



MAPP NA19OAR4310289



CAREER AGS-1749261
HDR OAC-1934668
AI Institute ICER-2019758

Explainable / Interpretable AI for Climate Science

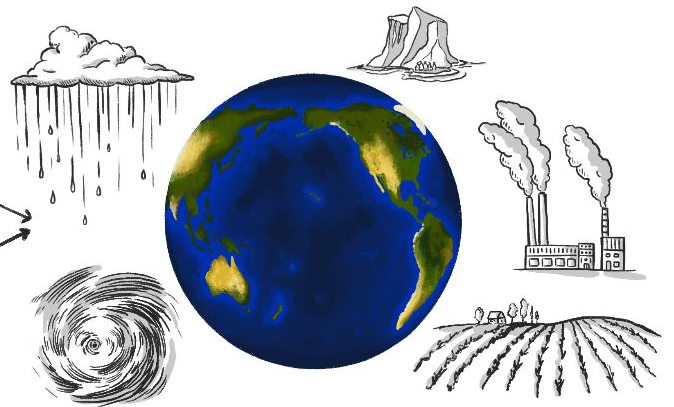
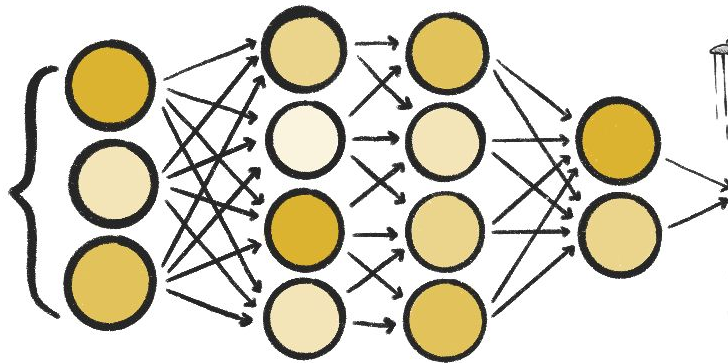
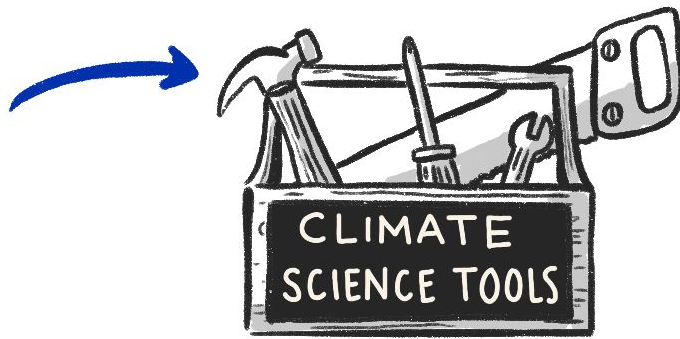


Dr. Elizabeth A. Barnes
Associate Professor
Department of Atmospheric Science
Colorado State University



ATMOSPHERIC SCIENCE
COLORADO STATE UNIVERSITY

MACHINE
LEARNING



ML for Climate Science

The field's interest, and research, has exploded in the past ~3 years!

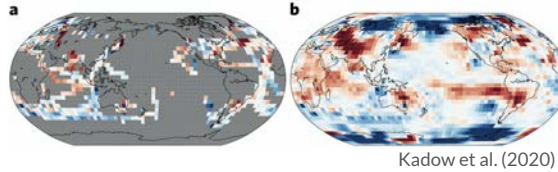
Applications of ML for atmospheric science dates back as far as the 1960's!

