MAPP NA19OAR4310289
MAPP NA16OAR4310064

Computational Graduate Fellowships
PCMDI

CAREER AGS-1749261
HDR OAC-1934668
AI Institute ICER-2019758
Graduate Research Fellowships

And with special thanks to David Wallerstein

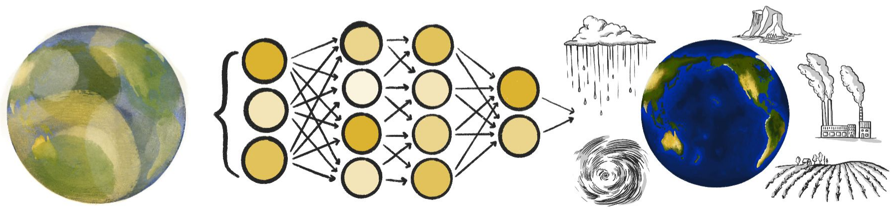
Explainable AI for Climate Science: Detection, Prediction and Discovery

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

email: eabarnes@colostate.edu
website: https://barnes.atmos.colostate.edu
twitter: @atmosbarnes
github: eabarnes1010

ATMOSPHERIC SCIENCE
COLORADO STATE UNIVERSITY

WCRP hybrid symposium on Frontiers in Subseasonal to Decadal Prediction
March 28, 2023



One of our jobs as scientists is to sift through piles of data and try to extract useful relationships that apply elsewhere, i.e. that are applicable “out of sample”.

This is what many machine learning methods are designed to do.

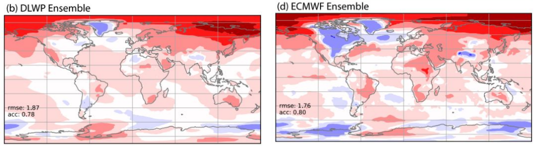
*throughout this talk I will use “AI” and “machine learning” or “ML” interchangeably

ML for S2D Predictability & Prediction

There are many ways that ML can be applied to try and improve understanding of S2D predictability

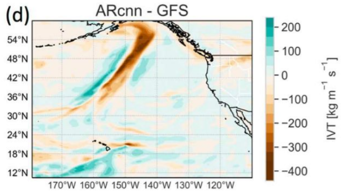
Deep Learning WP Models

e.g. *Weyn et al. (2021)*



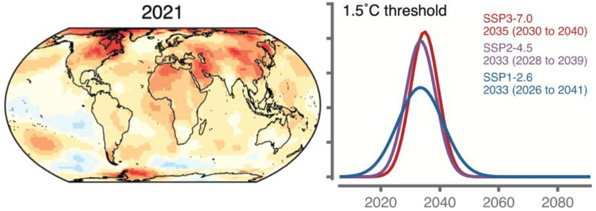
Predicting the Errors of Forecast Systems

e.g. *Chapman et al. (2019)*, Cahill et al. (in prep), Pan et al. (2021)



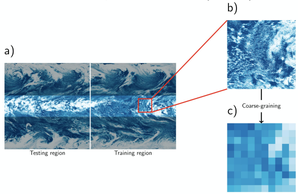
Post-processing of Multi-Model Ensembles

e.g. Haupt et al. (2021), Schumacher et al. (2021), Gronquist et al. (2020), *Diffenbaugh and Barnes (in review)*



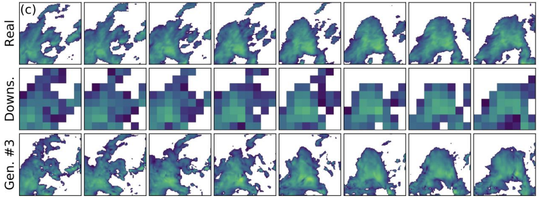
Improved Model Parameterizations

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



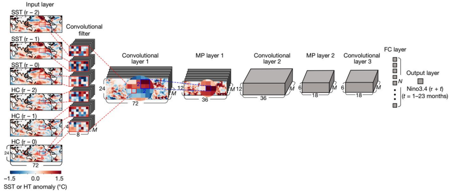
Downscaling for Regional Impacts

e.g. *Leinonen et al. (2020)*



Statistical Model Predictions

e.g. Mayer and Barnes (2021); Hassanibesheli et al. (2022); *Ham et al. (2019)*



Opening the Black Box with XAI

More and more papers are coming out demonstrating the use of ML explainability methods for geoscience

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.

AFFILIATIONS: MCGOVERN AND JERGENSEN—University of Oklahoma, Norman, Oklahoma; LAGERQUIST—Cooperative Institute for Mesoscale Meteorological Studies, and University of Oklahoma, Norman, Oklahoma; GAGNE—National Center for Atmospheric Research, Boulder, Colorado; ELMORE AND SMITH—Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/National Severe Storms Laboratory, Norman, Oklahoma; HOMEYER—School of Meteorology, University of Oklahoma, Norman, Oklahoma
CORRESPONDING AUTHOR: Amy McGovern, amcgovern@ou.edu

The abstract for this article can be found in this issue, following the table of contents.

DOI:10.1175/BAMS-D-18-0195.1
A supplement to this article is available online (10.1175/BAMS-D-18-0195.2)
In final form 20 June 2019

Machine learning (ML) and deep learning (DL; LeCun et al. 2015) have recently achieved breakthroughs across a variety of fields, including the world's best Go player (Silver et al. 2016, 2017), medical diagnosis (Rakhlin et al. 2018), and galaxy classification (Dieleman et al. 2015). Simple forms of ML (e.g., linear regression) have been ...

JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS002002

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

Benjamin A. Toms[1], Elizabeth A. Barnes[1], and Imme Ebert-Uphoff[2,3]

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

Key Points:
- Interpretable neural networks can identify the coherent spatial patterns of known modes of Earth system variability
- The layerwise relevance propagation and backward optimization methods enable new ways to use neural networks for geoscientific research
- We propose that the interpretation of what a neural network has learned can be used as the ultimate scientific outcome of a trained network

Supporting Information:
- Supporting Information S1

Correspondence to:
B. A. Toms,
ben.toms@colostate.edu

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

Abstract Neural networks have become increasingly prevalent within the geosciences, although a common limitation of their usage has been a lack of methods to interpret what the networks learn and how they make decisions. As such, neural networks have often been used within the geosciences to most accurately identify a desired output given a set of inputs, with the interpretation of what the network learns used as a secondary metric to ensure the network is making the right decision for the right reason. Neural network interpretation techniques have become more advanced in recent years, however, and we therefore propose that the ultimate objective of using a neural network can also be the interpretation of what the network has learned rather than the output itself. We show that the interpretation of neural networks can enable the discovery of scientifically meaningful connections within geoscientific data. In particular, we use two methods for neural network interpretation called backward optimization and layerwise relevance propagation, both of which project the decision pathways of a network back onto the original input dimensions. To the best of our knowledge, LRP has not yet been applied to geoscientific research, and we believe it has great potential in this area. We show how these interpretation techniques can be used to reliably infer scientifically meaningful information from neural networks by applying them to common climate patterns. These results suggest that combining interpretable neural networks with novel scientific hypotheses will open the door to many new avenues in neural network-related geoscience research.

Plain Language Summary Neural networks, a form of machine learning, have become ...

Physics > Geophysics

[Submitted on 18 Mar 2021]

Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset

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

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 tru...

Article Contents
Capsule:
Abstract
Footnotes

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

Split-Screen PDF Share Cite Get Permissions

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.

Physics > Geophysics

[Submitted on 7 Feb 2022]

Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience

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

Convolutional neural networks (CNNs) have recently attracted great attention in geoscience due to their ability to capture non-linear system behavior and extract predictive spatiotemporal patterns. Given their black-box nature however, and the importance of prediction explainability, methods of explainable artificial intelligence (XAI) are gaining popularity as a means to explain the CNN decision-making strategy. Here, we establish an intercomparison of some of the most popular XAI methods and investigate their fidelity in explaining CNN decisions for geoscientific applications. Our goal is to raise awareness of the theoretical limitations of these methods and gain insight into the relative strengths and weaknesses to help guide best practices. The considered XAI methods are first applied to an idealized attribution benchmark, where the ground truth of explanation of the network is known a priori, to help objectively assess their performance. Secondly, we apply XAI to a climate-related prediction setting, namely to explain a CNN that is trained to predict the number of atmospheric rivers in daily snapshots of climate simulations. Our results highlight several important issues of XAI methods (e.g., gradient shattering, inability to distinguish the sign of attribution, ignorance to zero input) that have previously been overlooked in our field and, if not considered cautiously, may lead to a distorted picture of the CNN decision-making strategy. We envision that our analysis will motivate further investigation into XAI fidelity and will help towards a cautious implementation of XAI in geoscience, which can lead to further exploitation of CNNs and deep learning for prediction problems.

XAI Attribution Methods

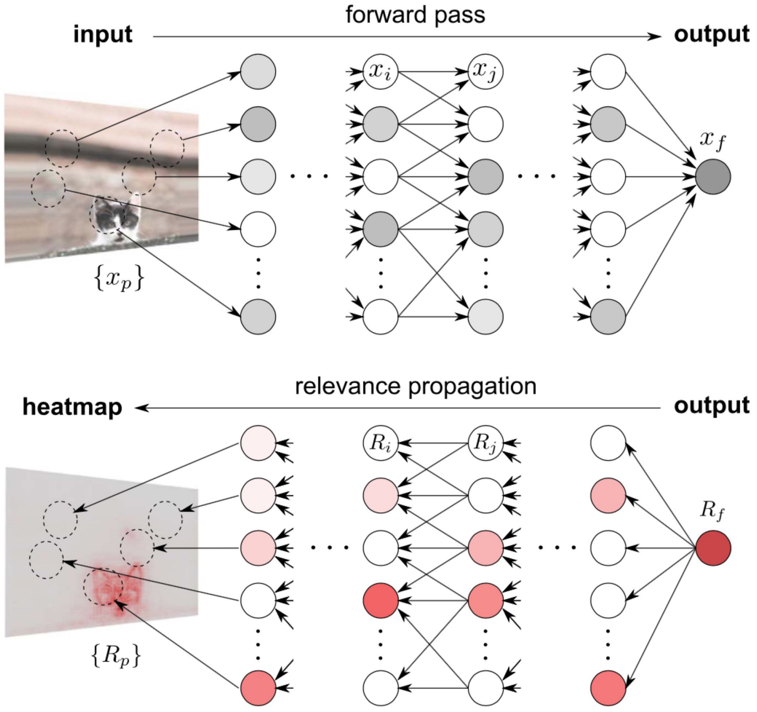
Attribution heatmaps are largely consistent with how many climate scientists pose questions

