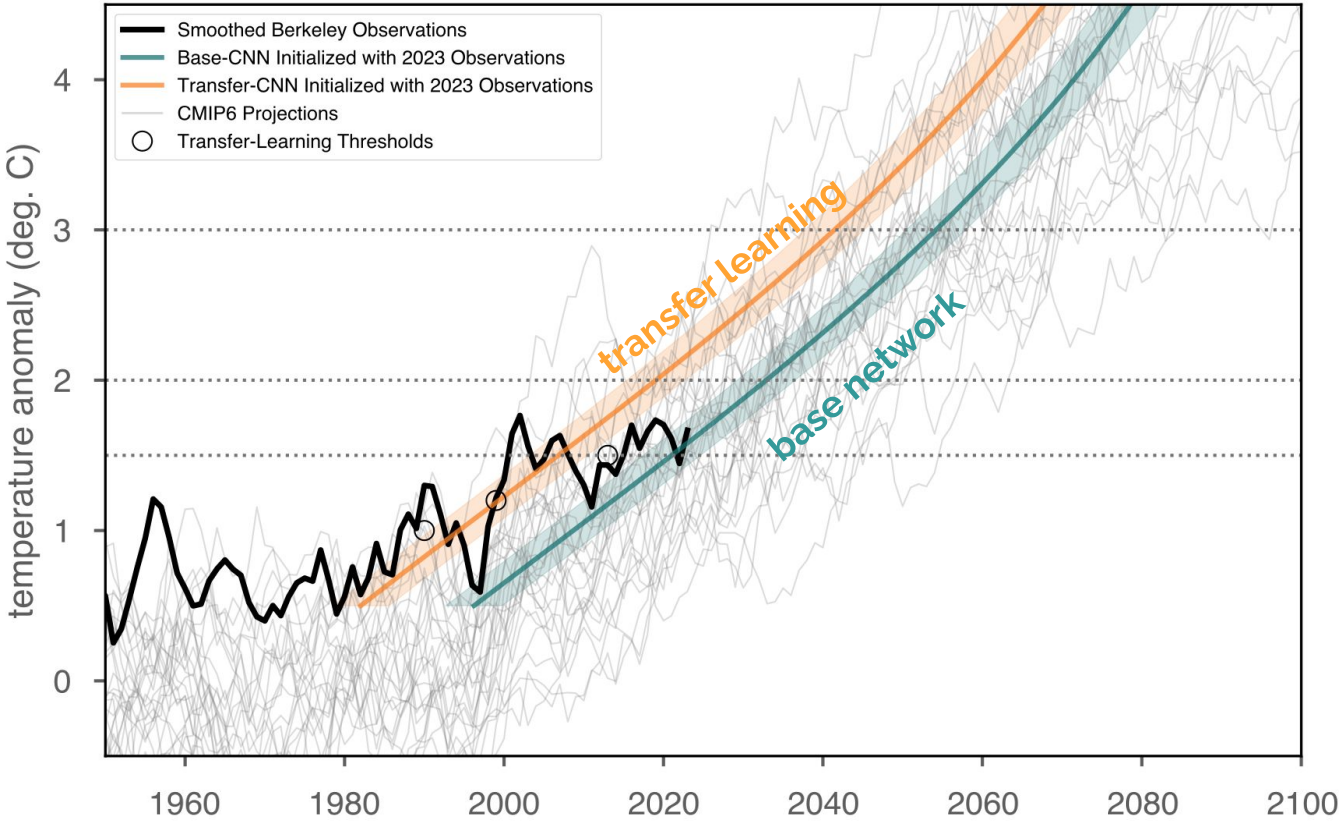
Transfer Learning



Regional transfer learning provides new insights

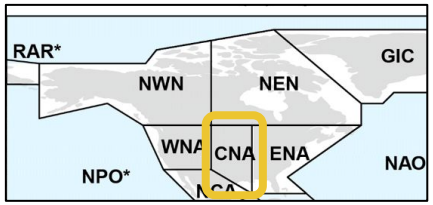
Barnes, Diffenbaugh & Seneviratne (2024)

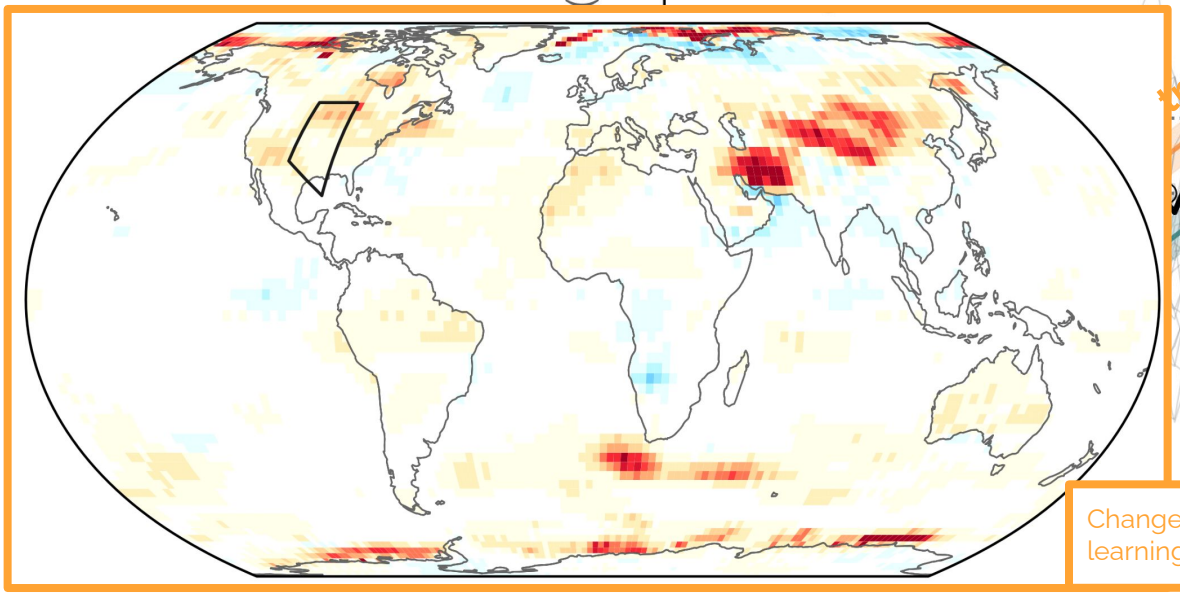




Regional transfer learning provides new insights

Barnes, Diffenbaugh & Seneviratne (2024)



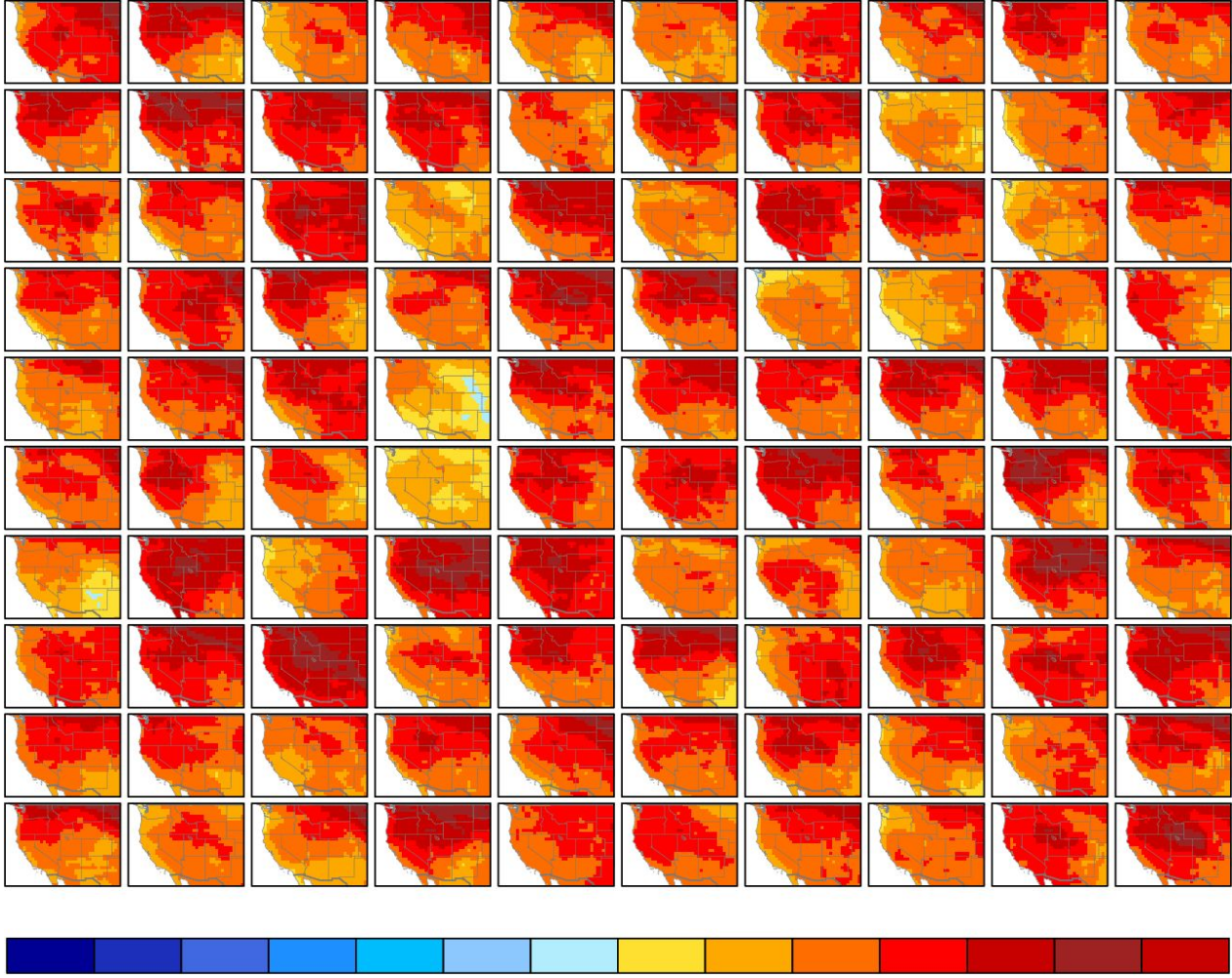


Changes in XAI SHAP values after transfer learning reveal relevant climate model errors

Regional transfer learning provides new insights

Barnes, Diffenbaugh & Seneviratne (2024)

Surface Temperature
Linear Trends
2021-2050

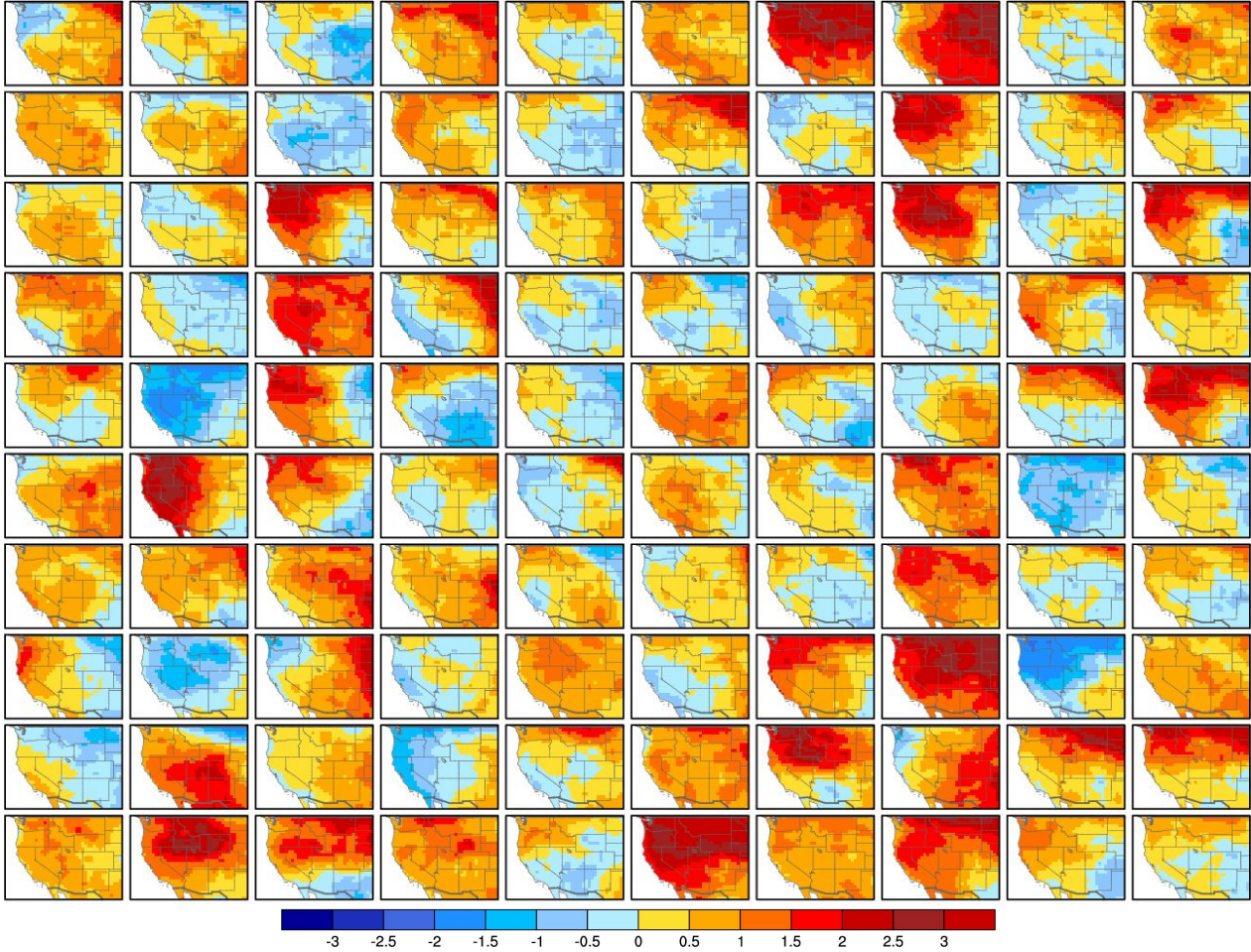


Annual mean surface temperature
As simulated by the CESM2 climate model over 100 different realizations