Range of applications:

- global weather prediction
- convective & radiative parameterizations
- downscaling
- extreme event detection
- data reconstruction
- weather prediction
- processing of remote sensing data

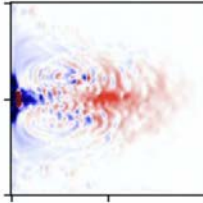
Climate Reconstruction

e.g. Kadow et al. (2020), DelSole and Nedza (2020)



Equation discovery

e.g. Zanna and Bolton (2020)



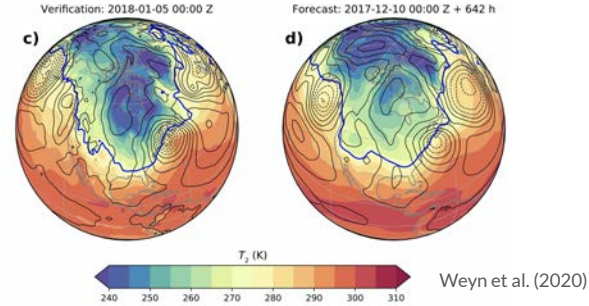
$$\hat{\mathbf{S}}_{\mathbf{u}}^{BT} \approx \kappa_{BT} \nabla \cdot \begin{pmatrix} \zeta^2 - \zeta D & \zeta \tilde{D} \\ \zeta \tilde{D} & \zeta^2 + \zeta D \end{pmatrix}$$

Climate change communication



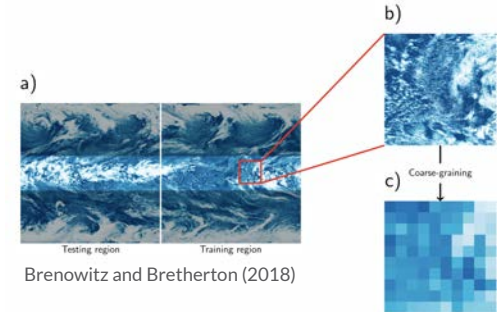
Weather Prediction

e.g. Gagne et al. (2019); Gagne et al. (2017); Chattopadhyay et al. (2019); Lagerquist et al. (2020)



Convective parameterizations

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



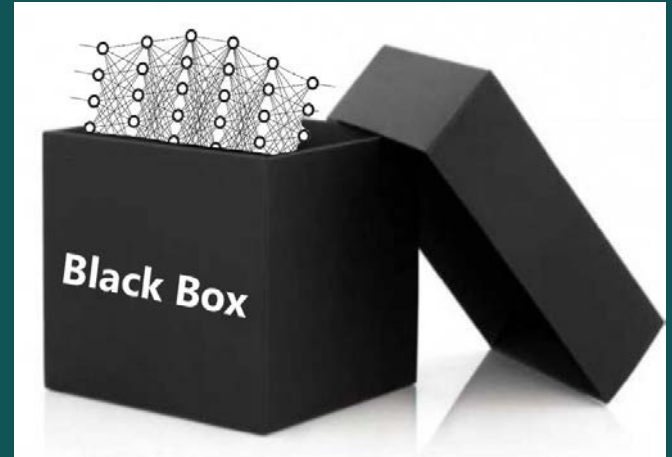
Reasons to use AI for climate science

- **Do it better**
 - e.g. convective parameterizations in models are not perfect, use ML to make them more accurate
- **Do it faster or cheaper**
 - e.g. radiation code in models is very slow - use ML methods to speed things up
- **Learn something new**
 - e.g. go looking for non-linear relationships you didn't know were there

Very relevant for research: may be slower and worse, but can still learn something (more to come on this...)

Opening the “Black Box”

Leveraging advances in
explainable / interpretable AI



Opening the Black Box

In the past few years multiple papers have come out demonstrating the use of ML explainability methods for geoscience

arXiv > physics > arXiv:2103.10005

Search...
Help | Advance

Physics > Geophysics

[Submitted on 18 Mar 2021]

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

Antonios Mamlakis, 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 trust and problem at hand. Many different methods have been introduced in the literature attributing the network's prediction to specific features in the input domain (like MNIST or ImageNet for image classification), or through deletion/insertion of ground truth for the attribution is lacking, making the assessment of problems in geosciences are rare. Here, we provide a framework, based on benchmark datasets for regression problems for which the ground truth of a dataset and train a fully-connected network to learn the underlying function attribution heatmaps from different XAI methods to the ground truth in order to compare. We believe that attribution benchmarks as the ones introduced here in the geosciences, and for accurate implementation of XAI methods, which

arXiv > physics > arXiv:2202.03407

Search...
Help | Advance

Physics > Geophysics

[Submitted on 7 Feb 2022]

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

Antonios Mamlakis, 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 inter-comparison 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.