Prediction

of 1 sample

Attribution

of 1 sample



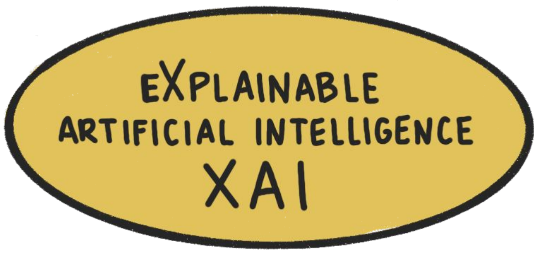
Pr(cat)=.8

Pr(cat)=.8

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

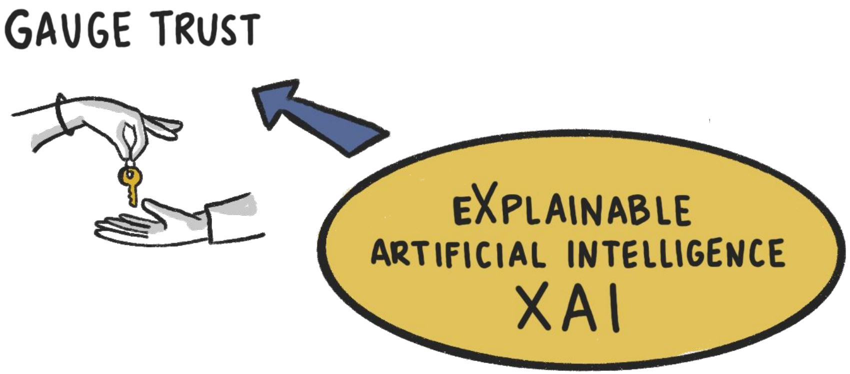
Reasons to care about XAI

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



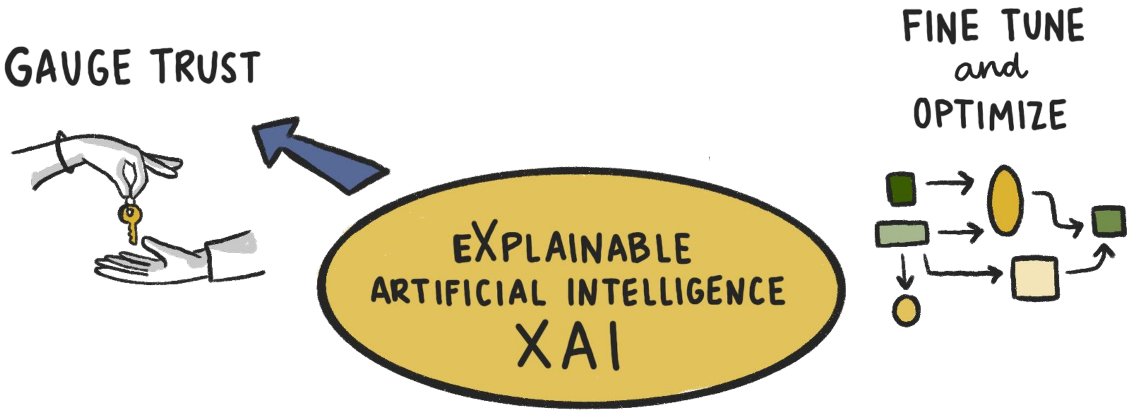
Reasons to care about XAI

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



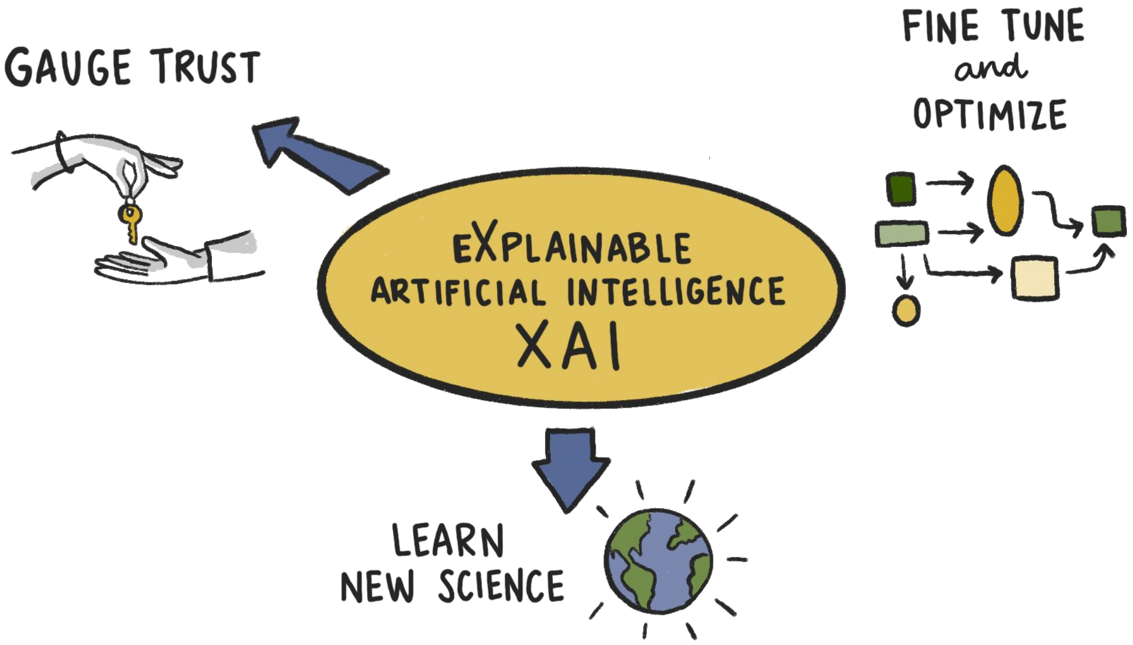
Reasons to care about XAI

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



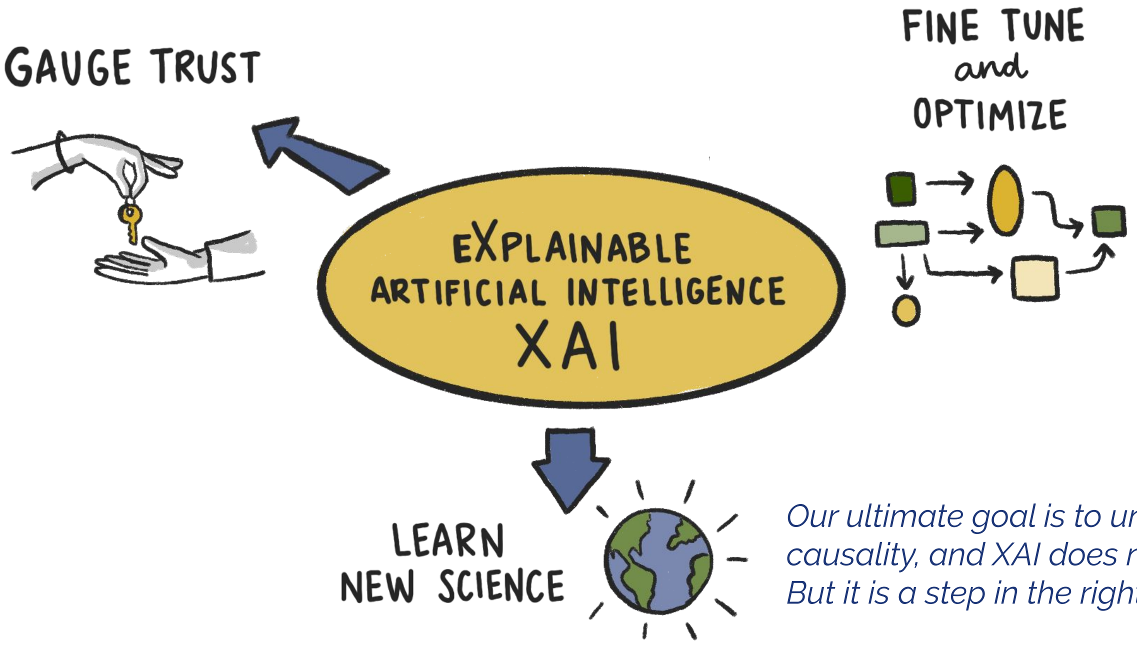
Reasons to care about XAI

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

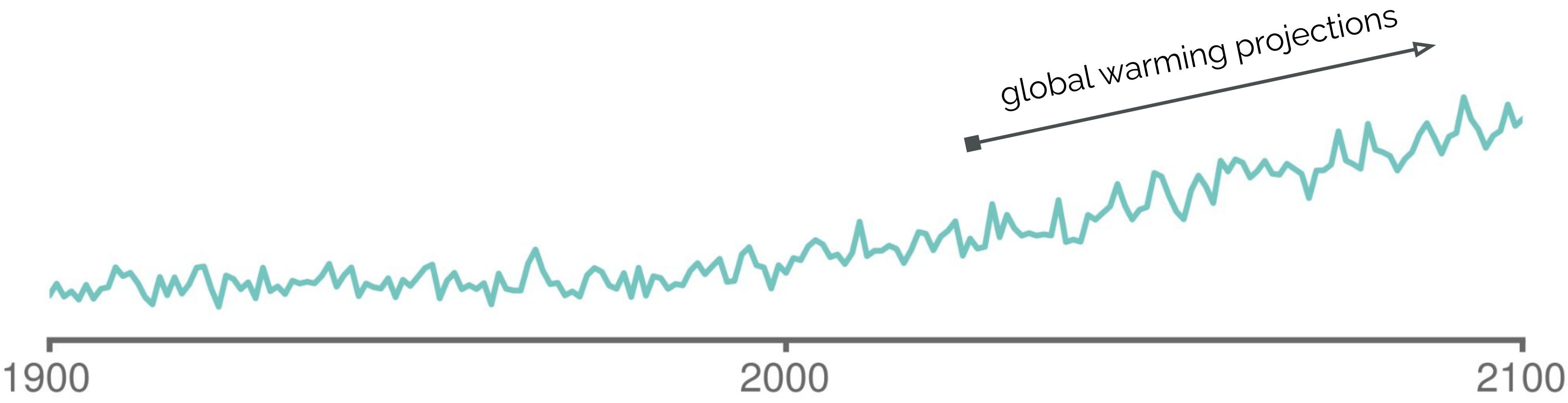
A scientist's ultimate goal is typically to understand "why?", but even if you don't care "why?" you should still care about XAI.



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

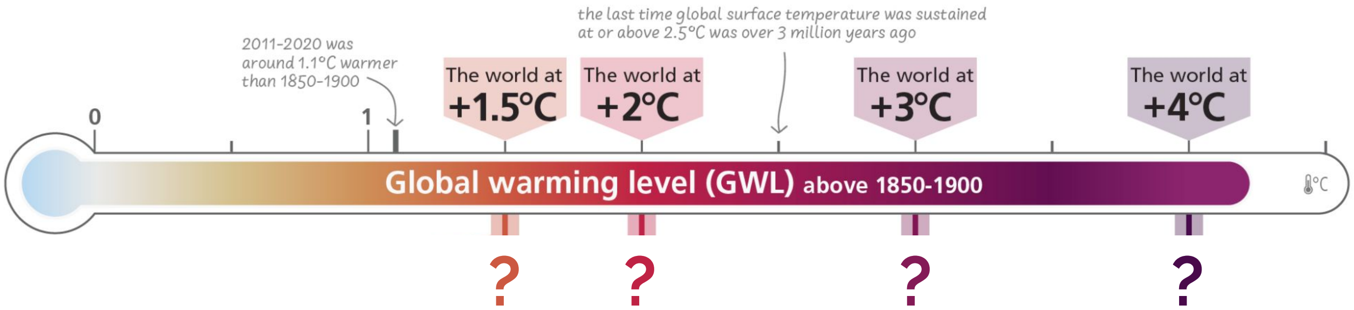
AI to leverage imperfect climate models to better **constrain future projections** by **fusing** simulations and observations.

e.g. Labe and Barnes (2022), Diffenbaugh and Barnes (2023), Rader et al. (2022), Labe and Barnes (2021), Barnes et al. (2020a), Barnes et al. (2019)



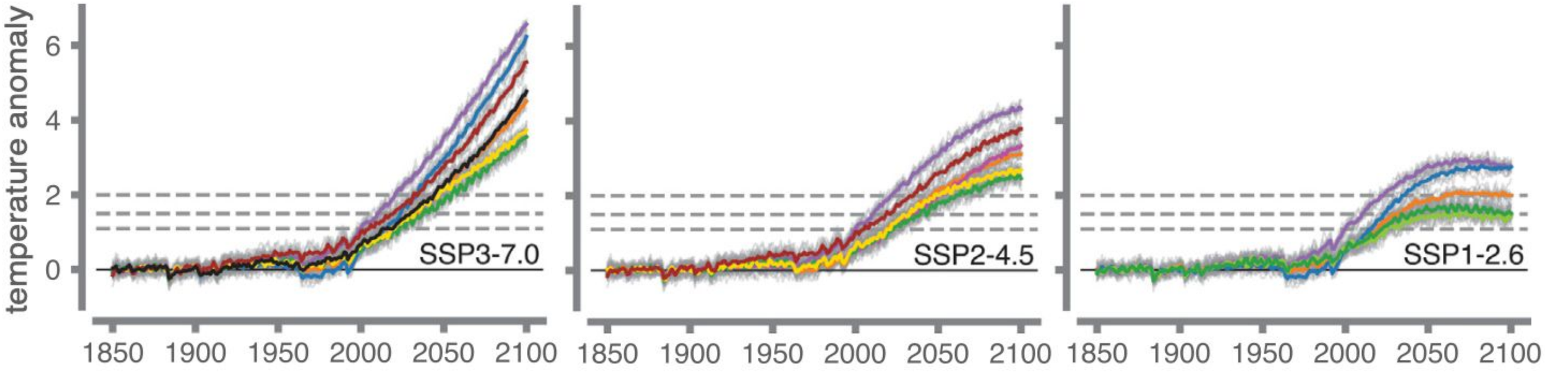
Surface temperature over Fort Collins, CO
CanESM2 simulation
historical + SSP3-7.0