figure by Dr. Lantao Sun

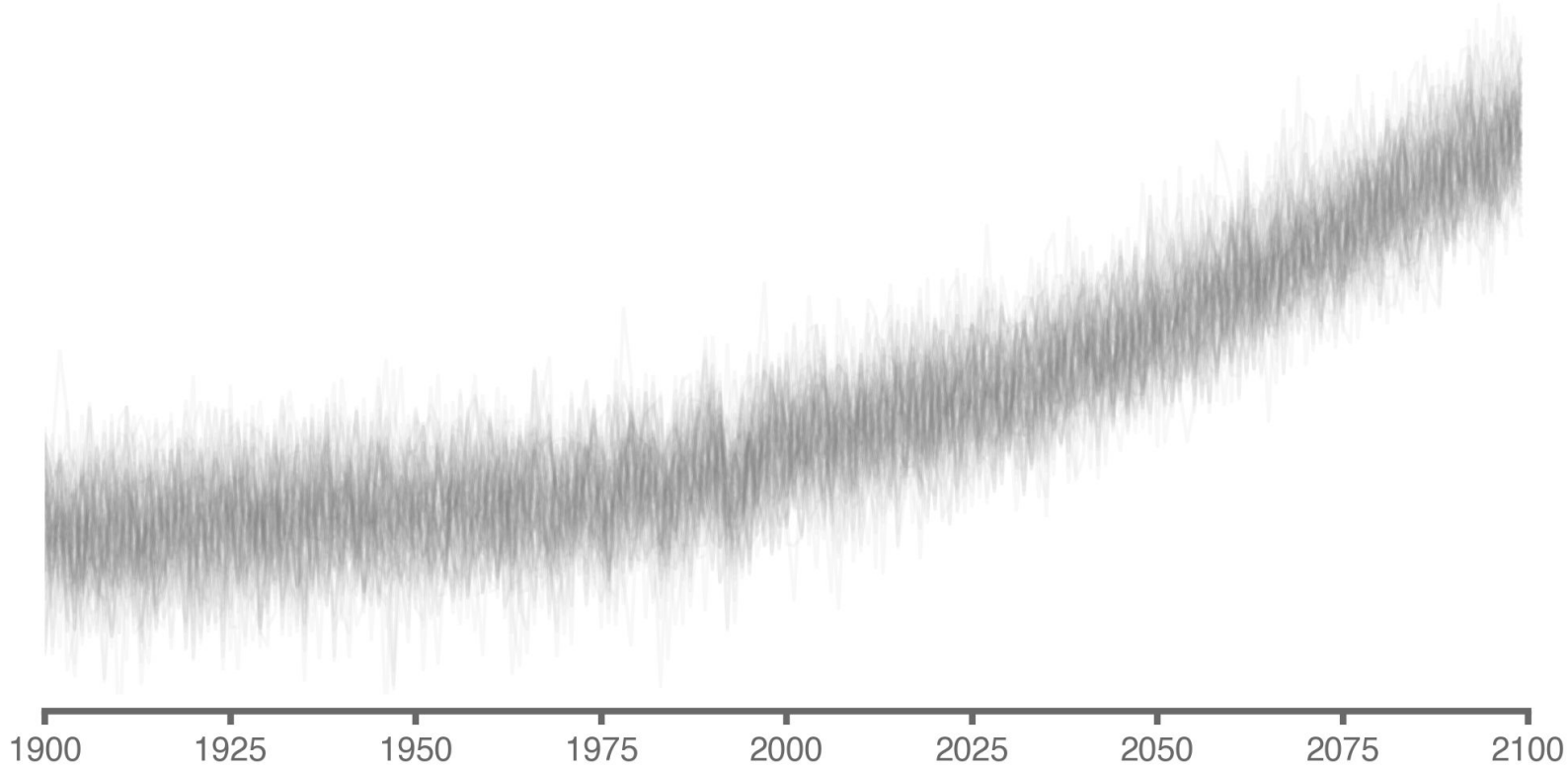
Surface Temperature
Linear Trends
2021-2030

Natural variability is much larger than the climate change signal on regional scales and confounds our knowledge of what the near future holds

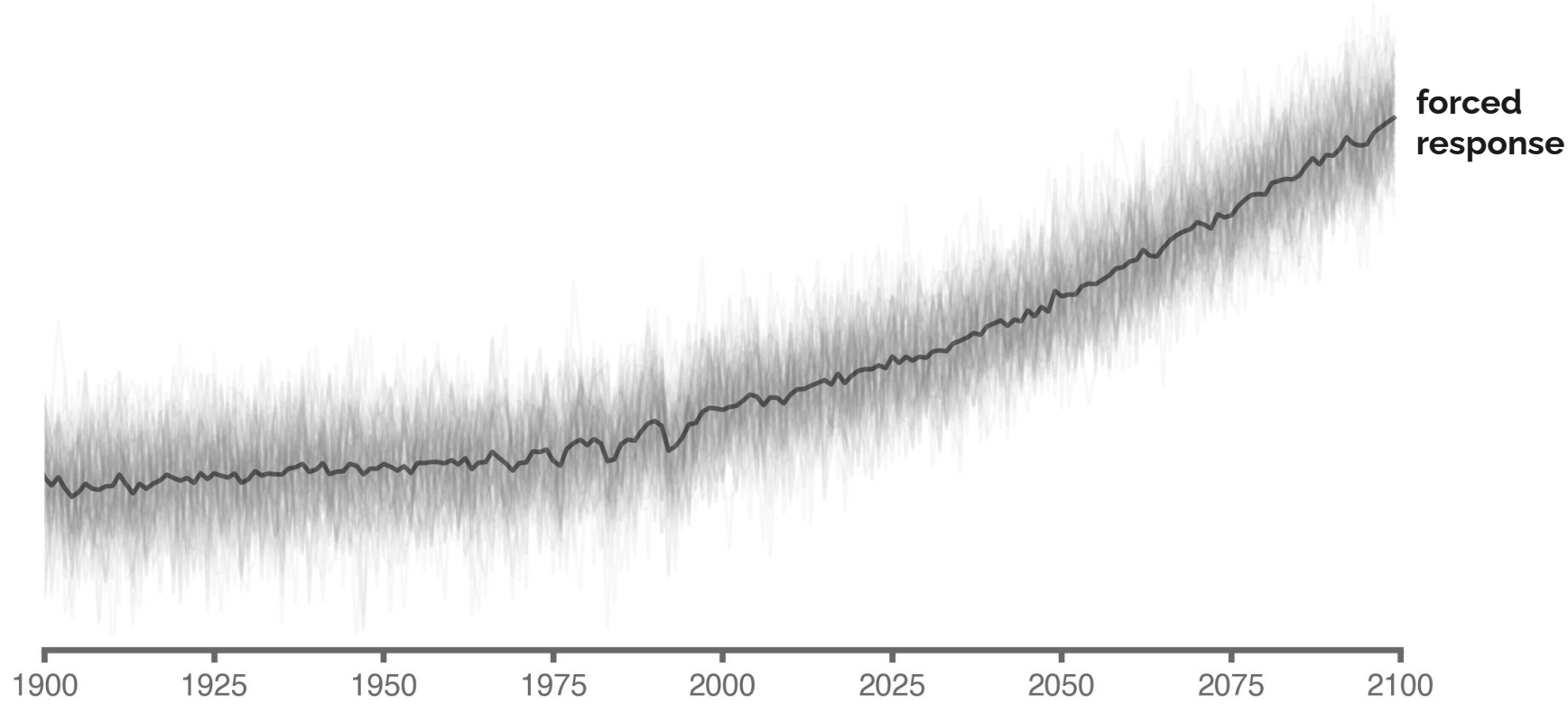


Annual mean surface temperature
As simulated by the CESM2 climate model over 100 different realizations

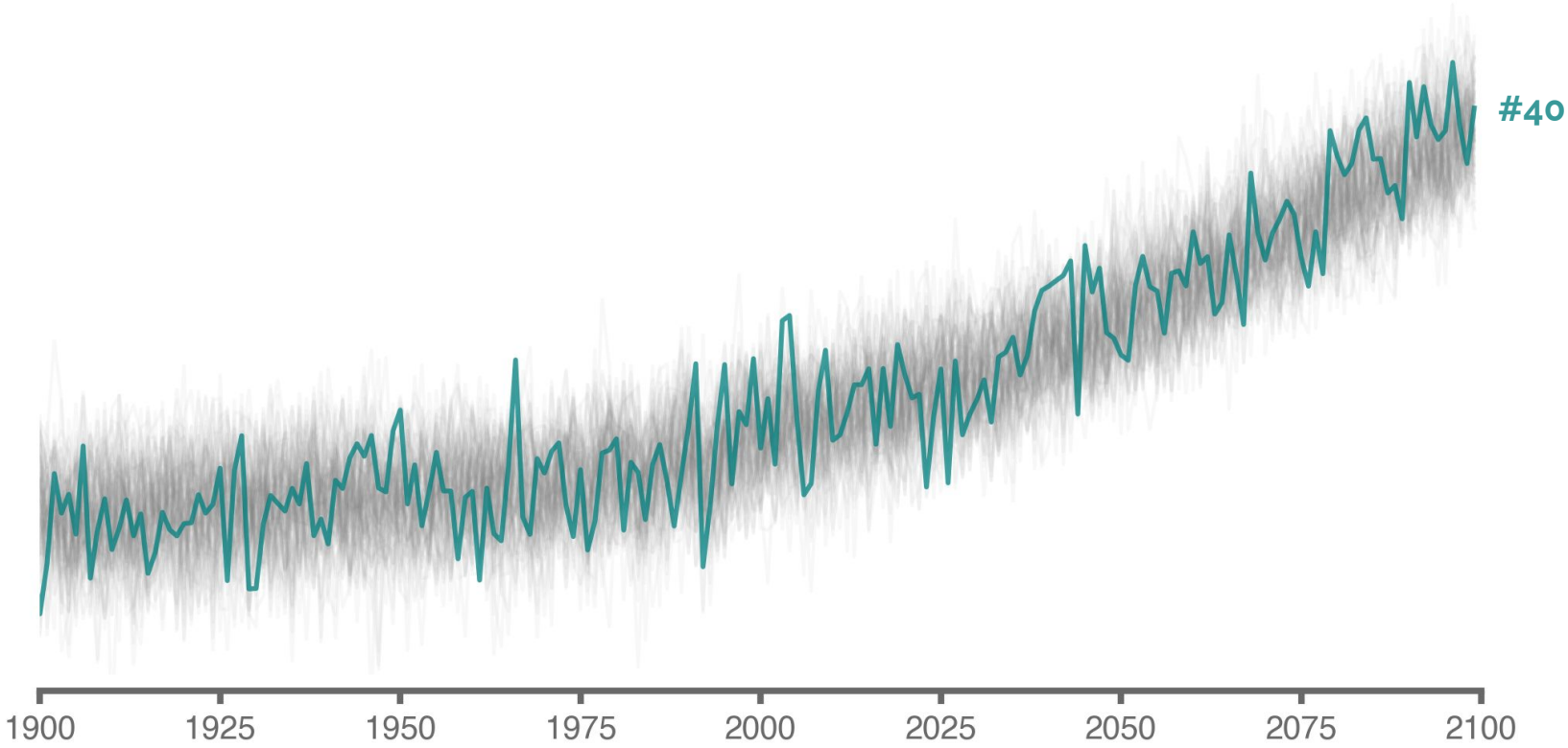
figure by Dr. Lantao Sun



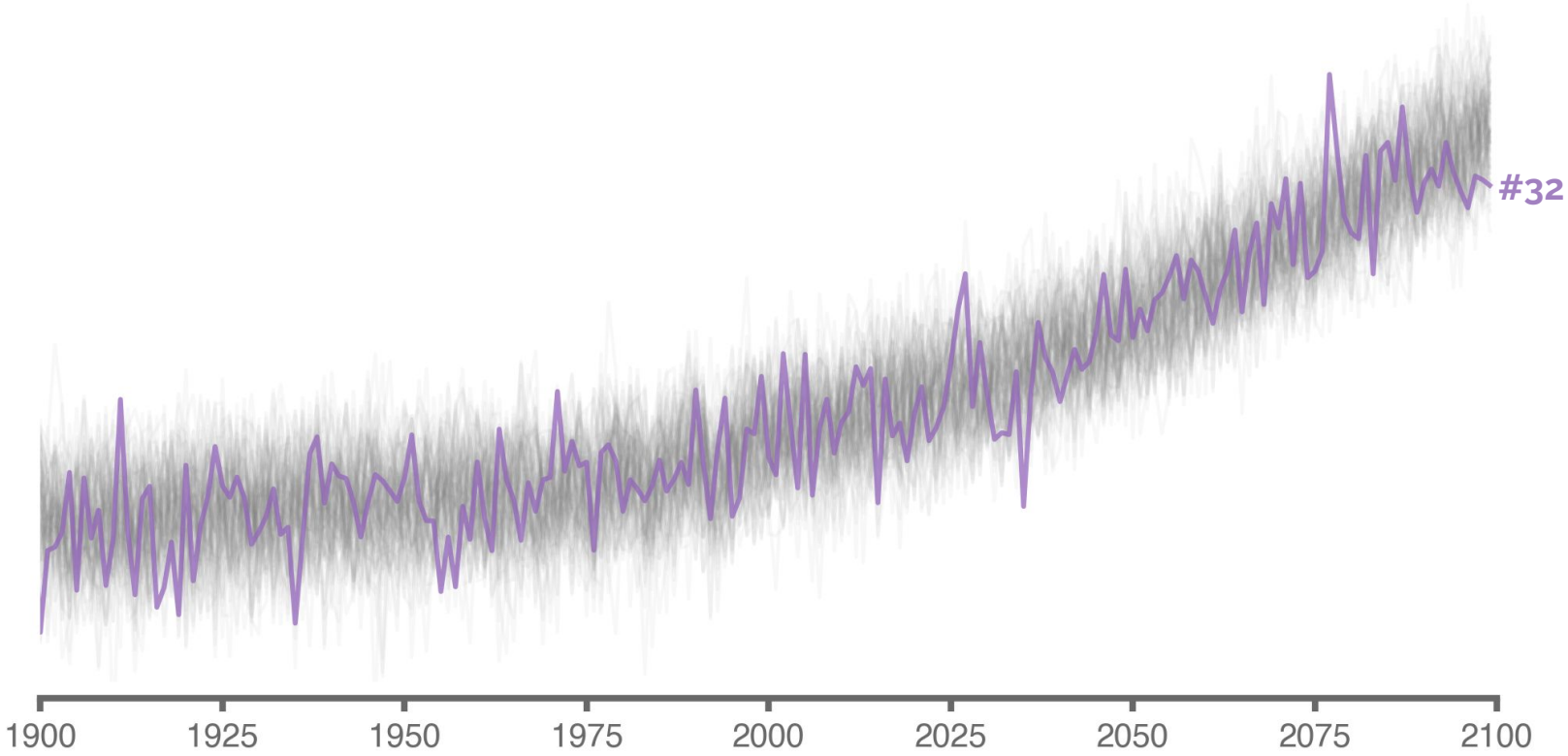
Surface temperature over Chicago, IL
MPI-ESM Large Ensemble; historical + RCP8.5