MAKING THE BLACK BOX MORE TRANSPARENT

Understanding the Physical Implications of Machine Learning

AMY MCGOVERN, RYAN LAGERQUIST, DAVID JOHN GAGNE II, G. EU JENSEN, KIMBERLY L. ELMORE, CAMERON R. HOEYER, AND TRAVIS SMITH

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

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 (Rakhtin et al. 2018), and galaxy

classification (Dieleman et al. 2015). Simple forms of ML (e.g., linear regression) have been around since at least the 1950s (Malo ML has been used extensively to forecast hazards since the mid-1990s. Kitzmann use linear regression to forecast the tornadoes, large hail, or damaging wind (1997) use linear regression to forecast ity and size. Marshall and Stumpf (1999) use neural networks to forecast the probability of tornado and damaging wind, respectively, and Wirt (2001) use neural networks to forecast hail probability at 1-day lead time et al. (2014) and Williams (2014) use r to forecast convectively induced airca while Cintino et al. (2014, 2018) us to forecast the probability of tornado and damaging wind. DL is also being used in meteorology, with applications in hail prediction (Gagne et al. 2019) an extreme weather patterns such as trop

AFFILIATIONS: McGovern and Jensen—University of Oklahoma, Norman, Oklahoma, mcgov@ou.edu; Lagerquist for Mesoscale Meteorological Studies, and University of Oklahoma, Norman, Oklahoma, gagne@ou.edu; Gagne—National Center for Atmospheric Research, Boulder, Colorado, gagne@ucar.edu; Smith—Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma, smith@ou.edu; Ebert-Uphoff, Barnes, and NOAA/National Severe Storms Laboratory, Norman, Oklahoma, ebert@ou.edu; Elmore—School of Meteorology, University of Oklahoma, Norman, Oklahoma, elmor@ou.edu; Hoefer, hoefer@ou.edu; and Travis Smith, smith@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-18-0195.1).
In final form 20 June 2019
Manuscript received 18 March 2018, in final form 18 June 2019

Abstract
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 (Rakhtin et al. 2018), and galaxy classification (Dieleman et al. 2015). Simple forms of ML (e.g., linear regression) have been around since at least the 1950s (Malo ML has been used extensively to forecast hazards since the mid-1990s. Kitzmann use linear regression to forecast the tornadoes, large hail, or damaging wind (1997) use linear regression to forecast ity and size. Marshall and Stumpf (1999) use neural networks to forecast the probability of tornado and damaging wind, respectively, and Wirt (2001) use neural networks to forecast hail probability at 1-day lead time et al. (2014) and Williams (2014) use r to forecast convectively induced airca while Cintino et al. (2014, 2018) us to forecast the probability of tornado and damaging wind. DL is also being used in meteorology, with applications in hail prediction (Gagne et al. 2019) an extreme weather patterns such as trop

atmospheric rivers, and synoptic-scale weather patterns (Gagne et al. 2019, Kunk et al. 2019b). The authors

Search...
Help | Advance

NOVEMBER 2019

15 NOVEMBER 2019

15 NOVEMBER 2019

15 NOVEMBER 2019

15 NOVEMBER 2019

15 NOVEMBER 2019

15 NOVEMBER 2019

AGU ADVANCING EARTH AND SPACE SCIENCE

JAMES Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS001620

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

Key Points:
• Interpretable neural networks can identify the coherent spatial patterns of known modes of Earth system variability.
• The Bayesian ensemble 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 criterion of a neural network.

Supporting Information:
• Supporting Information S1

Correspondence to:
A. M. McGovern, mcgov@ou.edu

Citation:
McGovern, A. M., Barnes, E. A., & Ebert-Uphoff, I. (2019). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling Earth Systems*, 11, e2019MS001620. <https://doi.org/10.1029/2019MS001620>

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 layer-wise 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 isolate scientifically meaningful information from neural networks to 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.