Time Remaining Until Critical Warming Thresholds are Reached



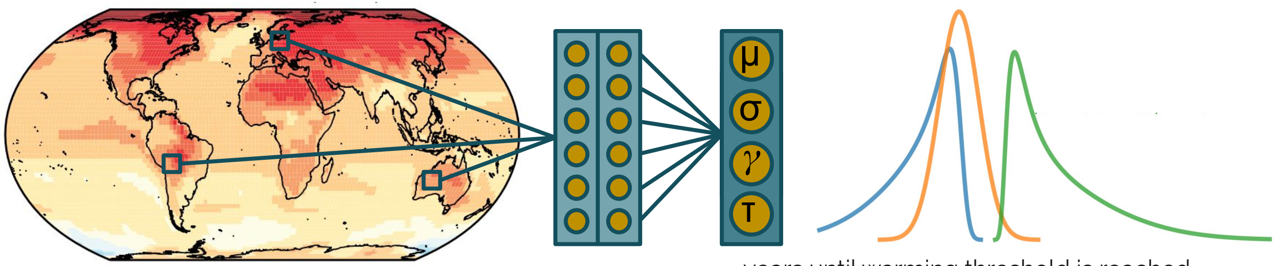
adapted from IPCC AR6 (2023); Summary for Policymakers

Time Remaining Until Critical Warming Thresholds are Reached



Diffenbaugh & Barnes (2023)

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

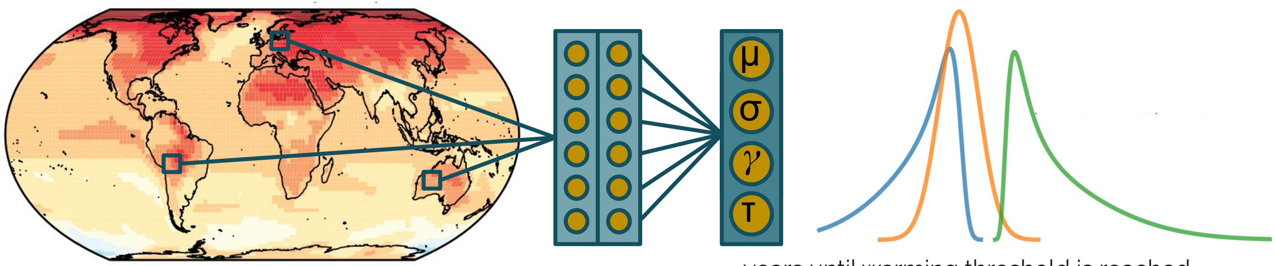


years until warming threshold is reached
+ uncertainties

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

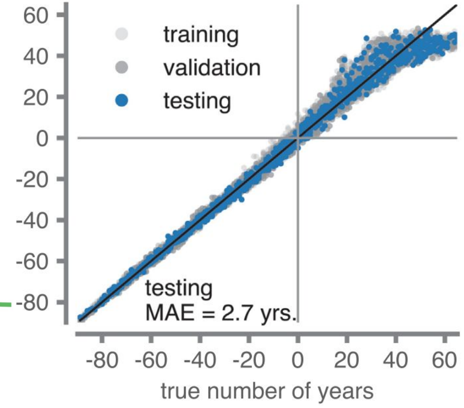
Diffenbaugh & Barnes (2023)

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



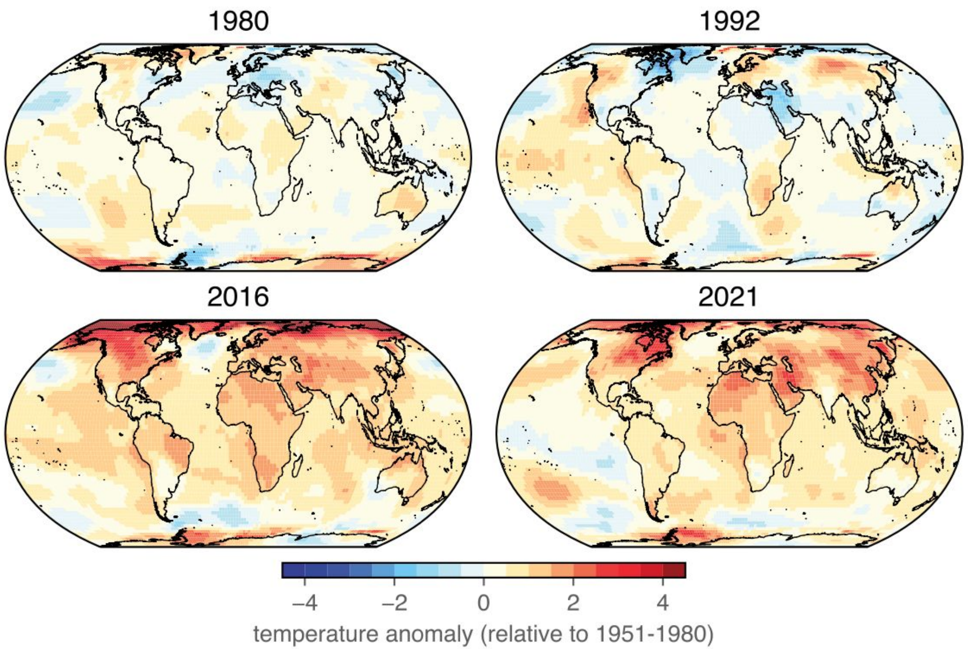
years until warming threshold is reached
+ uncertainties

Climate Model Results



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

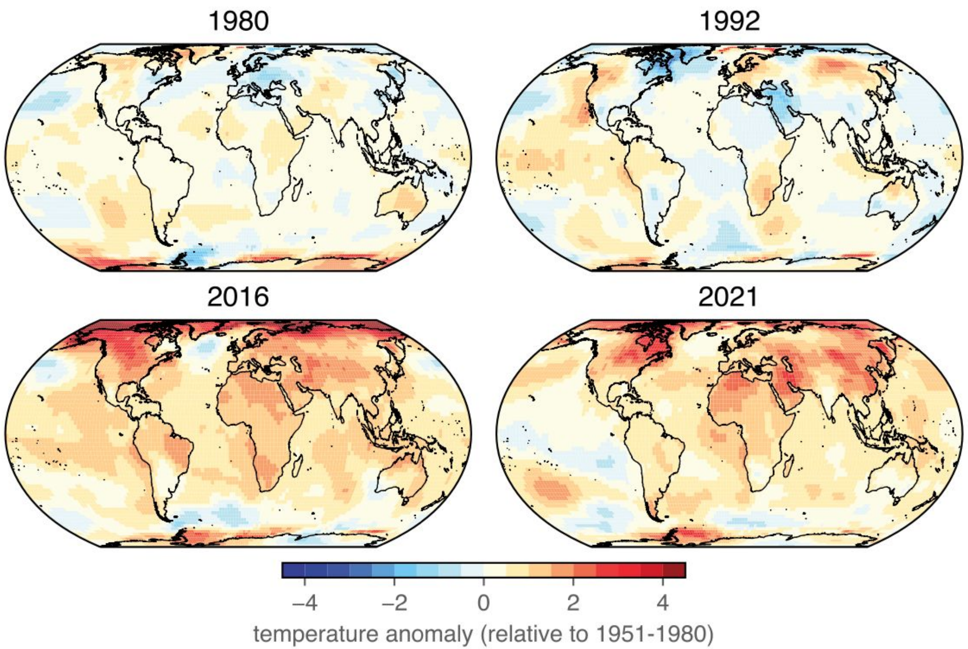
Diffenbaugh & Barnes (2023)

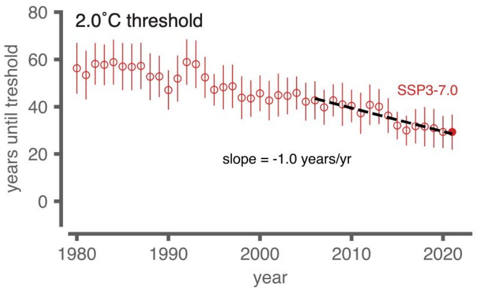


Observations
Berkeley Earth Surface Temperature

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

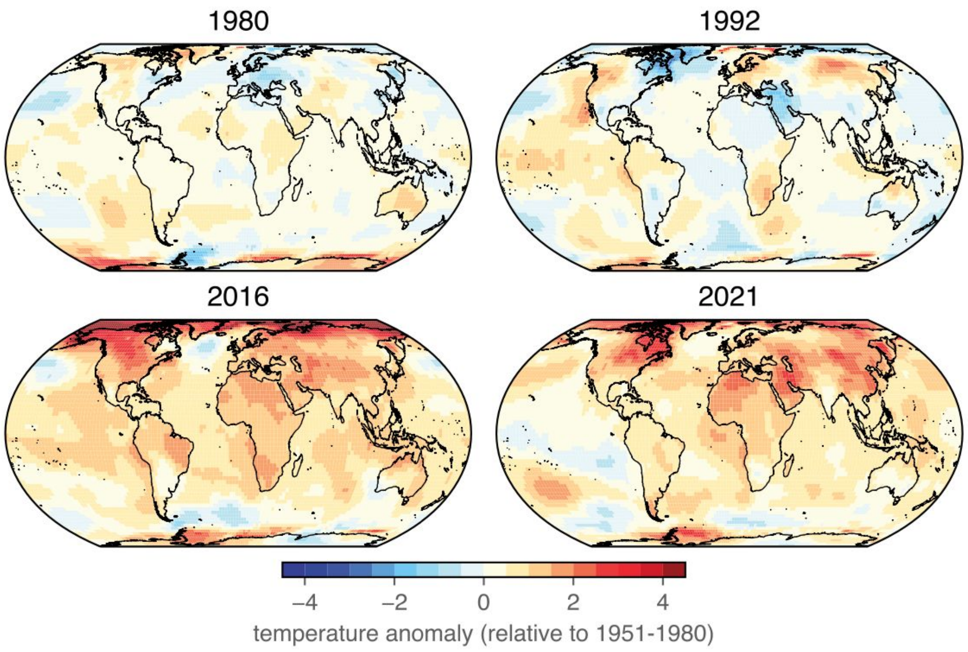
Diffenbaugh & Barnes (2023)

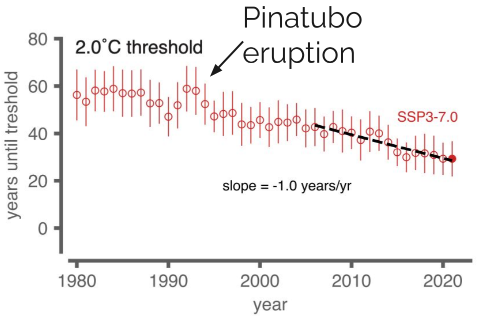




Use the trained AI model to predict thresholds based on maps of the observed climate

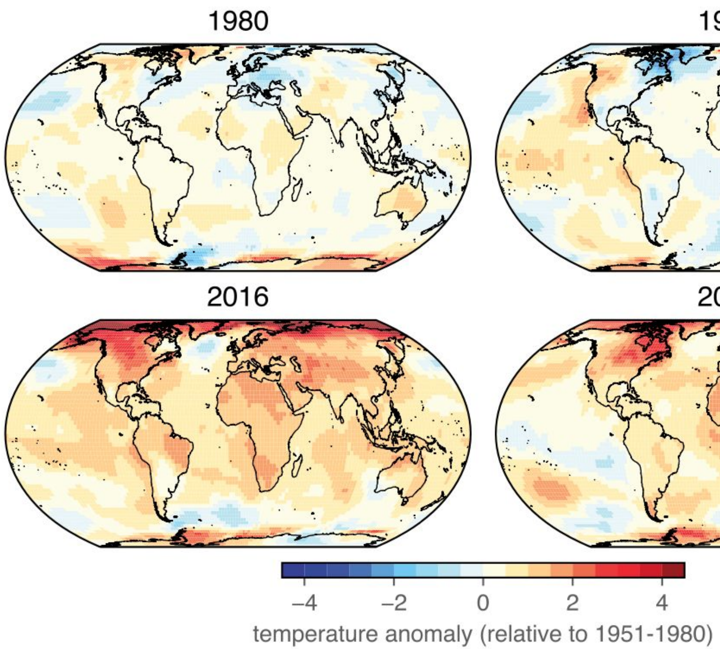
Diffenbaugh & Barnes (2023)