Surface temperature over Chicago, IL
MPI-ESM Large Ensemble; historical + RCP8.5

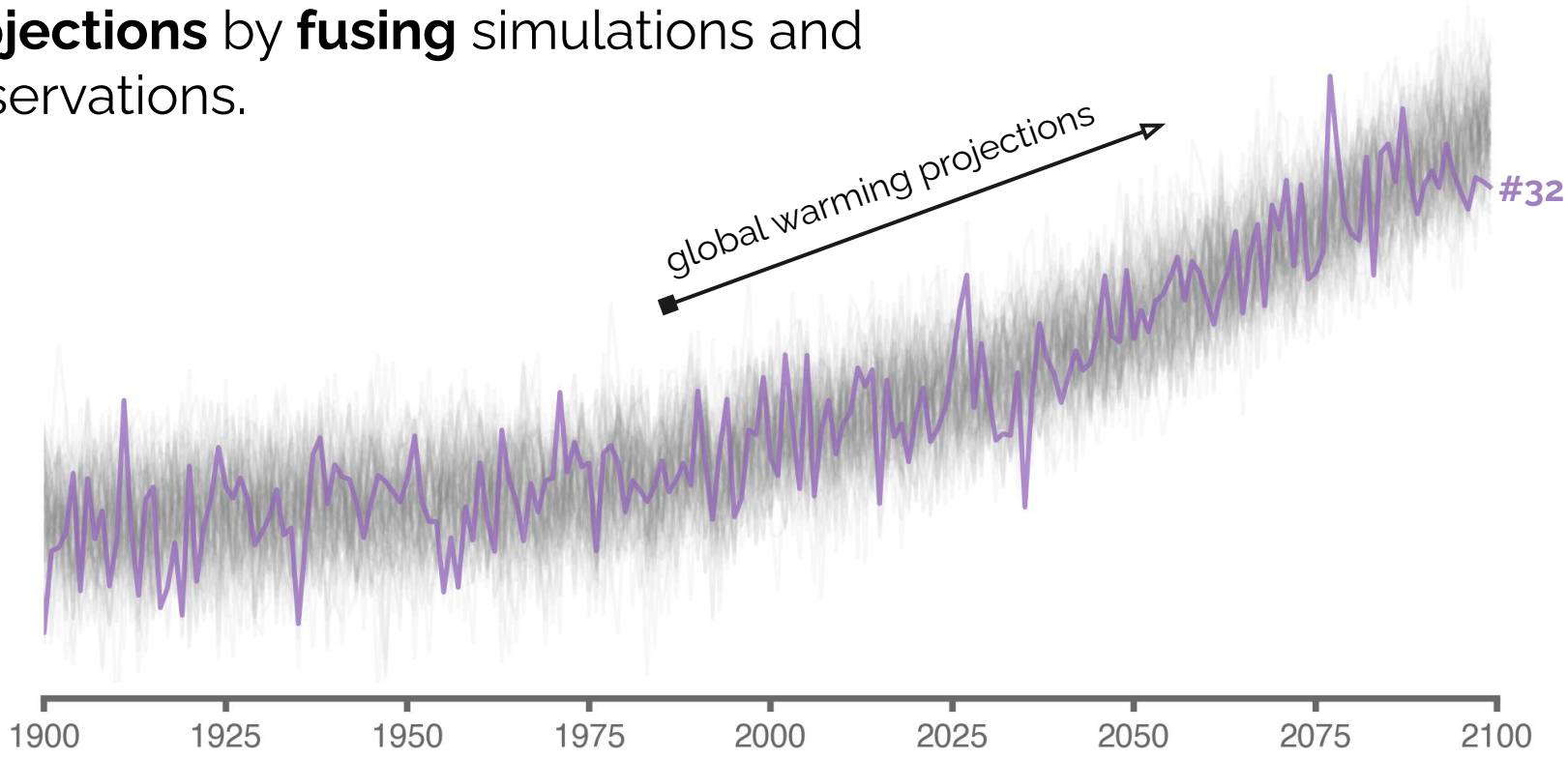


Surface temperature over Chicago, IL
MPI-ESM Large Ensemble; historical + RCP8.5



Surface temperature over Chicago, IL
MPI-ESM Large Ensemble; historical + RCP8.5

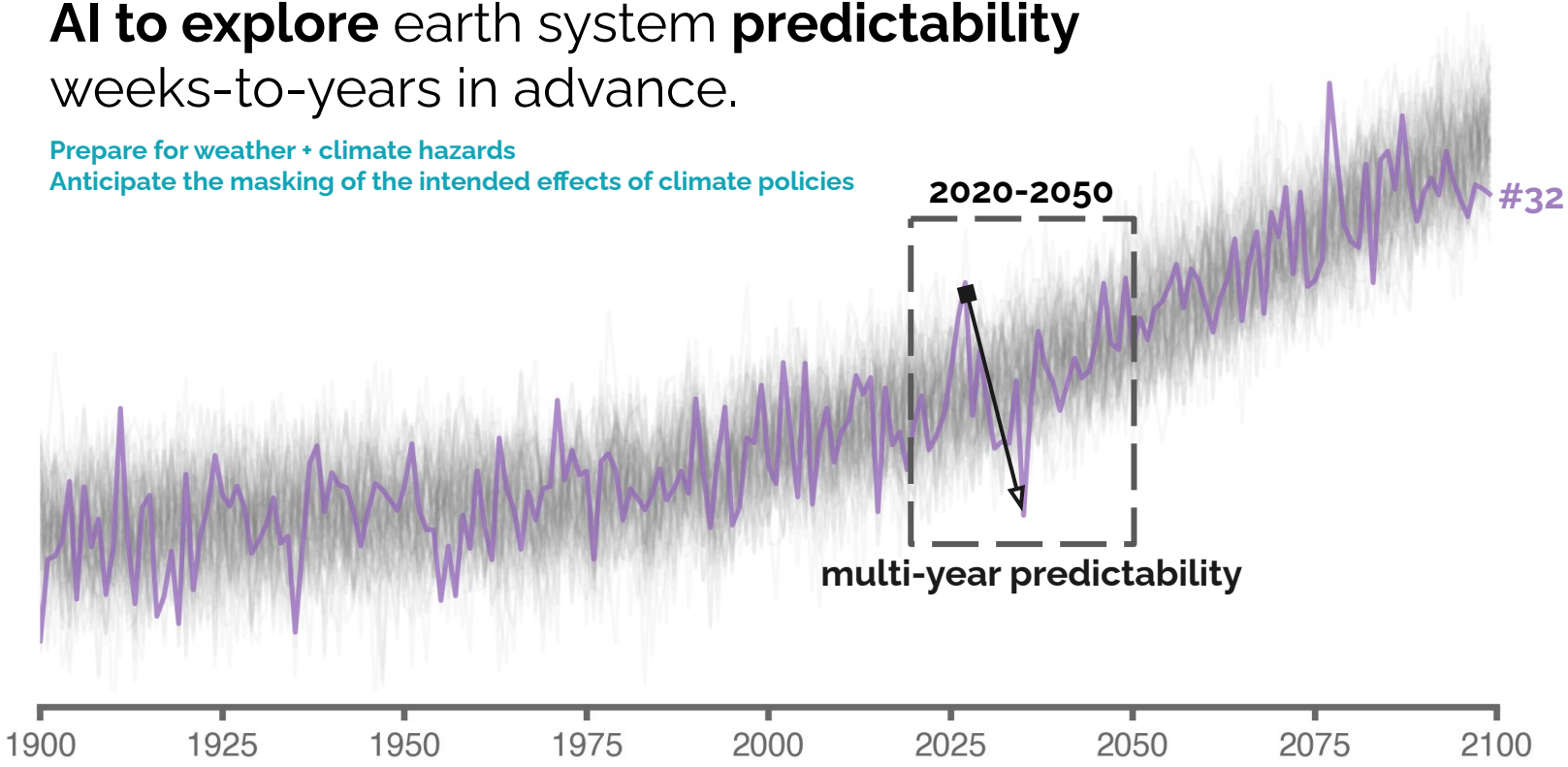
AI to leverage imperfect climate models to better **constrain future projections** by **fusing** simulations and observations.



Surface temperature over Chicago, IL
MPI-ESM Large Ensemble; historical + RCP8.5

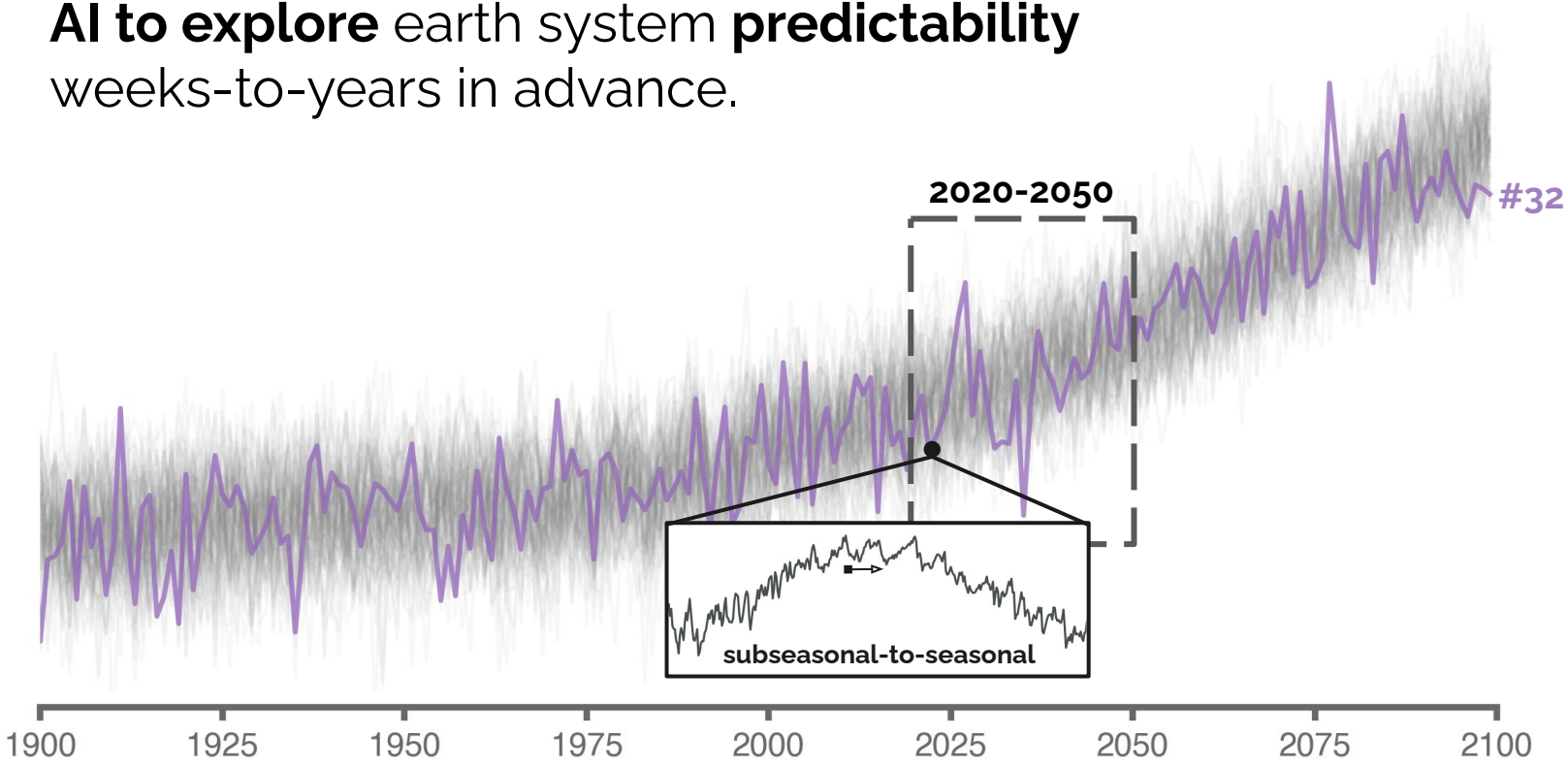
AI to explore earth system **predictability** weeks-to-years in advance.

Prepare for weather + climate hazards
Anticipate the masking of the intended effects of climate policies



Surface temperature over Chicago, IL
MPI-ESM Large Ensemble; historical + RCP8.5

AI to explore earth system **predictability**
weeks-to-years in advance.



Surface temperature over Chicago, IL
MPI-ESM Large Ensemble; historical + RCP8.5

But, climate prediction is incredibly challenging. We cannot expect to make perfect predictions all of the time.

But, climate prediction is incredibly challenging. We cannot expect to make perfect predictions all of the time.

Instead, we must look for specific states of the earth system, i.e. “forecasts of opportunity”, that lead to enhanced predictable behavior.

But, climate prediction is incredibly challenging. We cannot expect to make perfect predictions all of the time.

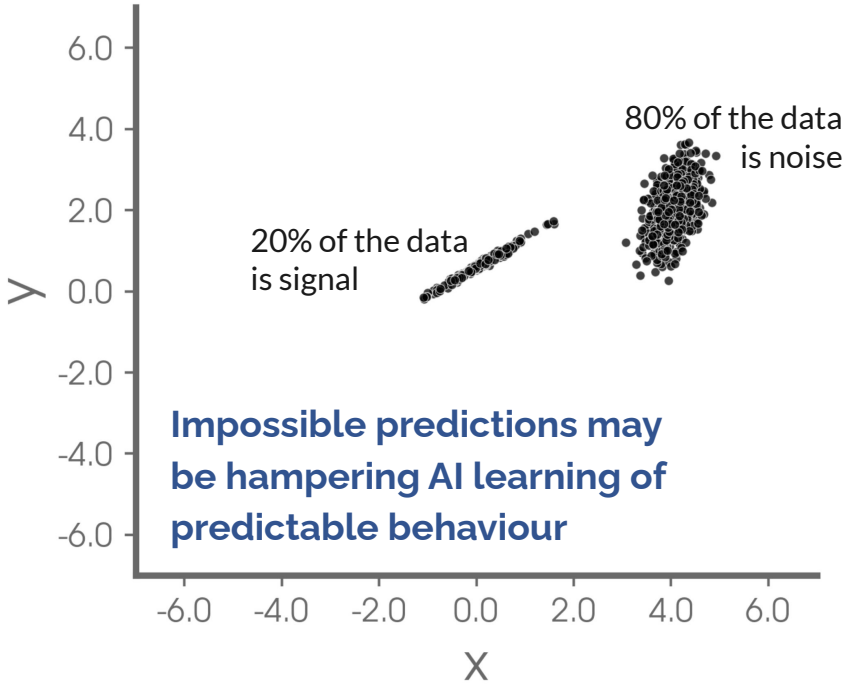
Instead, we must look for specific states of the earth system, i.e. “forecasts of opportunity”, that lead to enhanced predictable behavior.

AI can help us with this.

But, climate prediction is incredibly challenging. We cannot expect to make perfect predictions all of the time.

Instead, we must look for specific states of the earth system, i.e. “forecasts of opportunity”, that lead to enhanced predictable behavior.

AI can help us with this.