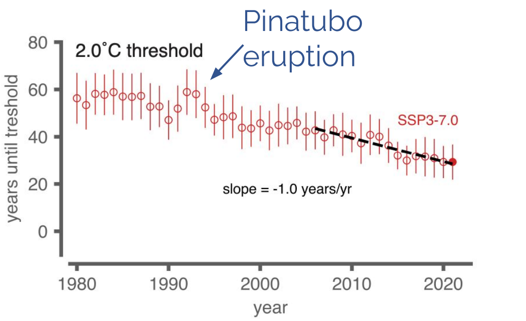


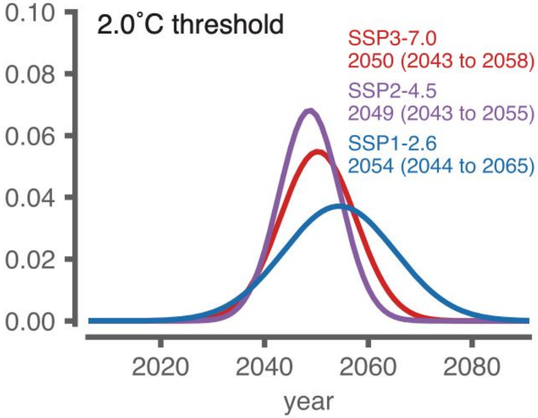


Use the trained AI model to predict thresholds based on maps of the observed climate

Diffenbaugh & Barnes (2023)

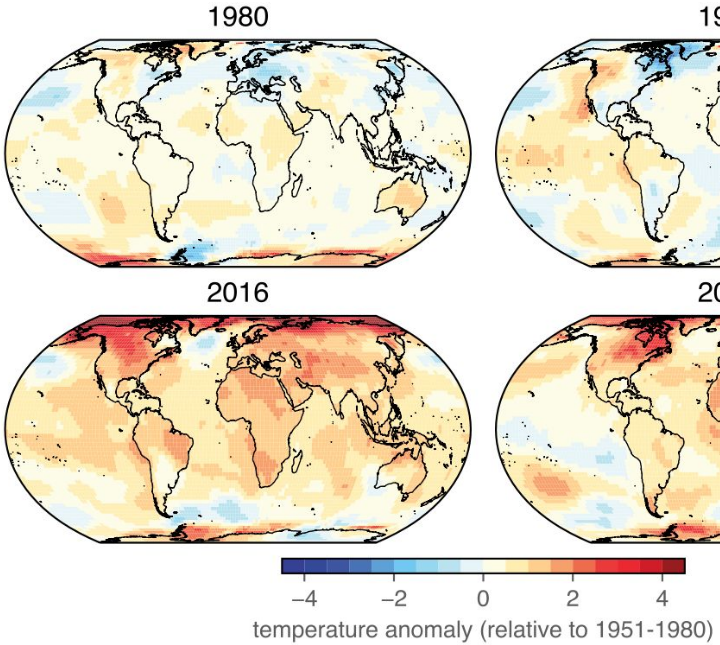




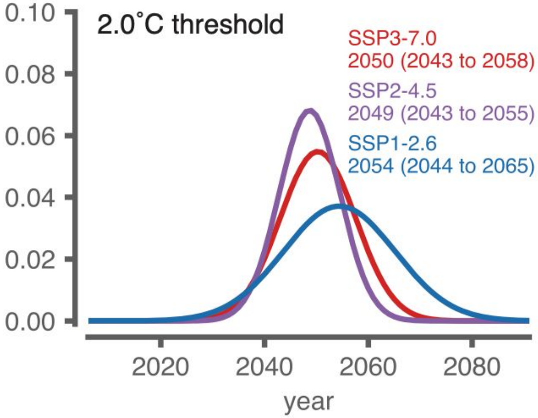


Use the trained AI model to predict thresholds based on maps of the observed climate

Diffenbaugh & Barnes (2023)

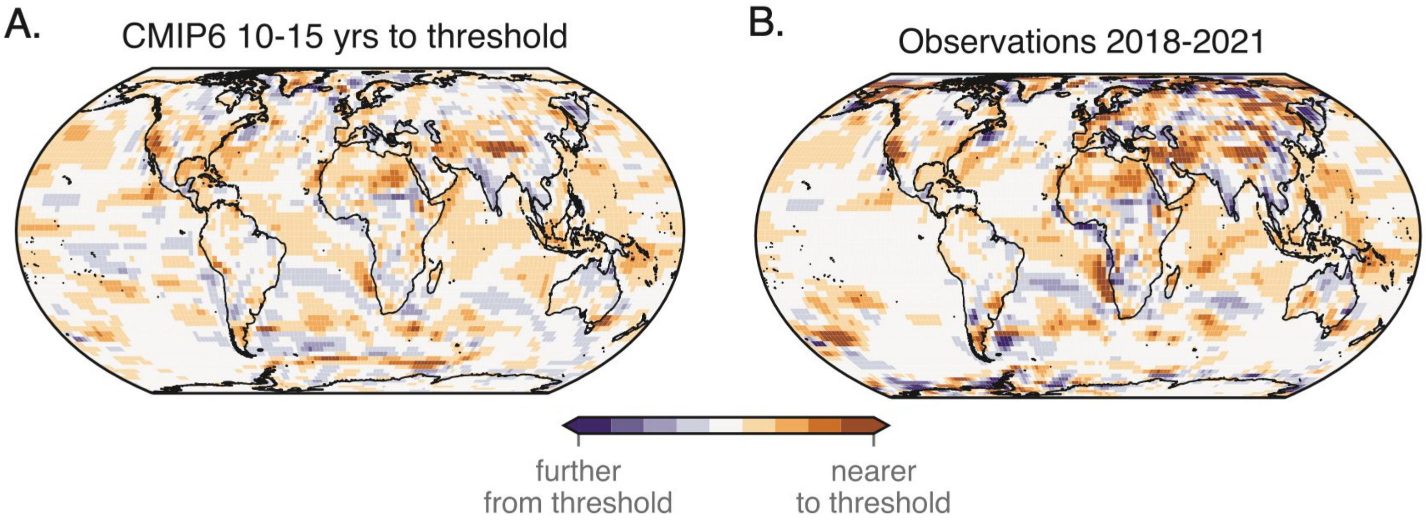


higher likelihood of reaching 2°C in the Low scenario than indicated in some previous assessments—although the possibility it could be avoided is not ruled out.



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

Diffenbaugh & Barnes (2023)



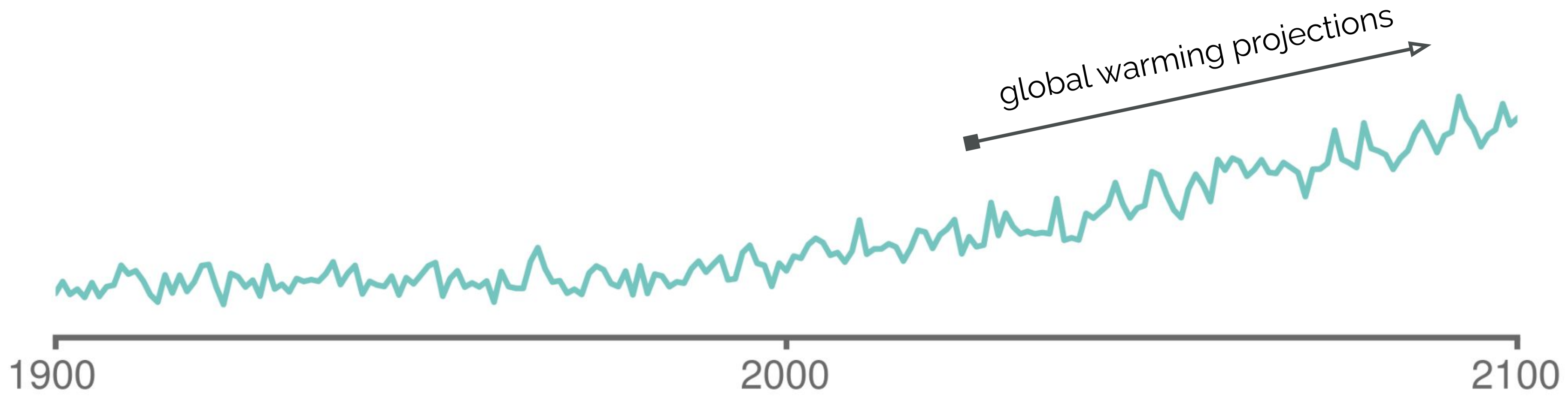
XAI

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

Diffenbaugh & Barnes (2023)

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

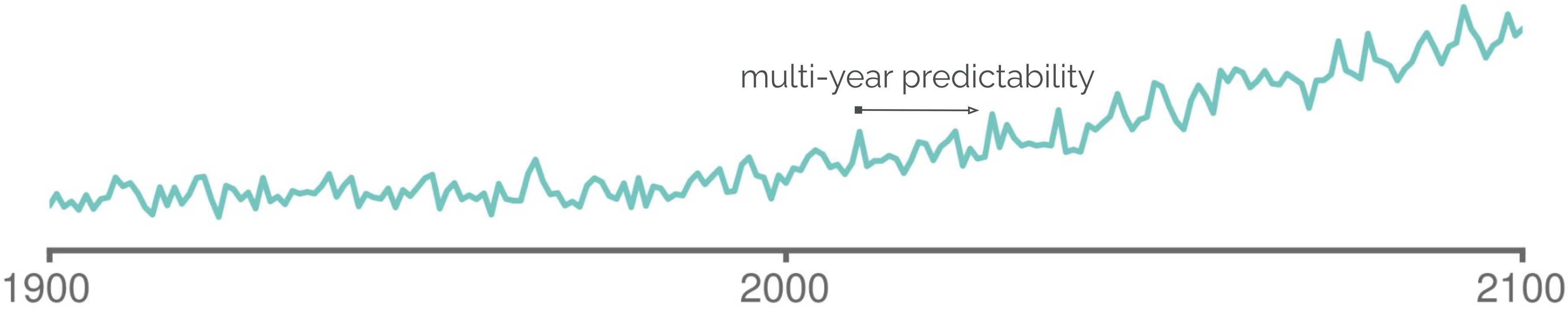
e.g. Labe and Barnes (2022), Diffenbaugh and Barnes (2023), Rader et al. (2022), Labe and Barnes (2021), Barnes et al. (2020a), Barnes et al. (2019)



Surface temperature over Fort Collins, CO
CanESM2 simulation
historical + SSP3-7.0

AI to explore earth system **predictability weeks-to-years in advance** to study prediction, dynamics, and change.

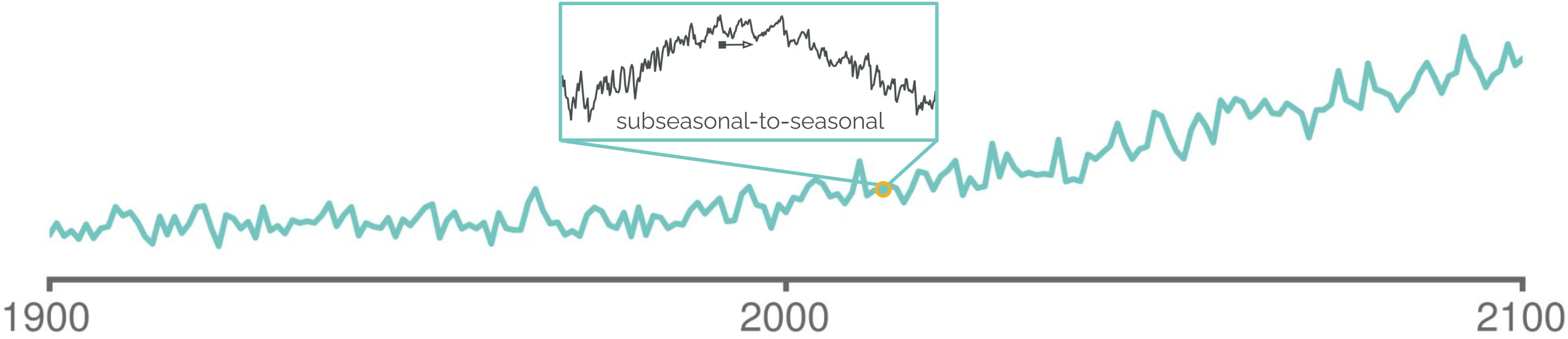
e.g. Gordon and Barnes (2022), Labe and Barnes (2022), Gordon, Barnes and Hurrell (2021), Toms, Barnes and Hurrell (2021), Mayer and Barnes (2022), Mayer and Barnes (2021), Barnes et al. (2020), Barnes et al. (2020)



Surface temperature over Fort Collins, CO
CanESM2 simulation
historical + SSP3-7.0

AI to explore earth system **predictability weeks-to-years in advance** to study prediction, dynamics, and change.

e.g. Gordon and Barnes (2022), Labe and Barnes (2022), Gordon, Barnes and Hurrell (2021), Toms, Barnes and Hurrell (2021), Mayer and Barnes (2022), Mayer and Barnes (2021), Barnes et al. (2020), Barnes et al. (2020)



Surface temperature over Fort Collins, CO
CanESM2 simulation
historical + SSP3-7.0

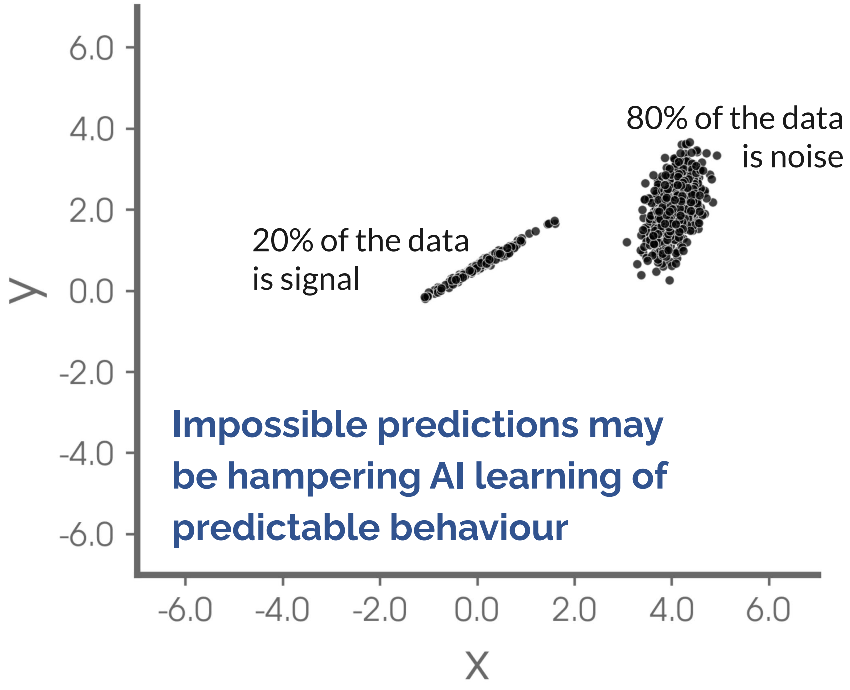
But, climate prediction is incredibly challenging. We cannot expect to make perfect predictions all of the time.

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.

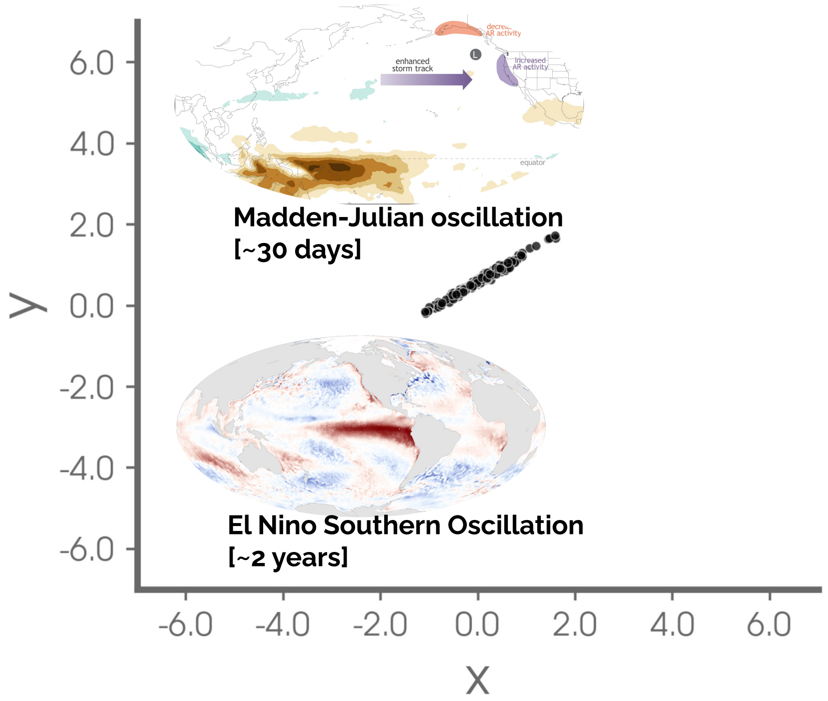


Barnes, Barnes and Gordillo (2021)
Barnes and Barnes (2021a, 2021b)

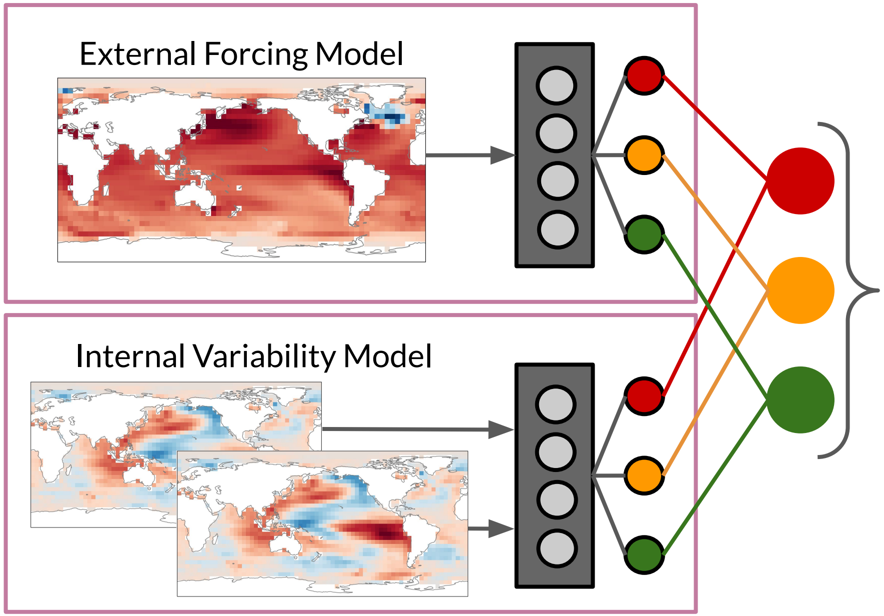
But, climate prediction is incredibly challenging. We cannot expect to make perfect predictions all of the time.

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)

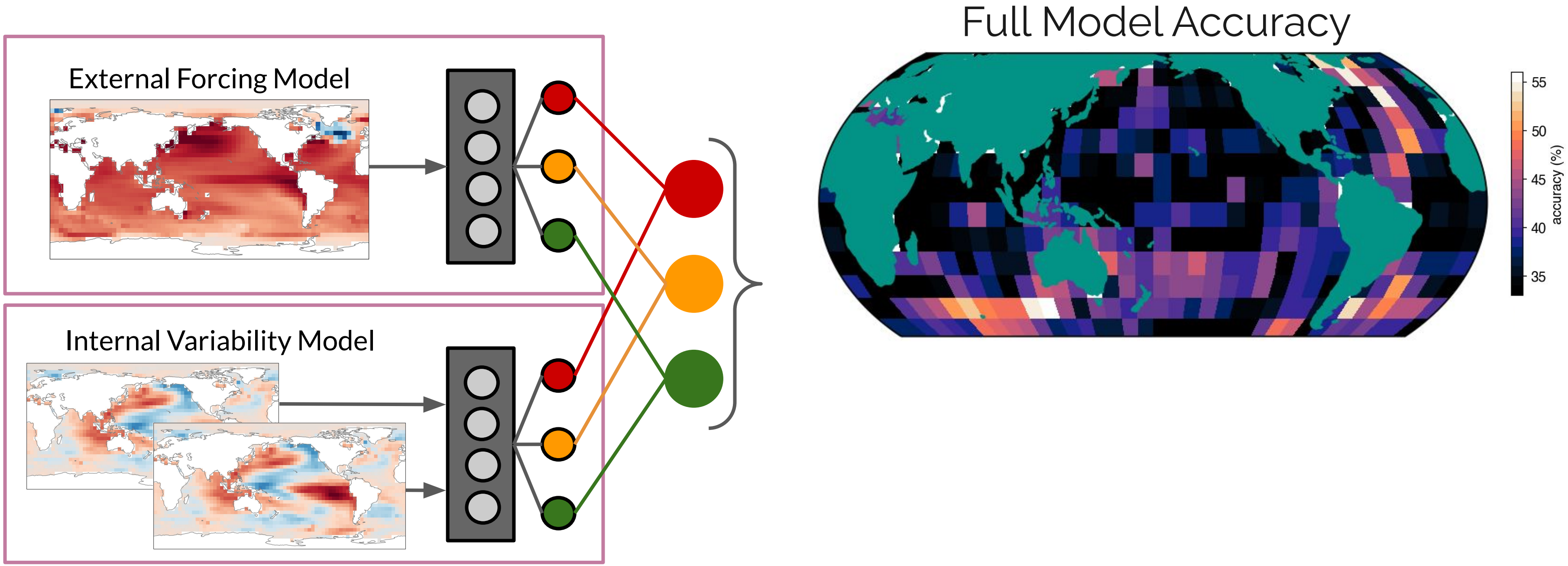


prediction of
future regional
sea surface
temperature

thoughtful choices of inputs and outputs can allow attribution of sources of predictability

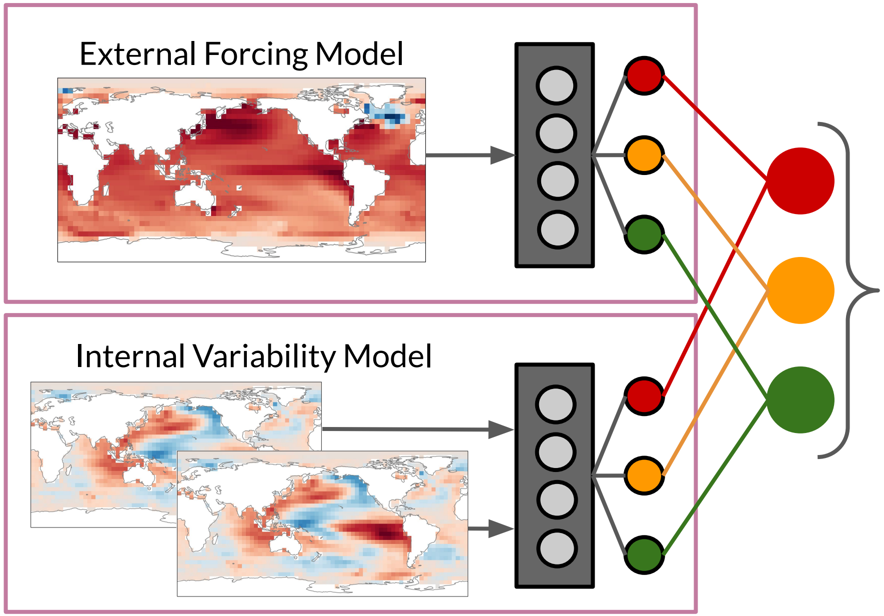
Attributing external + internal sources of predictability

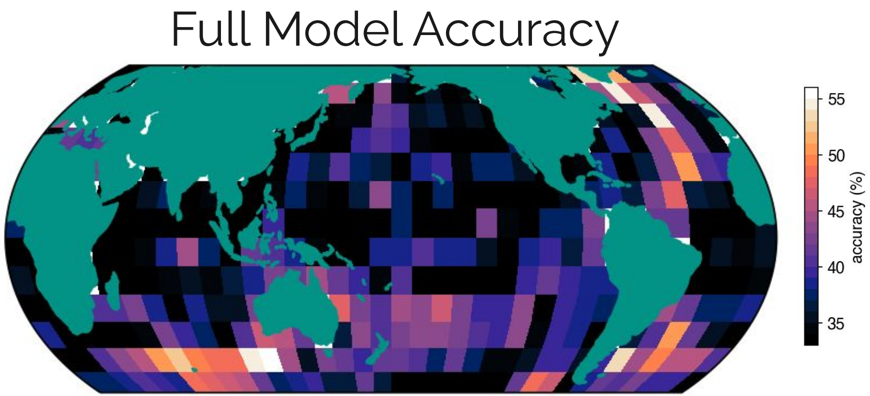
Gordon & Barnes (in prep)