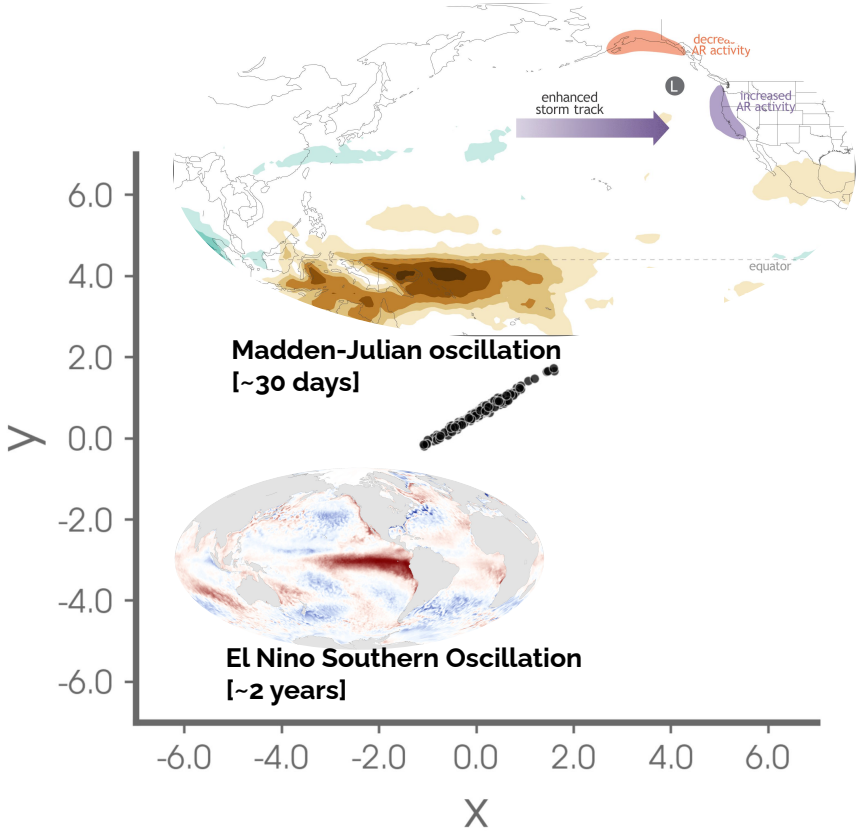


Barnes, Barnes and Gordillo (2021)
Barnes and Barnes (2021a, 2021b)

But, climate prediction is incredibly challenging. We cannot expect to make perfect predictions all of the time.

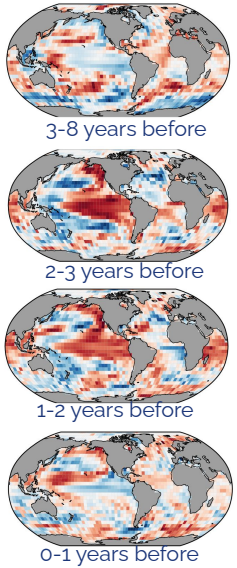
Instead, we must look for specific states of the earth system, i.e. “forecasts of opportunity”, that lead to enhanced predictable behavior.

AI can help us with this.



Barnes, Barnes and Gordillo (2021)
Barnes and Barnes (2021a, 2021b)

past sea-surface temperatures



future sea surface temperatures*
for one grid point
[0-5 years]

warm
neutral
cool

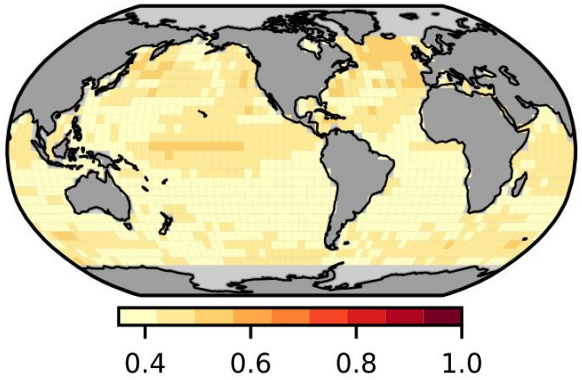
*can predict a range of variables

Predict ocean temperatures 5 years later

Davenport, Barnes & Gordon (2024)

CLIMATE MODEL DATA

Overall Accuracy



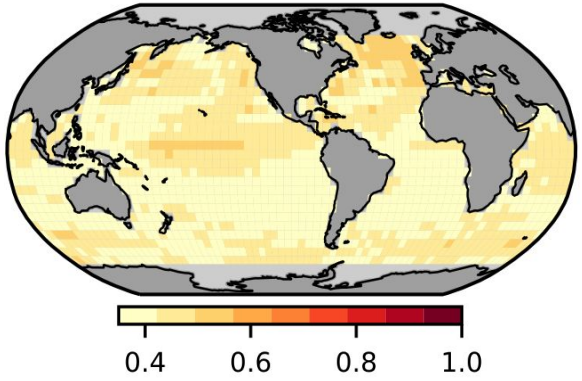
Trained on climate model **MPI-ESM-1-2-LR [3,630 years of data]**
Evaluated on climate model **MPI-ESM-1-2-LR**

Focusing on when the AI is most confident leads to accurate predictions

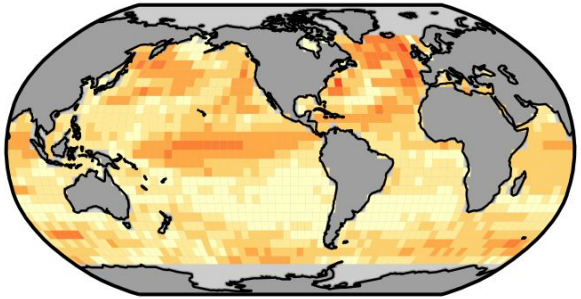
Davenport, Barnes & Gordon (2024)

CLIMATE MODEL DATA

Overall Accuracy



Accuracy for 40% most confident predictions



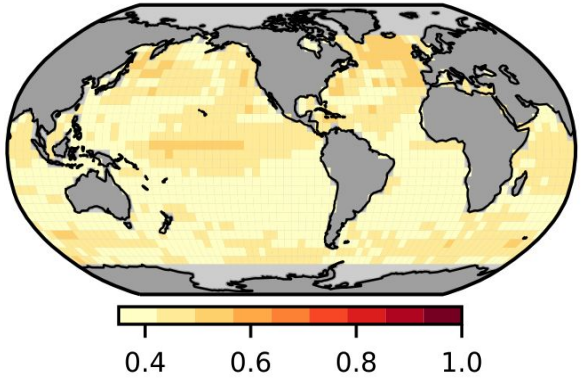
Trained on climate model **MPI-ESM-1-2-LR [3,630 years of data]**
Evaluated on climate model **MPI-ESM-1-2-LR**

Focusing on when the AI is most confident leads to accurate predictions

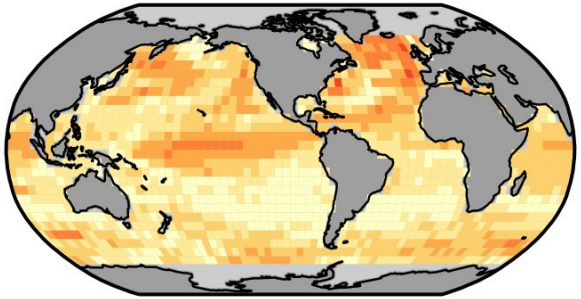
Davenport, Barnes & Gordon (2024)

CLIMATE MODEL DATA

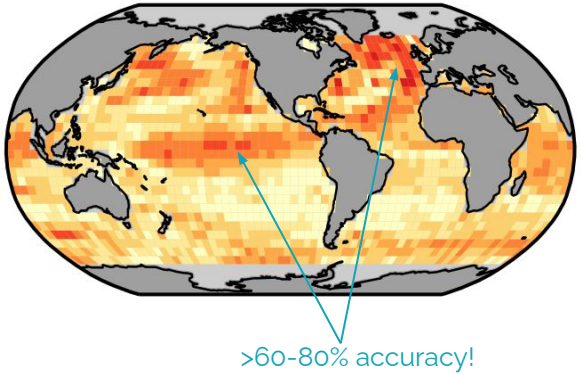
Overall Accuracy



Accuracy for 40% most confident predictions



Accuracy for 20% most confident predictions

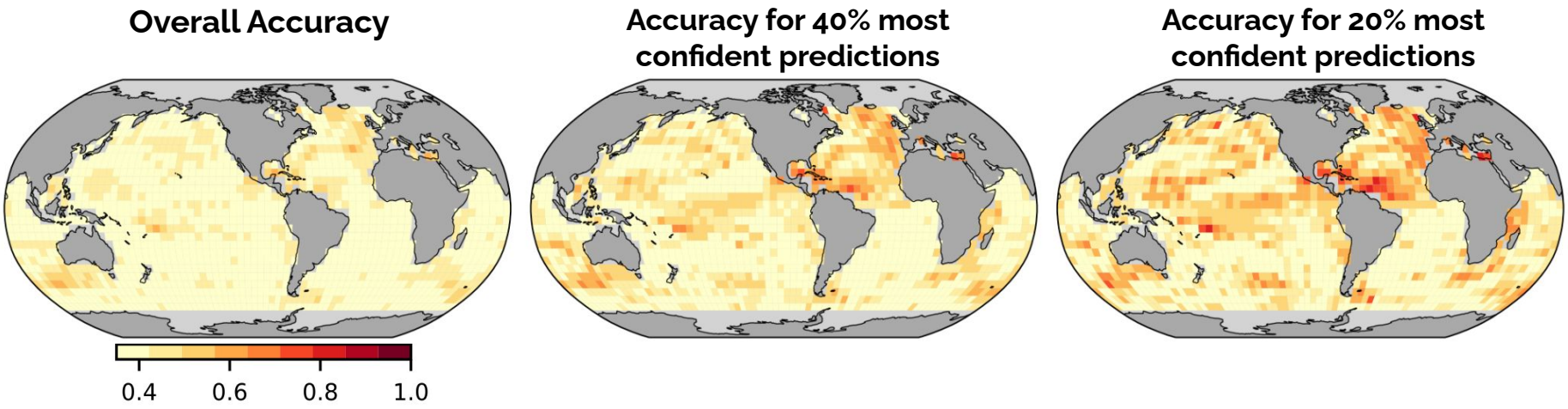


Trained on climate model **MPI-ESM-1-2-LR [3,630 years of data]**
Evaluated on climate model **MPI-ESM-1-2-LR**

Focusing on when the AI is most confident leads to accurate predictions

Davenport, Barnes & Gordon (2024)

OBSERVATIONS



Trained on climate model **MPI-ESM-1-2-LR [3,630 years of data]**
Evaluated on **observations [ERSSTv5; 169 years of data]**

Leveraging climate model data provides accurate predictions of the real world

Davenport, Barnes & Gordon (2024)

But how is the network making its prediction?

…what if we could learn which regions of the globe are most relevant to each prediction?

Opening the Black Box with XAI

In the past few years multiple papers have come out demonstrating the use of AI explainability methods for earth science

MAKING THE BLACK BOX MORE TRANSPARENT

Understanding the Physical Implications of Machine Learning

Amy McGovern, Ryan Lagerquist, David John Gagne II, G. Eli Jergensen, Kimberly L. Elmore, Cameron R. Homeyer, and Travis Smith

Machine learning model interpretation and visualization focusing on meteorological domains are introduced and analyzed.

JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE

Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability

Benjamin A. Toms, Elizabeth A. Barnes, and Imme Ebert-Uphoff

BAMS ISSUES EARLY ONLINE RELEASE COLLECTIONS FOR AUTHORS

RESEARCH ARTICLE | 31 AUGUST 2020

Evaluation, Tuning and Interpretation of Neural Networks for Working with Images in Meteorological Applications

Imme Ebert-Uphoff; Kyle Hilburn

Bull. Amer. Meteor. Soc. 1-49.

https://doi.org/10.1175/BAMS-D-20-0097.1

Capsule:

This article discusses strategies for the development of neural networks (aka *deep learning*) for meteorological applications. Topics include evaluation, tuning and interpretation of neural networks for working with meteorological images.

Abstract

The method of neural networks (aka deep learning) has opened up many new opportunities to utilize remotely sensed images in meteorology. Common applications include image classification, e.g., to determine whether an image contains a tropical cyclone, and image-to-image translation, e.g., to emulate radar imagery for satellites that only have passive channels. However, there are yet many open questions regarding the use of neural networks for working with meteorological images, such as best practices for evaluation, tuning and interpretation. This article highlights several strategies and practical considerations for neural network development that have not yet received much attention in the meteorological community, such as the concept of receptive fields, underutilized meteorological performance measures, and methods for neural network interpretation, such as synthetic experiments and layer-wise relevance propagation. We also consider the process of neural network interpretation as a whole, recognizing it as an iterative meteorologist-driven discovery process that builds on experimental design and hypothesis generation and testing. Finally, while most work on neural network interpretation in meteorology has so far focused on networks for image classification tasks, we expand the focus to also include networks for image-to-image translation.

Environmental Data Science (2022), 1: e8, 1–17

doi:10.1017/eds.2022.7

CAMBRIDGE UNIVERSITY PRESS

METHODS PAPER

Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset

Antonios Mamalakis, Imme Ebert-Uphoff and Elizabeth A. Barnes

Received: 29 November 2021; Revised: 05 April 2022; Accepted: 28 April 2022

Keywords: attribution benchmark; eXplainable Artificial Intelligence; geosciences; ground truth; neural networks; regression problems

Abstract

Despite the increasingly successful application of neural networks to many problems in the geosciences, their complex and nonlinear structure makes the interpretation of their predictions difficult, which limits model trust and does not allow scientists to gain physical insights about the problem at hand. Many different methods have been introduced in the emerging field of eXplainable Artificial Intelligence (XAI), which aims at attributing the network's prediction to specific features in the input domain. XAI methods are usually assessed by using benchmark datasets (such as MNIST or ImageNet for image classification). However, an objective, theoretically derived ground truth for the attribution is lacking for most of these datasets, making the assessment of XAI in many cases subjective. Also, benchmark datasets specifically designed for problems in geosciences are rare. Here, we provide a framework, based on the use of additively separable functions, to generate attribution benchmark datasets for regression problems for which the ground truth of the attribution is known a priori. We generate a large benchmark dataset and train a fully connected network to learn the underlying function that was used for simulation. We then compare estimated heatmaps from different XAI methods to the ground truth in order to identify examples where specific XAI methods perform well or poorly. We believe that attribution benchmarks as the ones introduced herein are of great importance for further application of neural networks in the geosciences, and for more objective assessment and accurate implementation of XAI methods, which will increase model trust and assist in discovering new science.

JULY 2024 BOMMER ET AL.

Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science

Philine Lou Bommer, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, Marina M.-C. Höhne

(Manuscript received 25 August 2023, in final form 8 March 2024, accepted 19 March 2024)

ABSTRACT: Explainable artificial intelligence (XAI) methods shed light on the predictions of machine learning algorithms. Several different approaches exist and have already been applied in climate science. However, usually missing ground truth explanations complicate their evaluation and comparison, subsequently impeding the choice of the XAI method. Therefore, in this work, we introduce XAI evaluation in the climate context and discuss different desired explanation properties, namely, robustness, faithfulness, randomization, complexity, and localization. To this end, we chose previous work as a case study where the decade of annual-mean temperature maps is predicted. After training both a multilayer perceptron (MLP) and a convolutional neural network (CNN), multiple XAI methods are applied and their skill scores in reference to a random uniform explanation are calculated for each property. Independent of the network, we find that XAI methods such as Integrated Gradients, layerwise relevance propagation, and input times gradients exhibit considerable robustness, faithfulness, and complexity while sacrificing randomization performance. Sensitivity methods, gradient, SmoothGrad, NoiseGrad, and FusionGrad, match the robustness skill but sacrifice faithfulness and complexity for the randomization skill. We find architecture-dependent performance differences regarding robustness, complexity, and localization skills of different XAI methods, highlighting the necessity for research task-specific evaluation. Overall, our work offers an overview of different evaluation properties in the climate science context and shows how to compare and benchmark different explanation methods, assessing their suitability based on strengths and weaknesses, for the specific research problem at hand. By that, we aim to support climate researchers in the selection of a suitable XAI method.