Attributing external + internal sources of predictability

Gordon & Barnes (in prep)

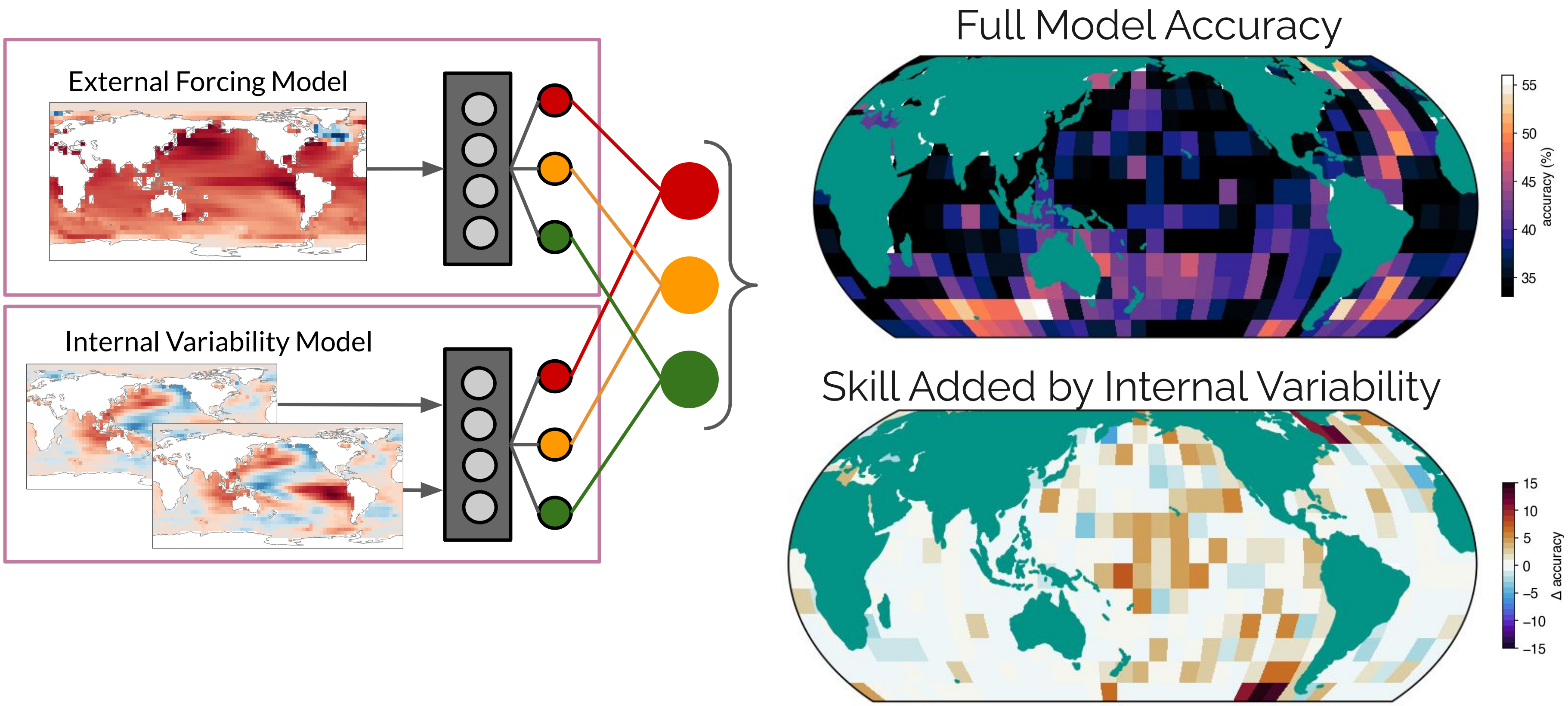




Full Model − External Only Model =

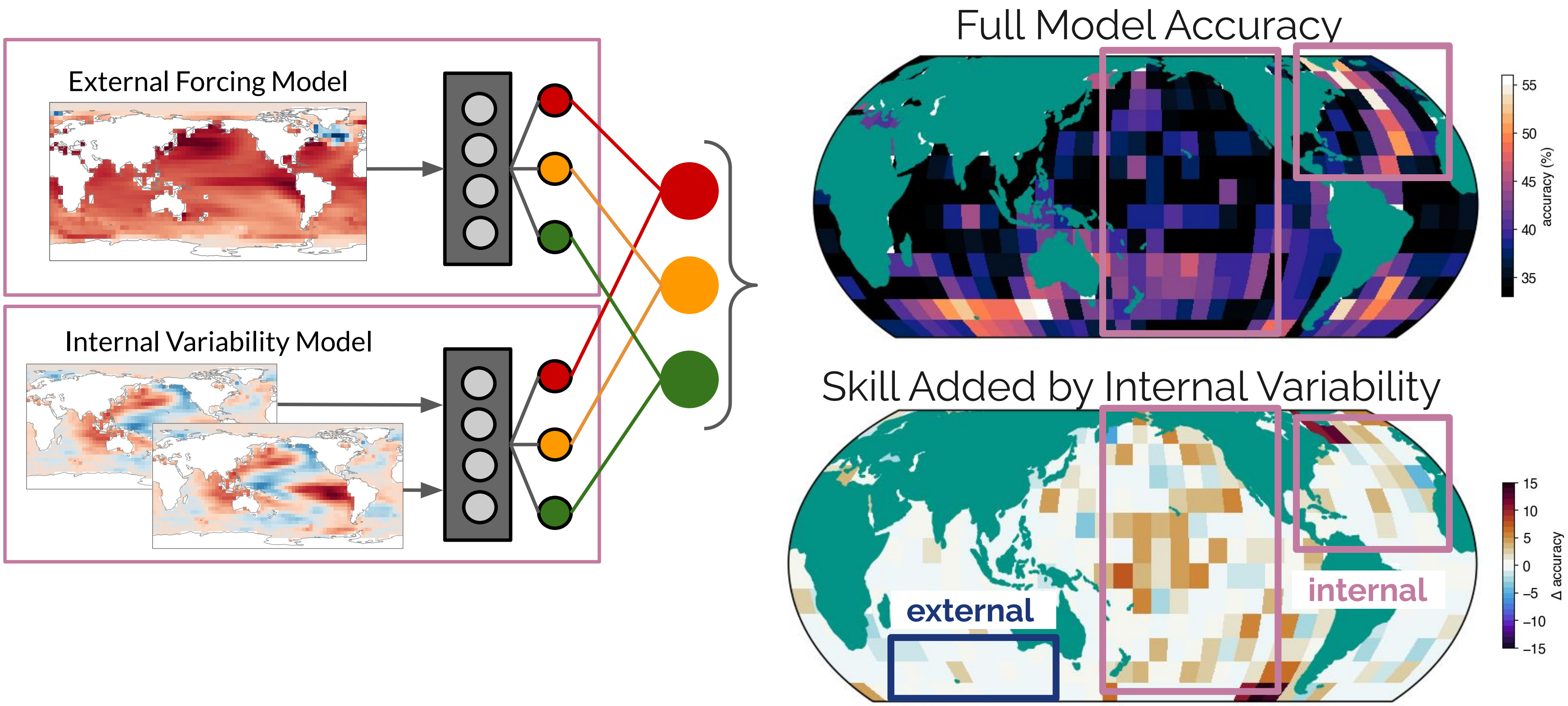
Attributing external + internal sources of predictability

Gordon & Barnes (in prep)



Attributing external + internal sources of predictability

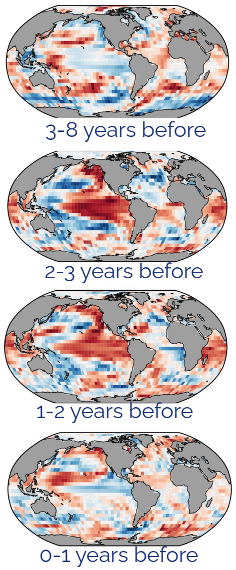
Gordon & Barnes (in prep)

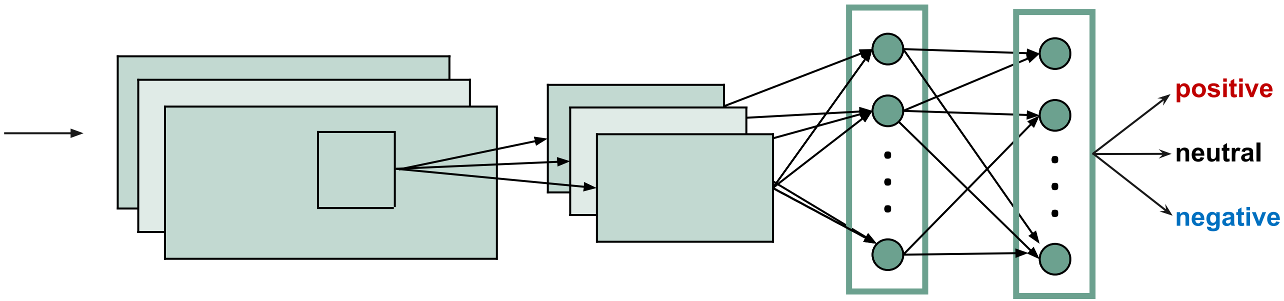


Attributing external + internal sources of predictability

Gordon & Barnes (in prep)

past sea-surface temperatures





future sea surface temperature anomalies*
for one point
[0-5 years]

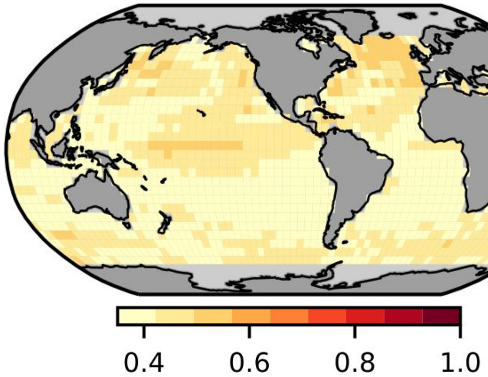
*can predict a range of variables

Predict ocean temperatures 5 years later

Davenport & Barnes (in prep)

CLIMATE MODEL DATA

Overall Accuracy

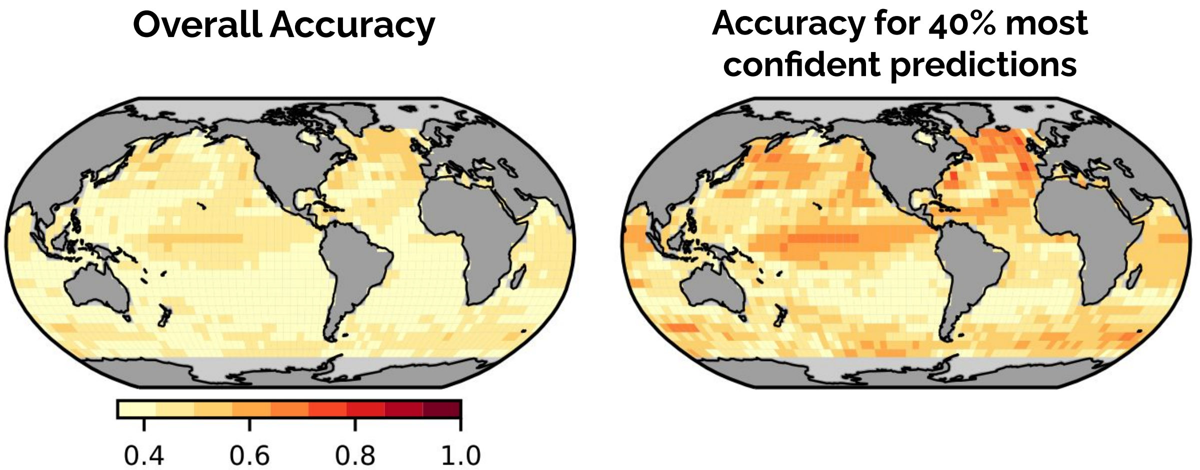


Trained on climate model **MPI-ESM-1-2-LR [3,630 years of data]**
Evaluated on climate model **MPI-ESM-1-2-LR**