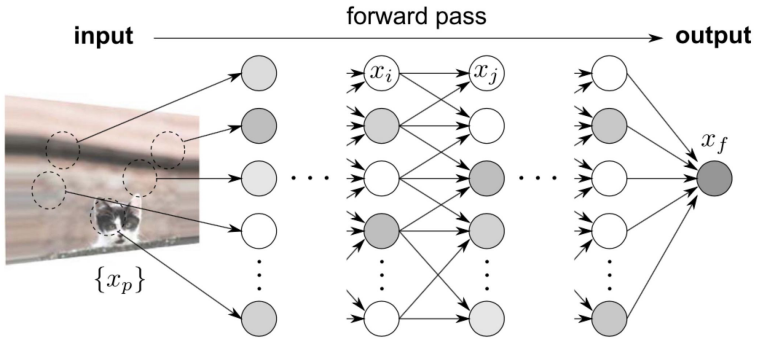
SIGNIFICANCE STATEMENT: Explainable artificial intelligence (XAI) helps to understand the reasoning behind the prediction of a neural network. XAI methods have been applied in climate science to validate networks and provide new insight into physical processes. However, the increasing number of XAI methods can overwhelm practitioners, making it difficult to choose an explanation method. Since XAI methods' results can vary, uninformed choices might cause misleading conclusions about the network decision. In this work, we introduce XAI evaluation to compare and assess the performance of explanation methods based on five desirable properties. We demonstrate that XAI evaluation reveals the strengths and weaknesses of different XAI methods. Thus, our work provides climate researchers with the tools to compare, analyze, and subsequently choose explanation methods.

XAI Attribution Methods

Attribution heatmaps are largely consistent with how many climate scientists pose questions

Prediction

of 1 sample



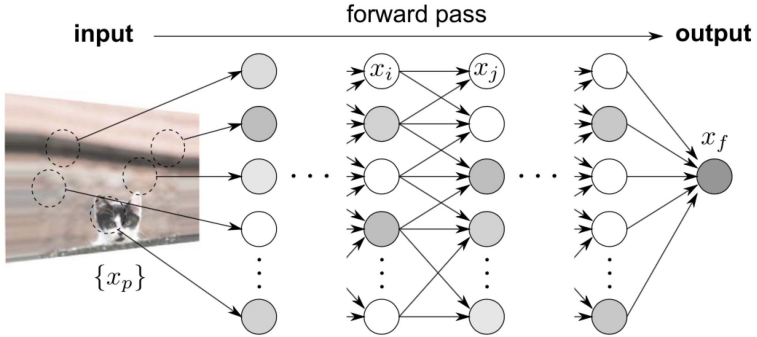
Pr(cat)=.8

e.g. Montavon et al. (2017), Pattern Recognition; Montavon et al. (2018), Digital Signal Processing

XAI Attribution Methods

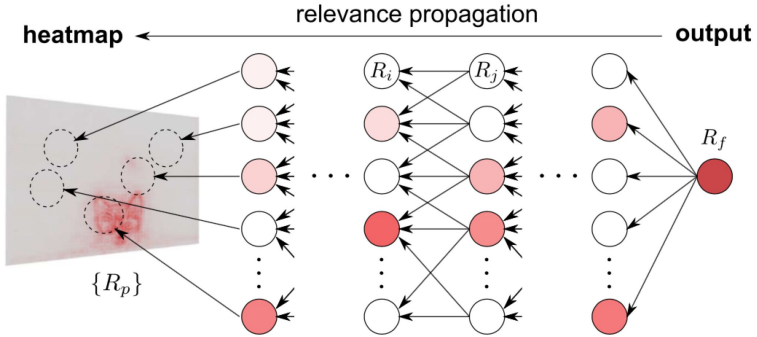
Attribution heatmaps are largely consistent with how many climate scientists pose questions

Prediction
of 1 sample



Pr(cat)=.8

Attribution
of 1 sample



Pr(cat)=.8

e.g. Montavon et al. (2017), Pattern Recognition; Montavon et al. (2018), Digital Signal Processing

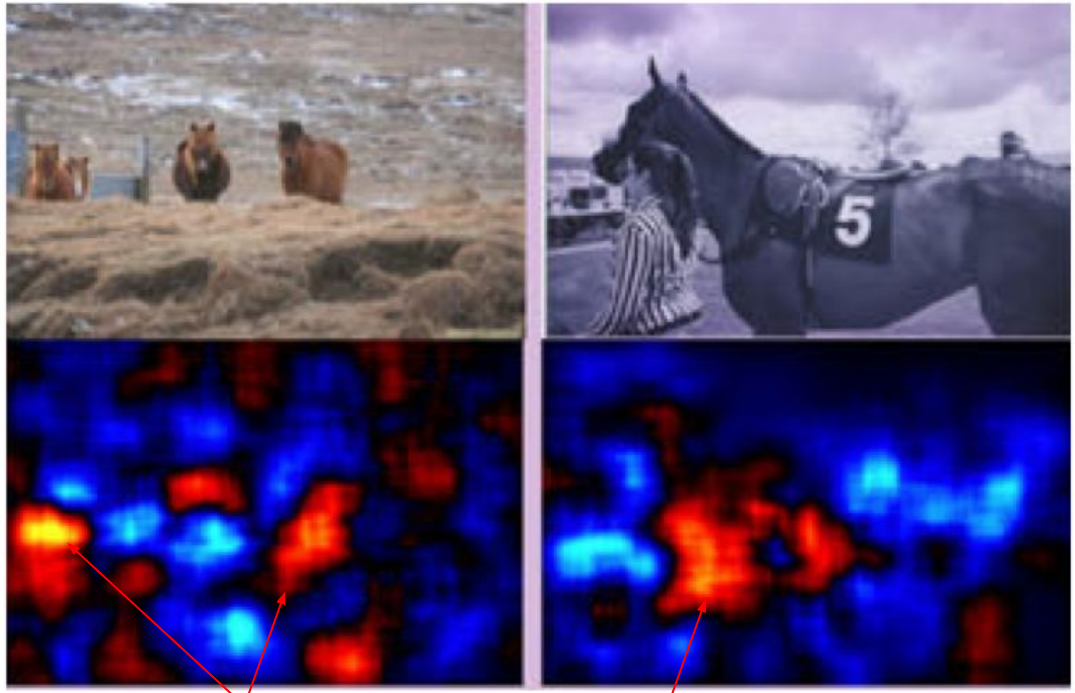
XAI to identify problematic strategies

Is the AI approach predicting the right answers for the right reasons?

Lapuschkin et al. “Unmasking Clever Hans Predictors and Assessing What Machines Really Learn.” Nature Communications, vol. 10, no. 1, Mar. 2019, p. 1096, doi:10.1038/s41467-019-08987-4.

XAI to identify problematic strategies

Is the AI approach predicting the right answers for the right reasons?



red shading: relevant regions for making the network think there is a horse

Lapuschkin et al. "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn." Nature Communications, vol. 10, no. 1, Mar. 2019, p. 1096, doi:10.1038/s41467-019-08987-4.

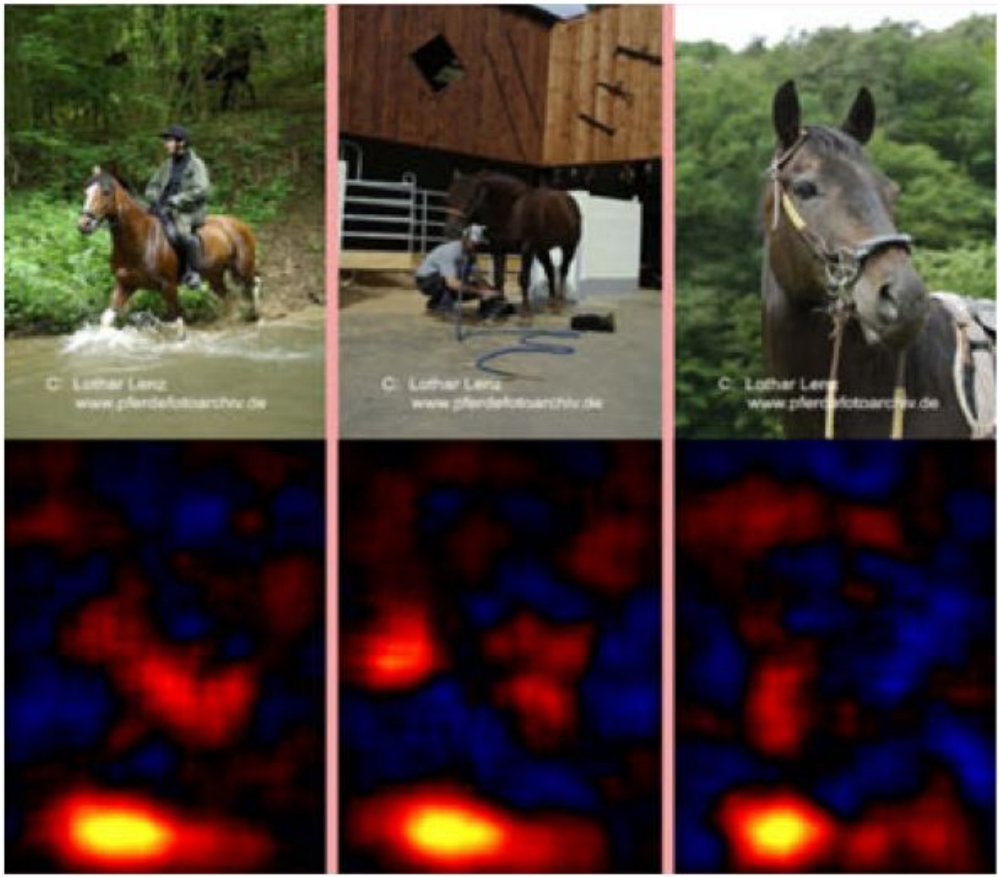
XAI to identify problematic strategies

Is the AI approach predicting the right answers for the right reasons?

Lapuschkin et al. “Unmasking Clever Hans Predictors and Assessing What Machines Really Learn.” Nature Communications, vol. 10, no. 1, Mar. 2019, p. 1096, doi:10.1038/s41467-019-08987-4.

XAI to identify problematic strategies

Is the AI approach predicting the right answers for the right reasons?

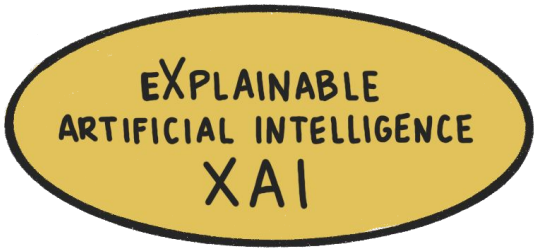


red shading: relevant regions for making the network think there is a horse

Lapuschkin et al. "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn." Nature Communications, vol. 10, no. 1, Mar. 2019, p. 1096, doi:10.1038/s41467-019-08987-4.

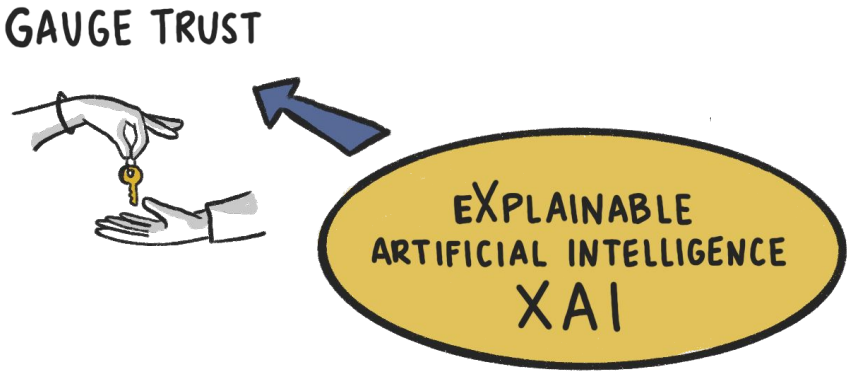
Reasons to care about XAI

A scientist's ultimate goal is typically to understand "why?", but even if you don't care "why?" you should still care about XAI.



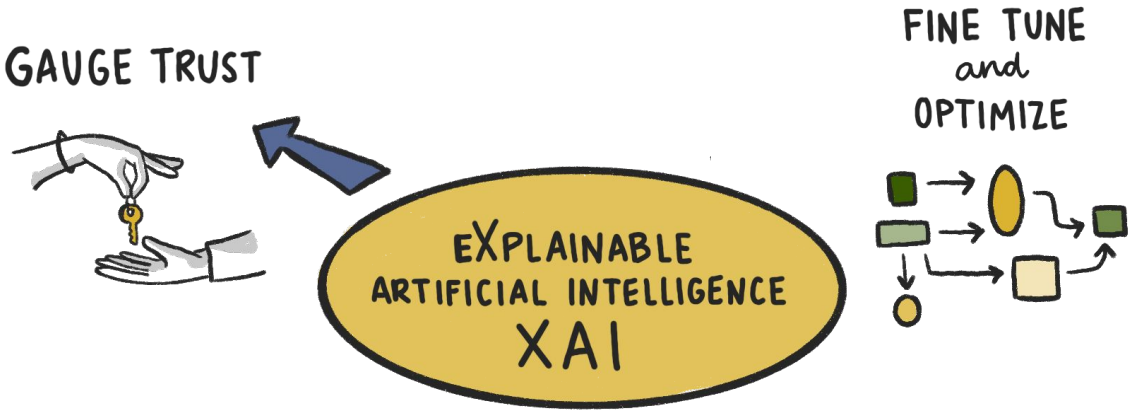
Reasons to care about XAI

A scientist's ultimate goal is typically to understand "why?", but even if you don't care "why?" you should still care about XAI.