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

Davenport & Barnes (in prep)

CLIMATE MODEL DATA



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

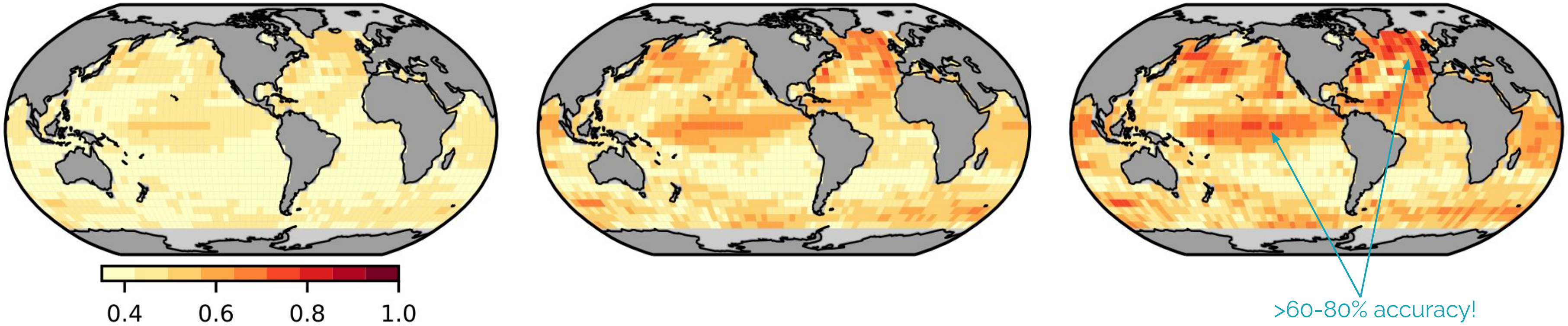
Davenport & Barnes (in prep)

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

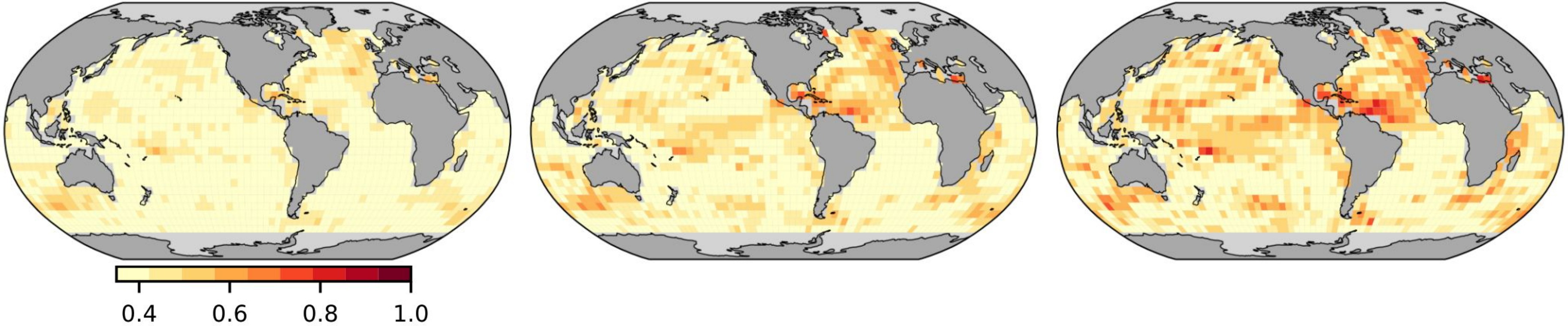
Davenport & Barnes (in prep)

OBSERVATIONS

Overall Accuracy

Accuracy for 40% most confident predictions

Accuracy for 20% most confident predictions

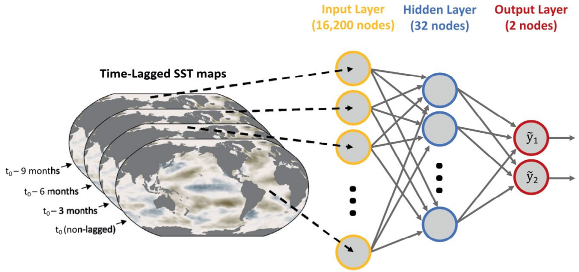


Trained on climate model **MPI-ESM-1-2-LR [3,630 years of data]**
Evaluated on **observations [ERSSTv5; 169 years of data]**

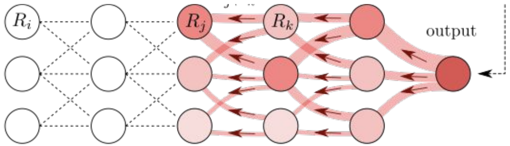
Leveraging climate model data provides accurate predictions of the real world

Davenport & Barnes (in prep)

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

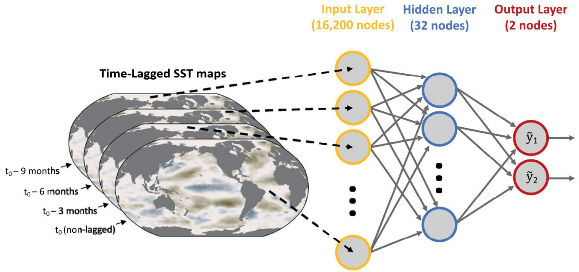


XAI

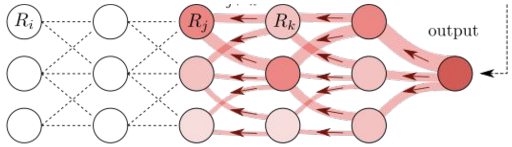
XAI reveals sources of predictability that vary in time and space


Toms, Barnes & Hurrell (2021)

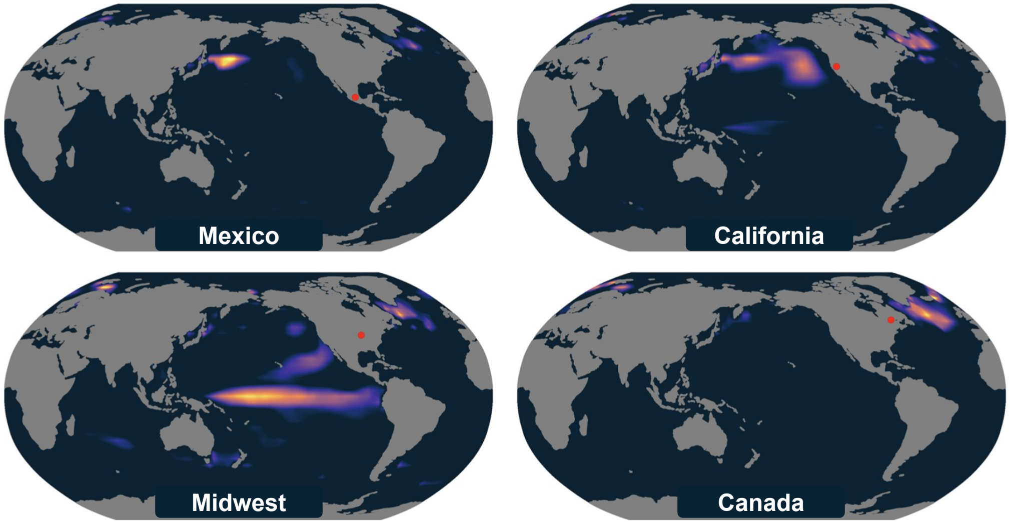
Predicting 5-year average surface temperature at each grid point
Applied to 1200 years of CESM2 control simulation



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



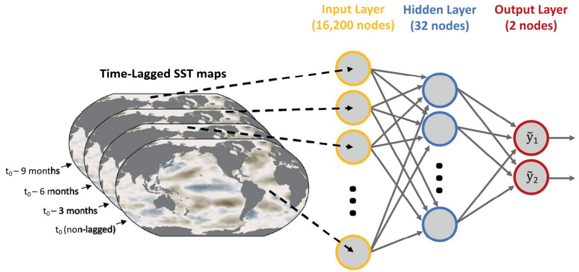
XAI



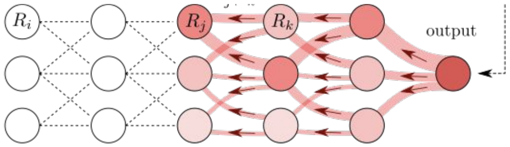
XAI reveals sources of predictability that vary in time and space


Toms, Barnes & Hurrell (2021)

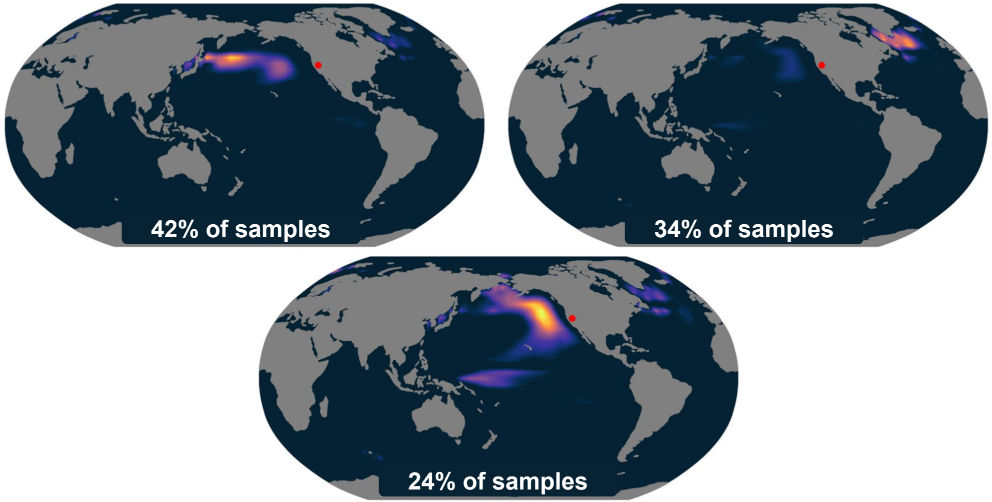
Predicting 5-year average surface temperature at each grid point
Applied to 1200 years of CESM2 control simulation



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



XAI

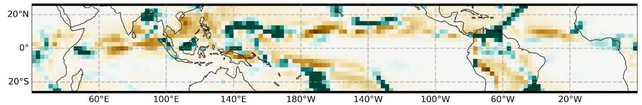


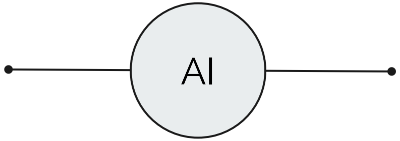
XAI reveals sources of predictability that vary in time and space

Toms, Barnes & Hurrell (2021)

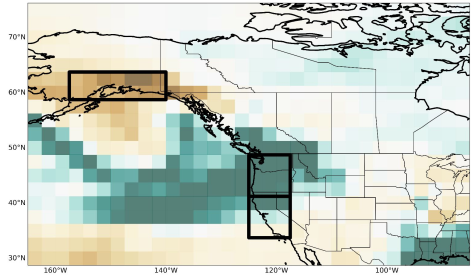
Input: Daily tropical precipitation

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





Output: Precipitation 3-4 weeks later

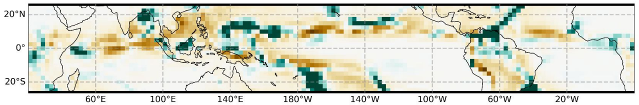


XAI allows us to quantify predictability in past and future climates and assess changes in sources.