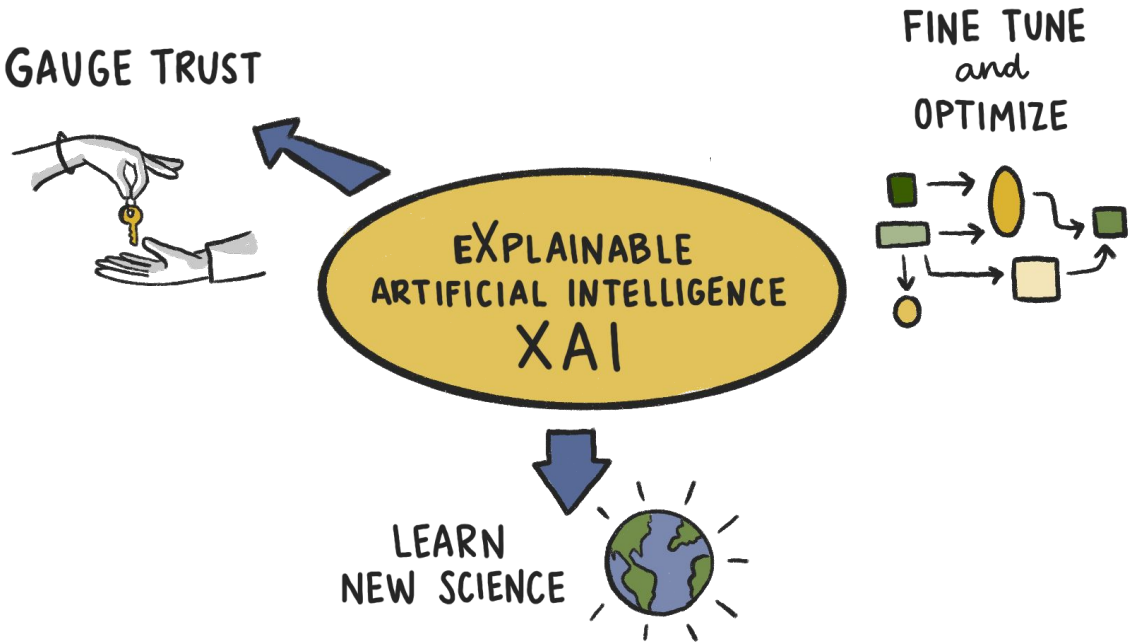
Reasons to care about XAI

A scientist's ultimate goal is typically to understand "why?", but even if you don't care "why?" you should still care about XAI.



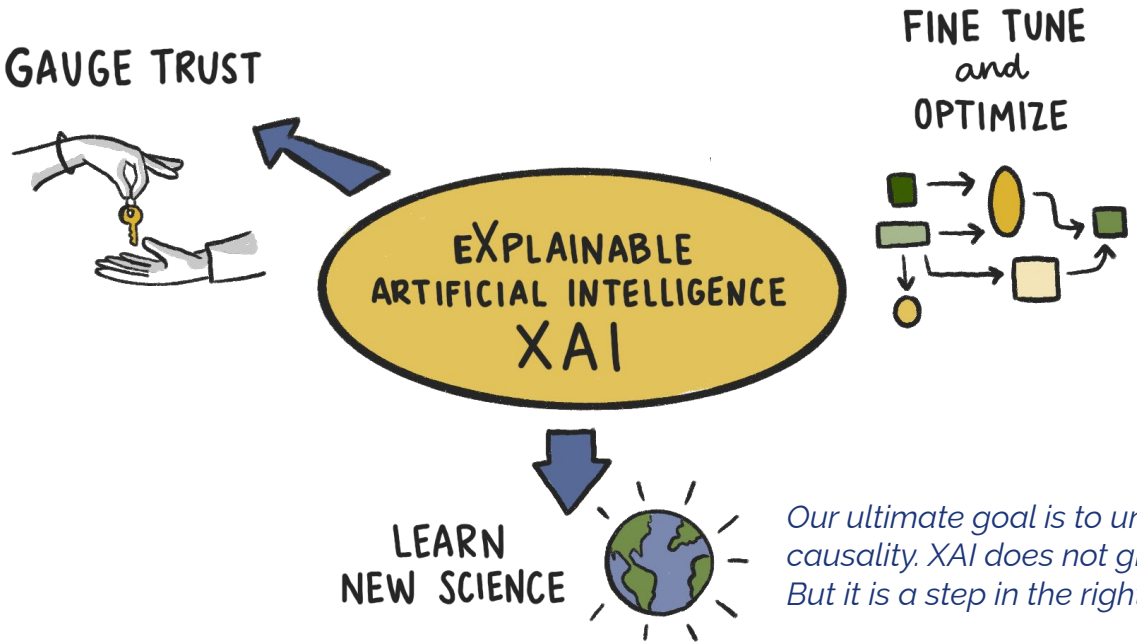
Reasons to care about XAI

A scientist's ultimate goal is typically to understand "why?", but even if you don't care "why?" you should still care about XAI.



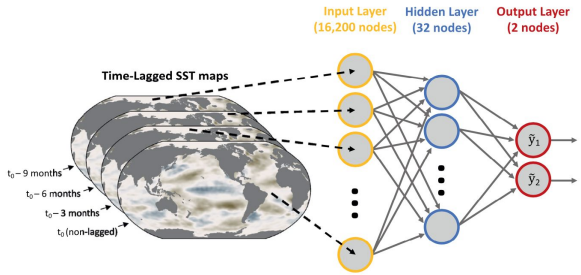
Reasons to care about XAI

A scientist's ultimate goal is typically to understand "why?", but even if you don't care "why?" you should still care about XAI.

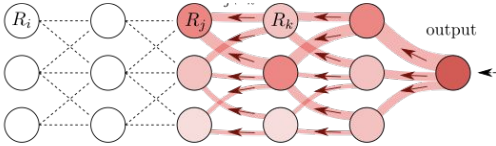


Our ultimate goal is to understand causality. XAI does not give us this. But it is a step in the right direction.

Predicting 5-year average surface temperature at each grid point
Applied to 1200 years of CESM2 control simulation



Prediction of average land surface temperature anomaly over the next 5 years

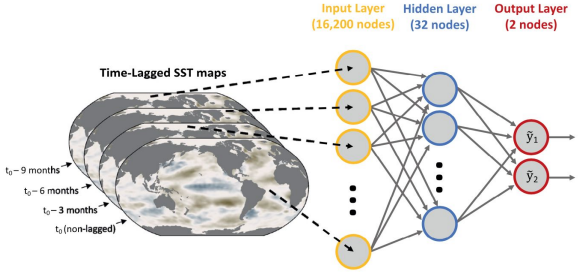


XAI

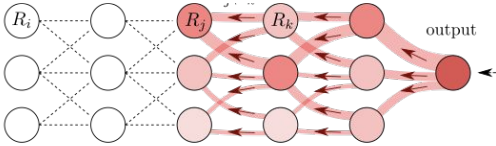
XAI reveals sources of predictability that vary in time and space

Toms, Barnes & Hurrell (2021)

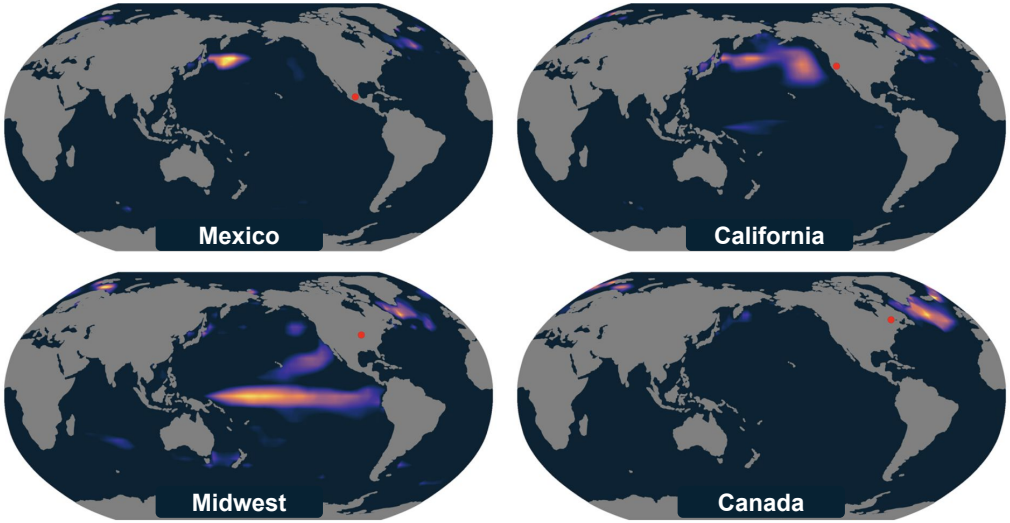
Predicting 5-year average surface temperature at each grid point
Applied to 1200 years of climate model CESM2 control simulation



Prediction of average land surface temperature anomaly over the next 5 years



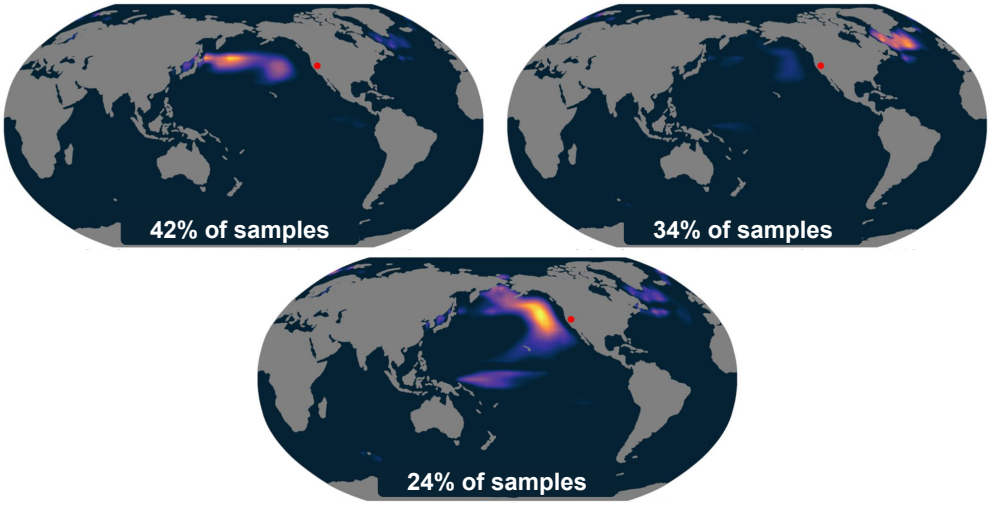
XAI

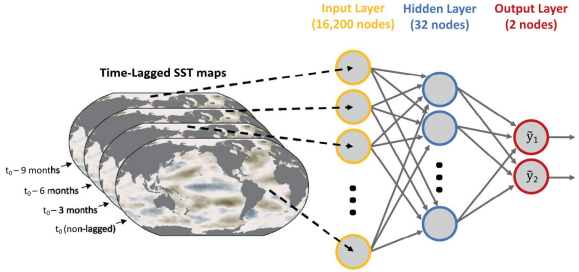


XAI reveals sources of predictability that vary in time and space

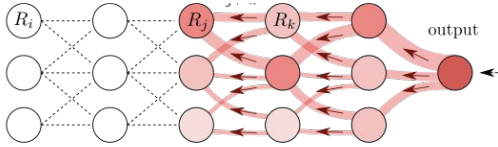
Toms, Barnes & Hurrell (2021)

Predicting 5-year average surface temperature at each grid point
Applied to 1200 years of climate model CESM2 control simulation





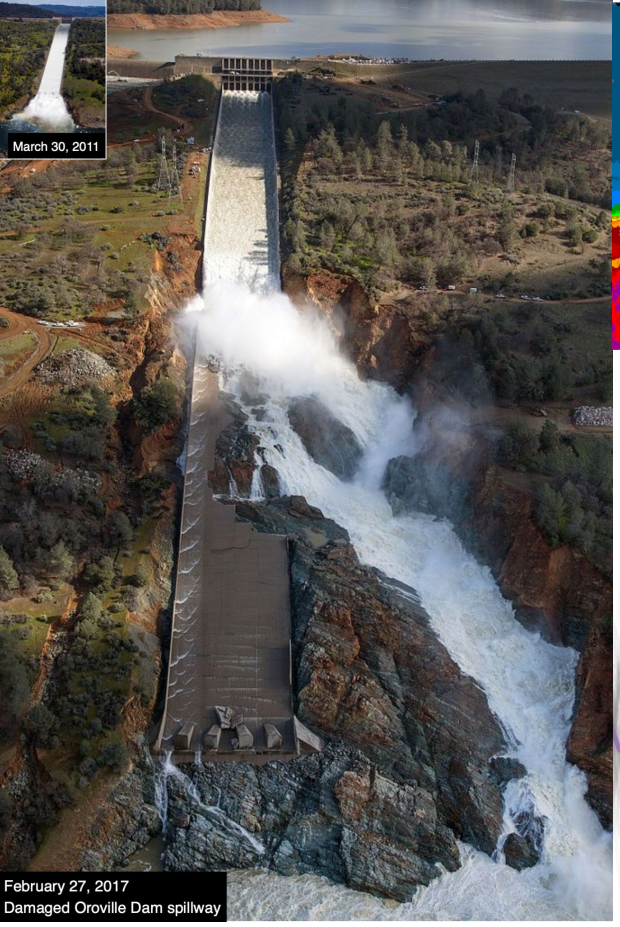
Prediction of average land surface temperature anomaly over the next 5 years

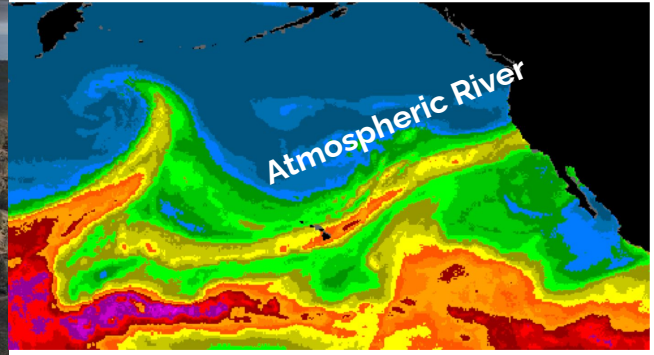


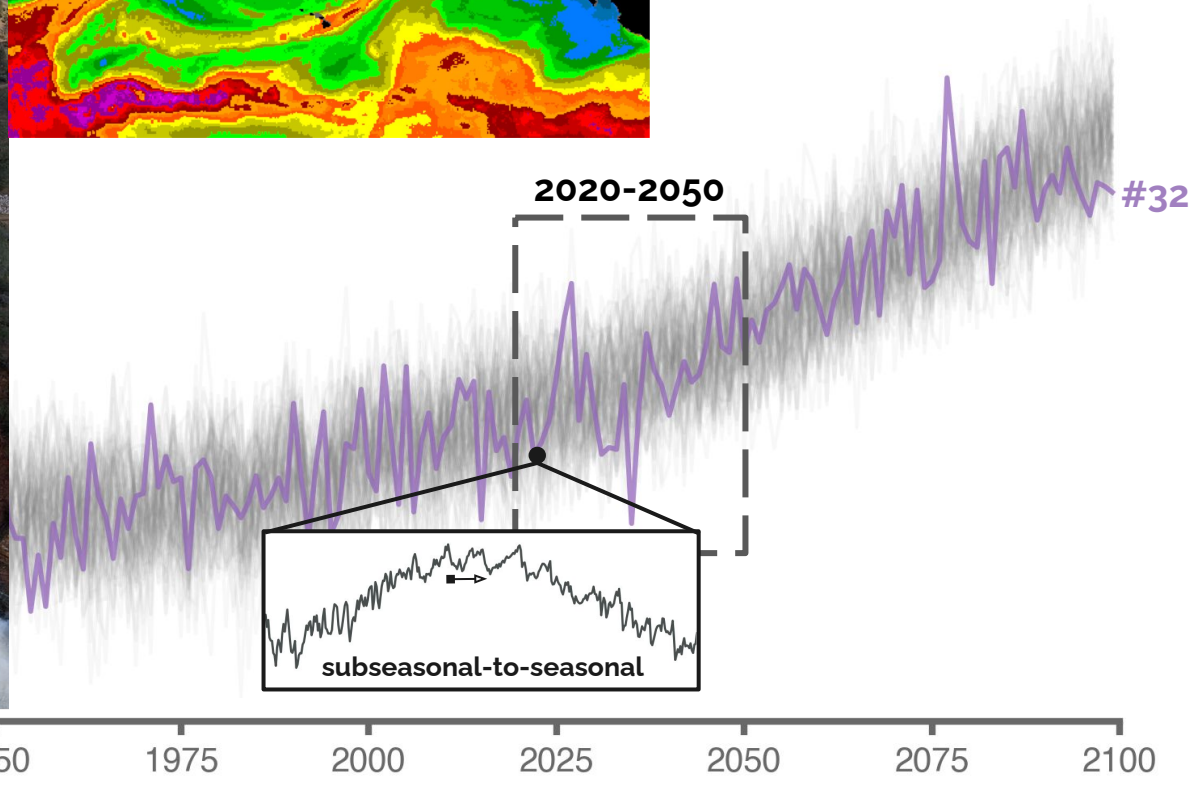
XAI

XAI reveals sources of predictability that vary in time and space

Toms, Barnes & Hurrell (2021)



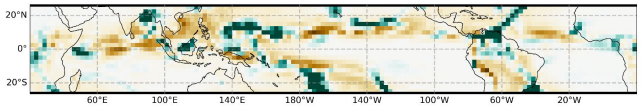




Surface temperature over Chicago, IL
MPI-ESM Large Ensemble; historical + RCP8.5

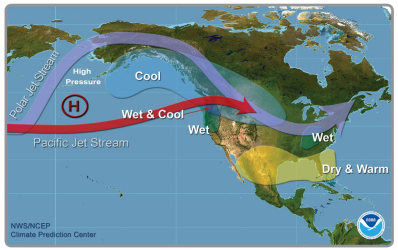
Input: Daily tropical precipitation

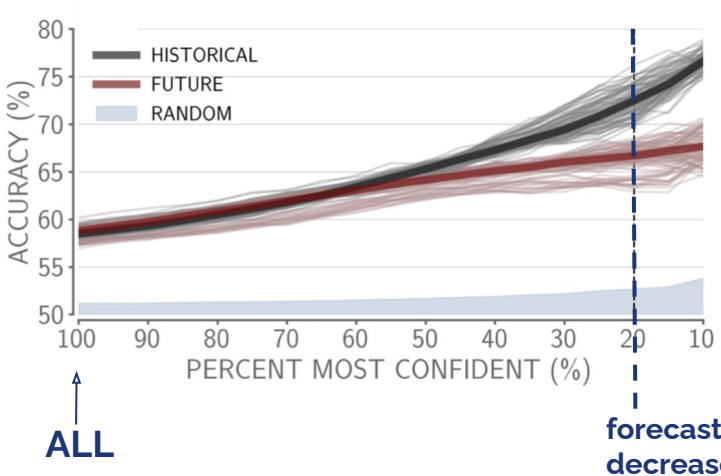
Trained on climate model **CESM2 [800 years of daily data]**



AI

Output: Pacific circulation 3 weeks later



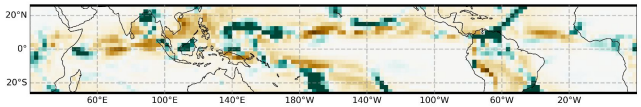


XAI allows us to quantify predictability in past and future climates and assess its sources.

Mayer & Barnes (2021, 2022)

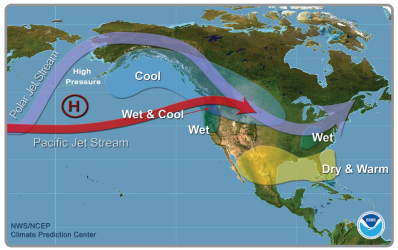
Input: Daily tropical precipitation

Trained on climate model **CESM2 [800 years of daily data]**



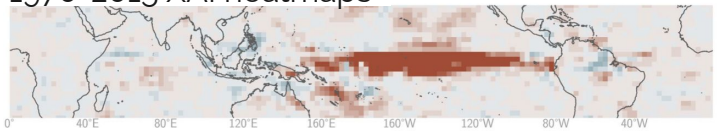
AI

Output: Pacific circulation 3 weeks later

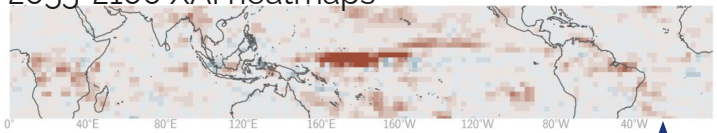


Most Relevant Regions

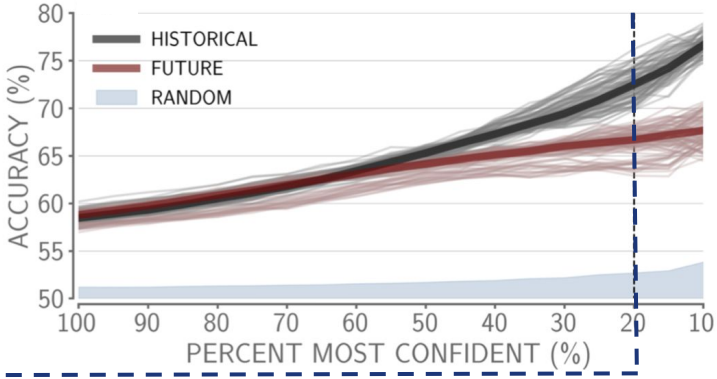
1970-2015 XAI heatmaps



2055-2100 XAI heatmaps



XAI



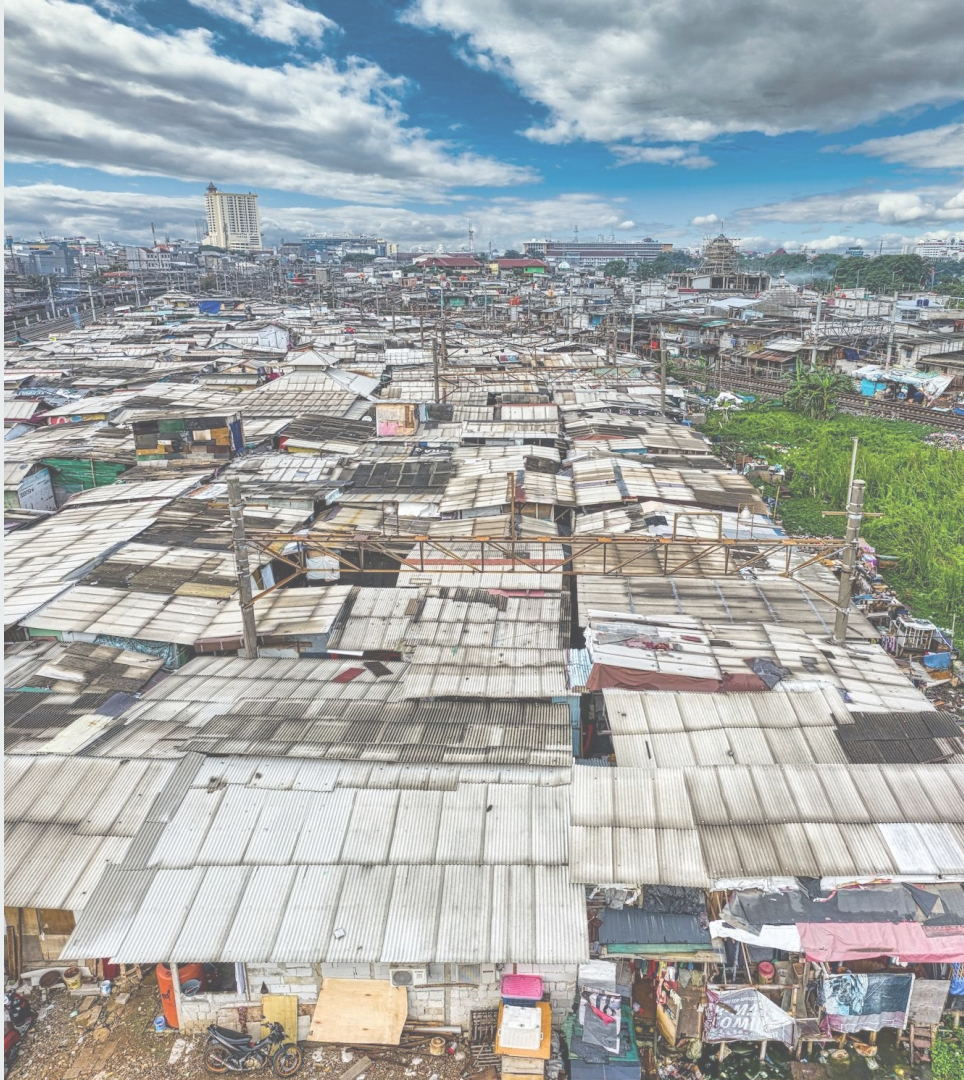
XAI allows us to quantify predictability in past and future climates and assess its sources.

Mayer & Barnes (2021, 2022)

The Earth system is coupled and complex.

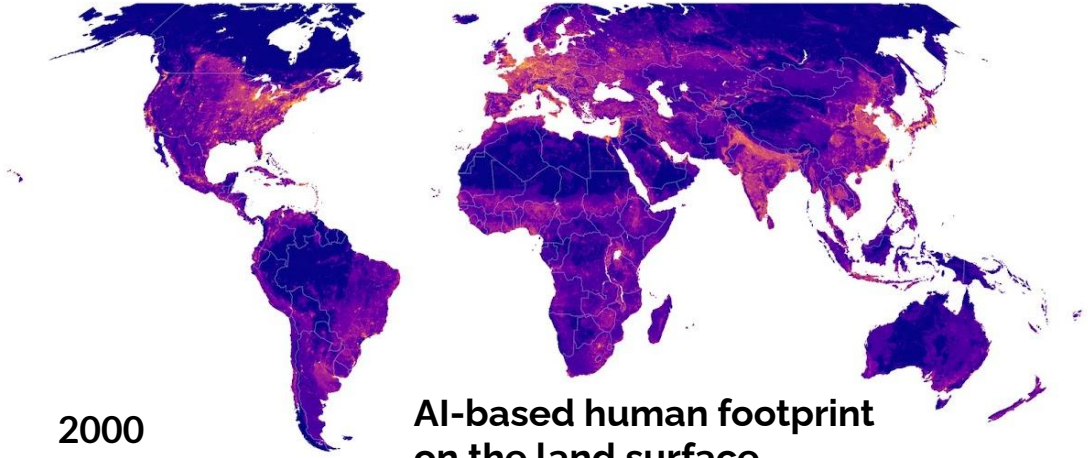
Human development and environmental factors cannot be easily decoupled.

Can we quantify the present and anticipate future changes in human development under future climate drivers?

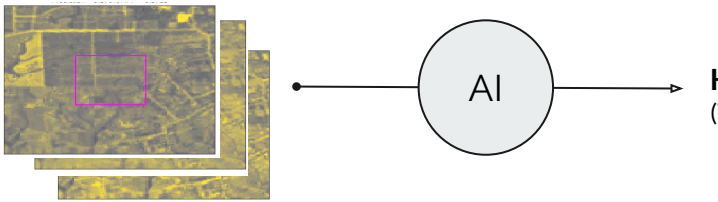


Human-Earth System Interactions

Quantify the present and predict the future transformations of the earth system by humanity in the presence of climate hazards



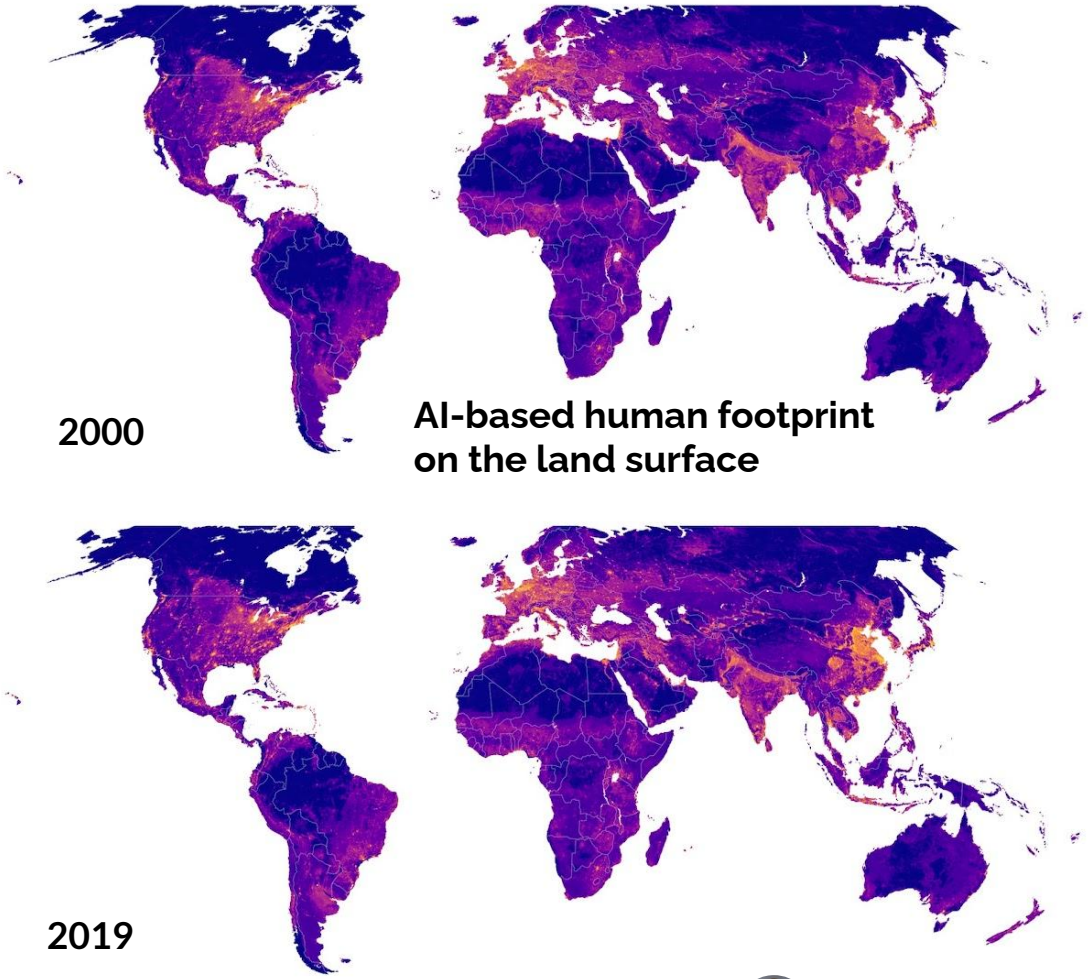
Landsat imagery from year 2000
3 bands x 120 pixels x 120 pixels



Keys, Barnes & Carter (2021)

Human-Earth System Interactions

Quantify the present and predict the future transformations of the earth system by humanity in the presence of climate hazards