Arcodia & Barnes (in prep)

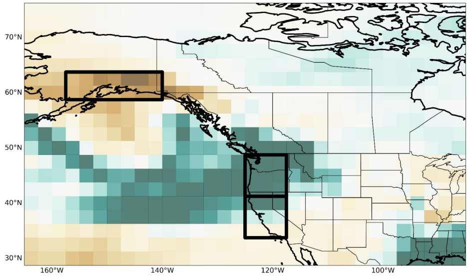
Input: Daily tropical precipitation

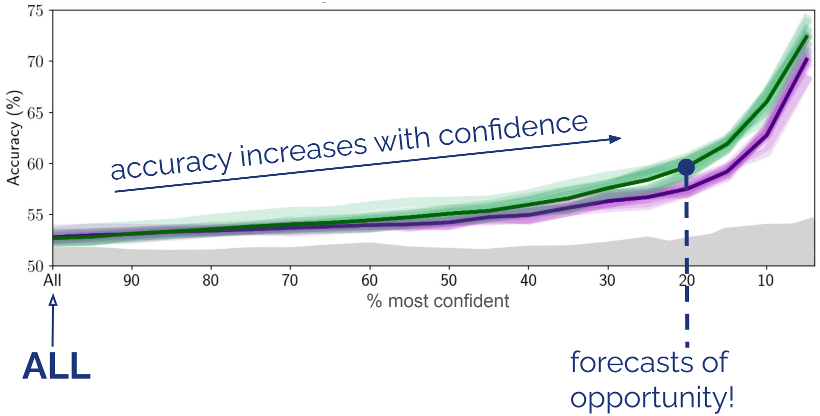
Trained on climate model **CESM2 [800 years of daily data]**



AI

Output: Precipitation 3-4 weeks later



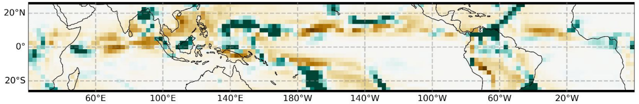


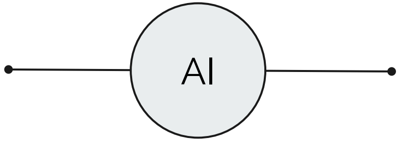
XAI allows us to quantify predictability in past and future climates and assess changes in sources.

Arcodia & Barnes (in prep)

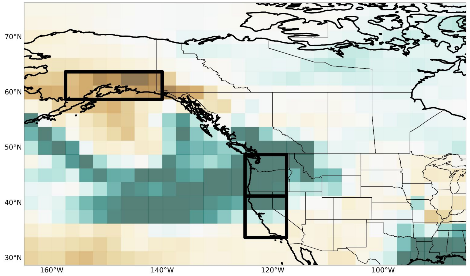
Input: Daily tropical precipitation

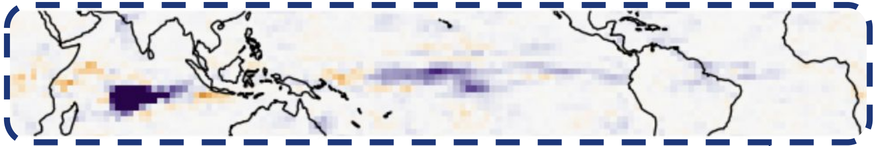
Trained on climate model **CESM2 [800 years of daily data]**

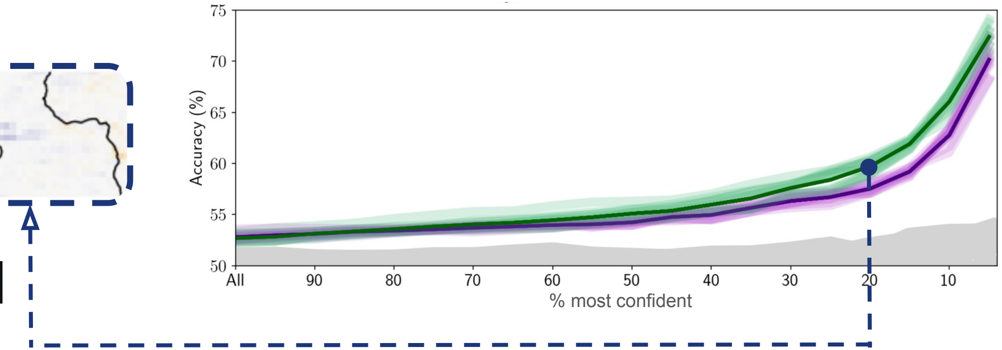




Output: Precipitation 3-4 weeks later







XAI

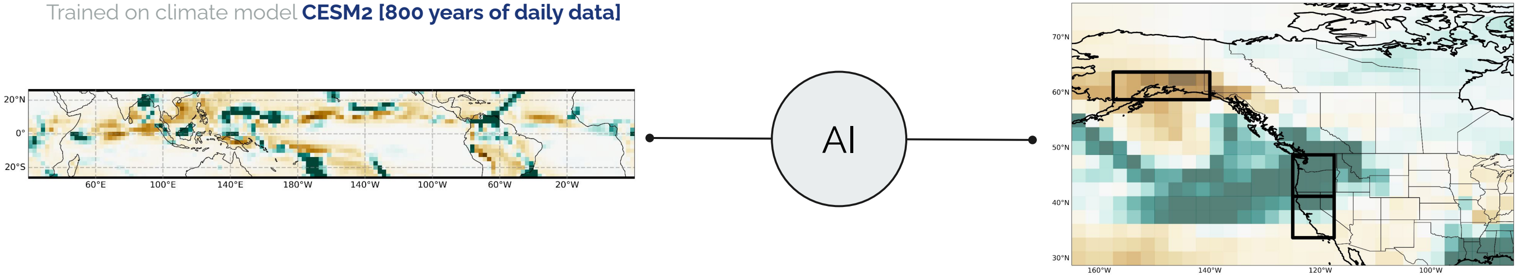
XAI allows us to quantify predictability in past and future climates and assess changes in sources.

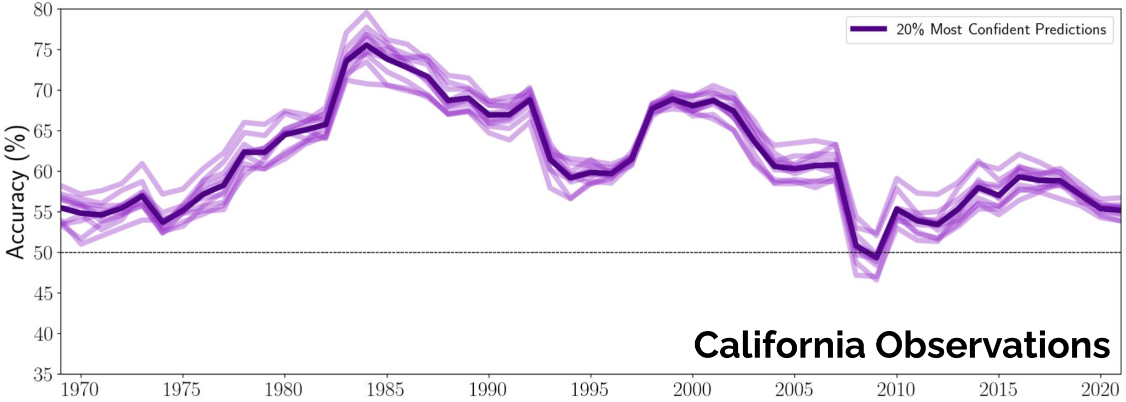
Arcodia & Barnes (in prep)

Input: Daily tropical precipitation

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

Output: Precipitation 3-4 weeks later



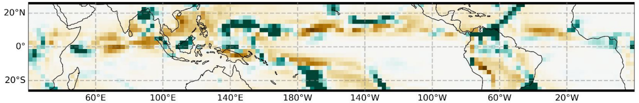


XAI allows us to quantify predictability in past and future climates and assess changes in sources.

Arcodia & Barnes (in prep)

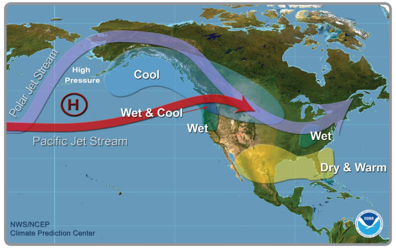
Input: Daily tropical precipitation

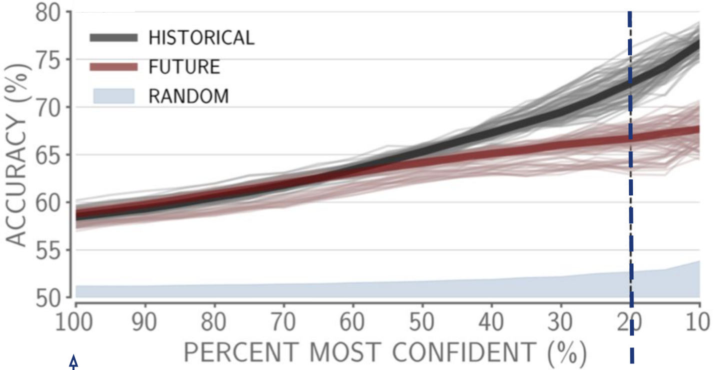
Trained on climate model **CESM2 [800 years of daily data]**



AI

Output: Pacific circulation 3 weeks later





ALL

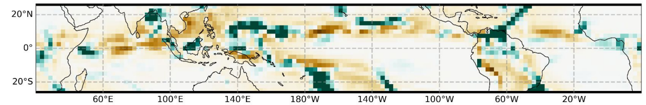
forecasts of opportunity
decrease in skill!

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

Mayer & Barnes (2021, 2022)

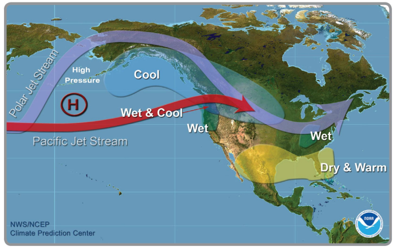
Input: Daily tropical precipitation

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

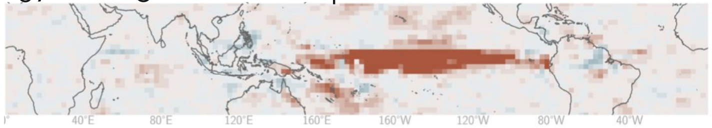


AI

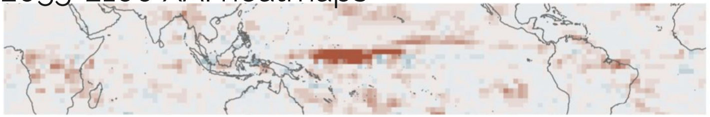
Output: Pacific circulation 3 weeks later

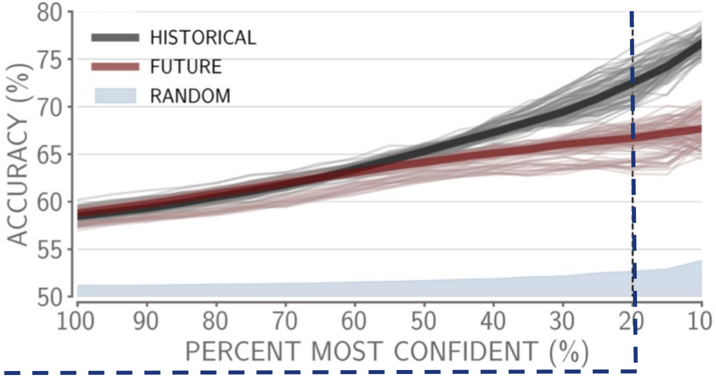


1970-2015 XAI heatmaps



2055-2100 XAI heatmaps

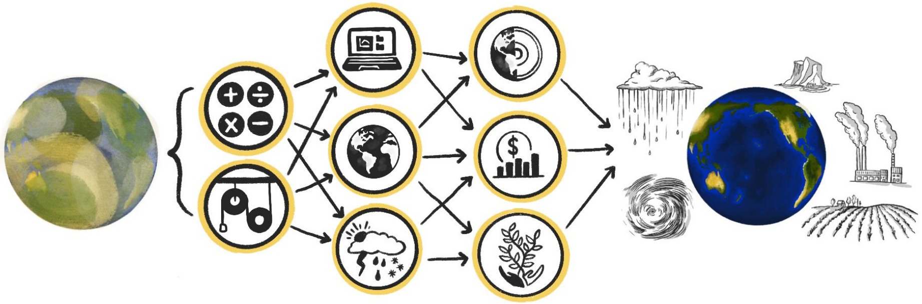




XAI

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

Mayer & Barnes (2021, 2022)



The future of actionable climate predictions requires the **mixing of knowledge**.

And ultimately we want more than just a prediction - we want to know “**why?**”

Explainable AI has a lot to offer climate prediction.

The future of actionable climate predictions requires the **mixing of knowledge**.

And ultimately we want more than just a prediction - we want to know “**why?**”

Thank you.

eabarnes@colostate.edu

Explainable AI has a lot to offer climate prediction.

The future of actionable climate predictions requires the **mixing of knowledge**.

And ultimately we want more than just a prediction - we want to know "**why?**"