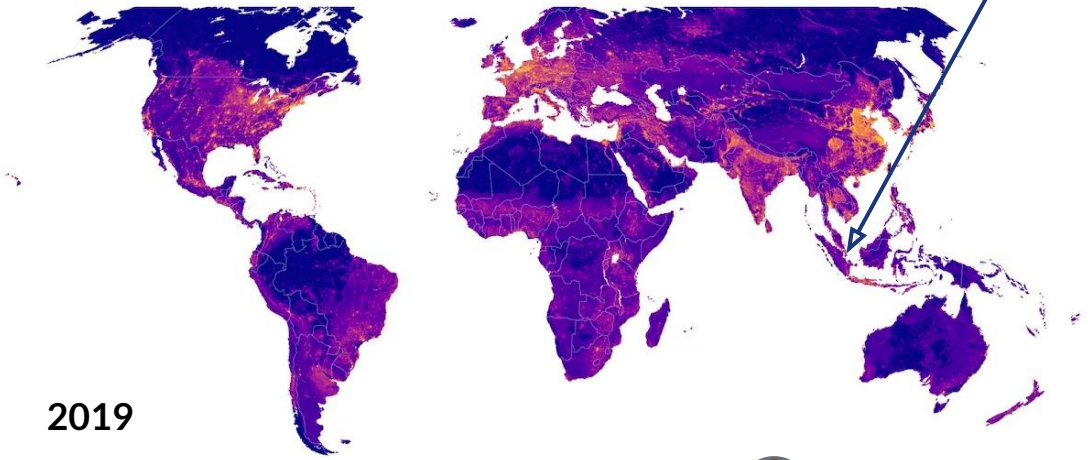


Keys, Barnes & Carter (2021)

Human-Earth System Interactions

Quantify the present and predict the future transformations of the earth system by humanity in the presence of climate hazards

2000 2019 XAI

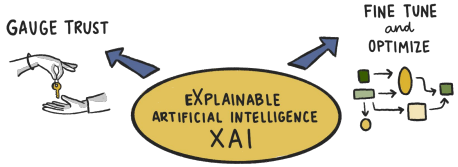


2019

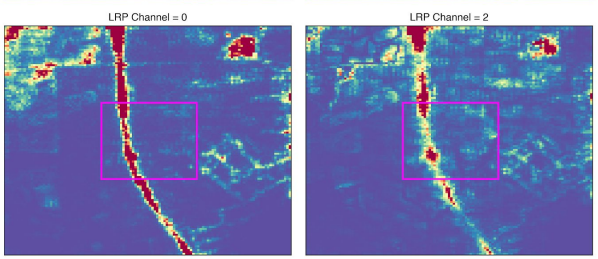
Keys, Barnes & Carter (2021)

Human-Earth System Interactions

Quantify the present and predict the future transformations of the earth system by humanity in the presence of climate hazards



XAI helped us determine which Landsat channels were most important for the network's decision; aiding us in refining our inputs.

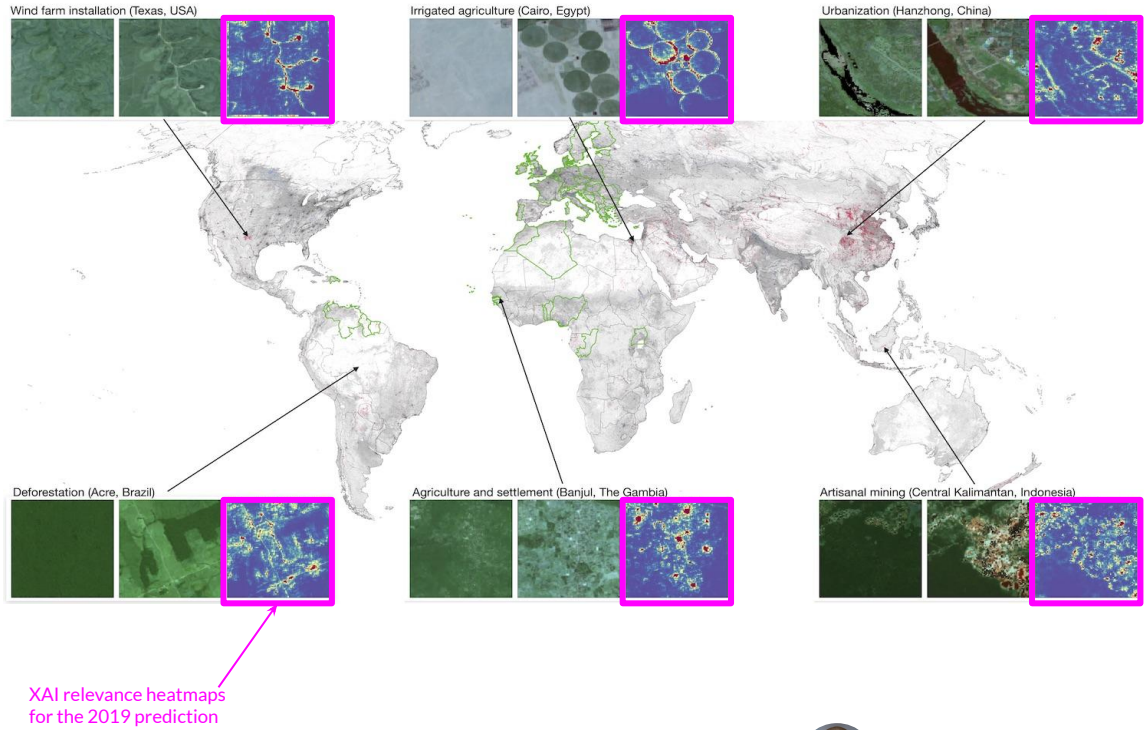


Keys, Barnes & Carter (2021)

Human-Earth System Interactions

Quantify the present and predict the future transformations of the earth system by humanity in the presence of climate hazards

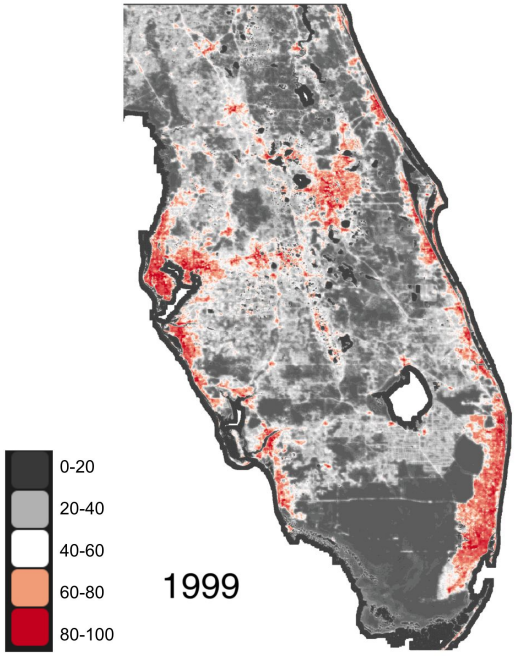
Changes in human footprint between 2000 and 2019



Keys, Barnes & Carter (2021)

Human-Earth System Interactions

Quantify the present and predict the future transformations of the earth system by humanity in the presence of climate hazards



Assess how climate hazards may interfere with global sustainable development in the coming years



Orihuela-Pinto, Keys, Davenport & Barnes (in prep)

Interpretable AI

AI models that are interpretable by design

- Explainable AI tells us ***where***, but not ***how***.
- Interpretable AI **explicitly incorporates the decision-making process** into its structure

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x
nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Rudin (2019)

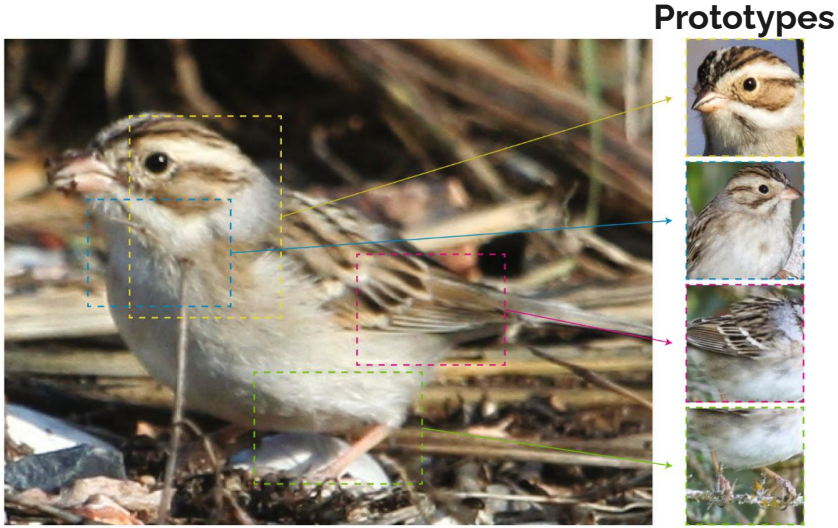
Interpretable AI

AI models that are interpretable by design

- Explainable AI tells us ***where***, but not ***how***.
- Interpretable AI **explicitly incorporates the decision-making process** into its structure

Our Current Goal:
work toward building AI models that mimic scientific human reasoning to improve intrinsic interpretability

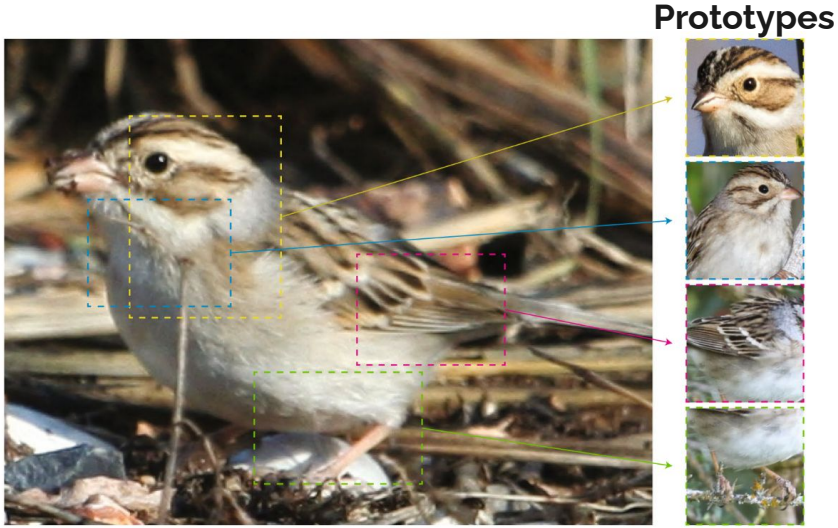
This Looks Like ***That***



Rudin (2019)
Chen et al. (2019)

Building interpretable AI for climate applications

This Looks Like ***That*** **Anywhere**



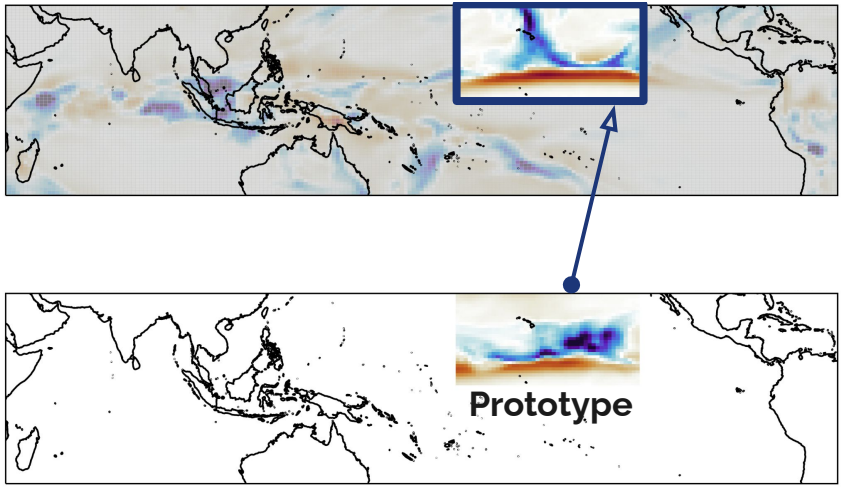
Rudin (2019)
Chen et al. (2019)

Building interpretable AI for climate applications

This Looks Like *That* ~~Anywhere~~ There

Building interpretable AI for climate applications

This Looks Like ***That*** ~~**Anywhere**~~
There

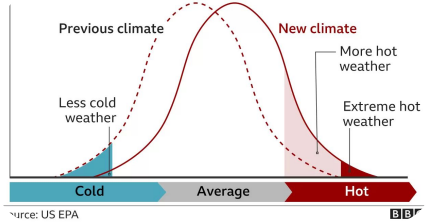


the network is constrained to make its predictions based on only #N prototypes – it needs to learn which ones

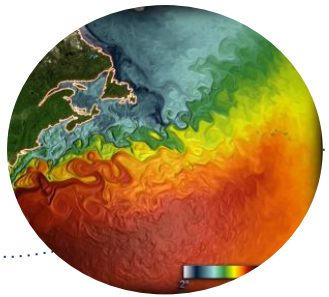
Building interpretable AI for climate applications

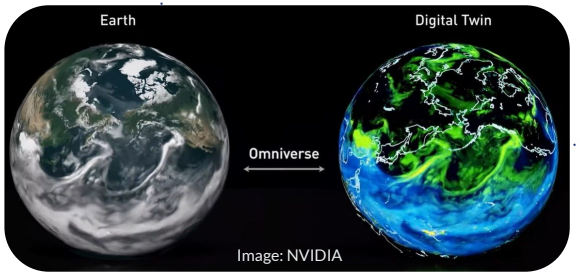
Gordillo and Barnes (under review)

Climate change influence the performance of AI models



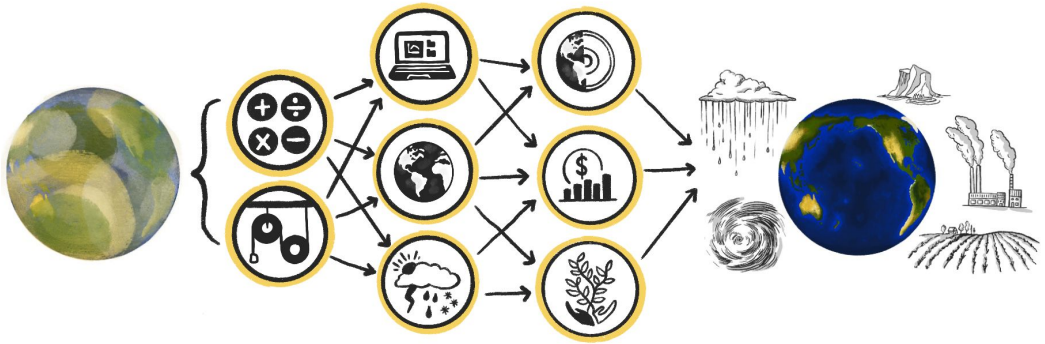
Role of AI in high-resolution Earth system modeling





AI for weather & climate model replacement

Additional New Frontiers for XAI + Climate



Thank you.

eabarnes@colostate.edu
https://barnes.atmos.colostate.edu
github: eabarnes1010