BAMS ISSUES EARLY ONLINE RELEASE COLLECTIONS FOR AUTHORS

Article Contents

Abstract
Caption
Footnotes

RESEARCH ARTICLE | 31 AUGUST 2020

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

Imme Ebert-Uphoff, Antonios Mamlakis, Elizabeth A. Barnes, and Travis Smith

15 AUGUST 2020

10.1175/BAMS-D-20-0097.1

Split-Screen PDF Share Cite Get Permissions

Applications across all areas of geoscientific research (e.g., machine learning, 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 layer-wise 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 isolate scientifically meaningful information from neural networks to 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.

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.

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.

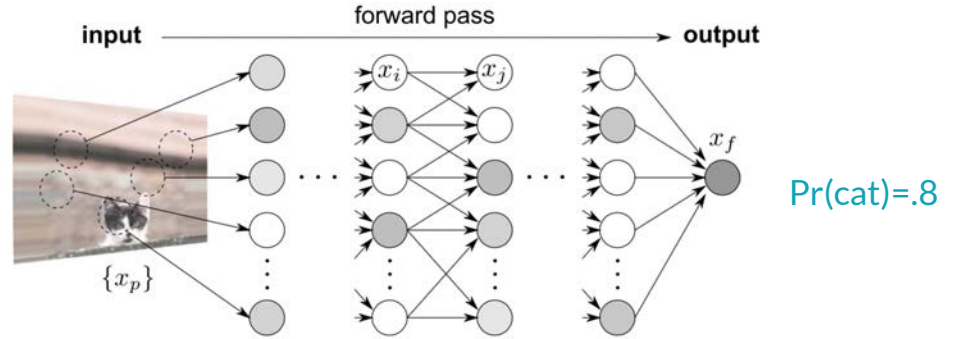


XAI Attribution Methods

Attribution methods produce a heatmap of the most relevant regions of the input for each prediction

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

Prediction
of 1 sample

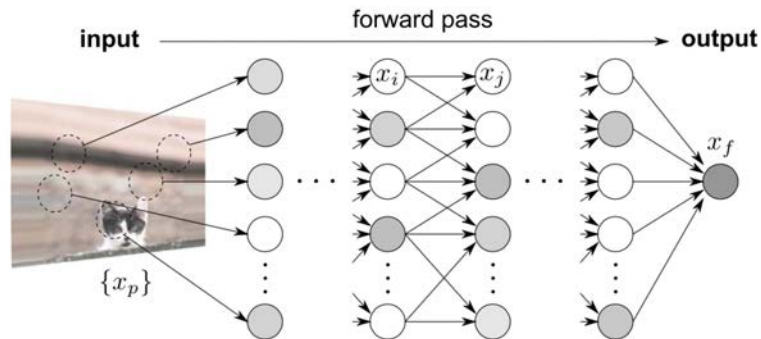


XAI Attribution Methods

Attribution methods produce a heatmap of the most relevant regions of the input for each prediction

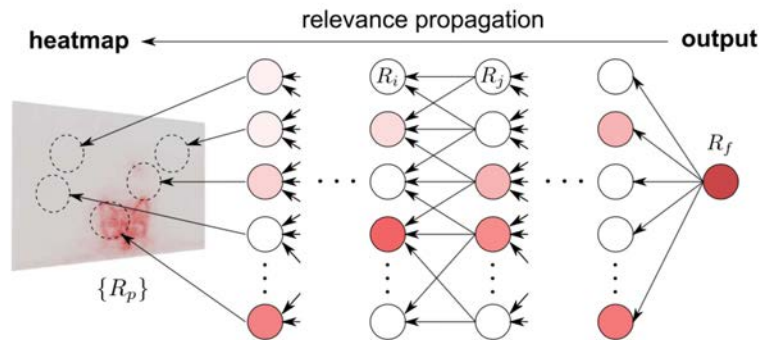
Attribution heatmaps are largely consistent with how many climate scientists pose questions

Prediction
of 1 sample



$\Pr(\text{cat})=.8$

Attribution
of 1 sample



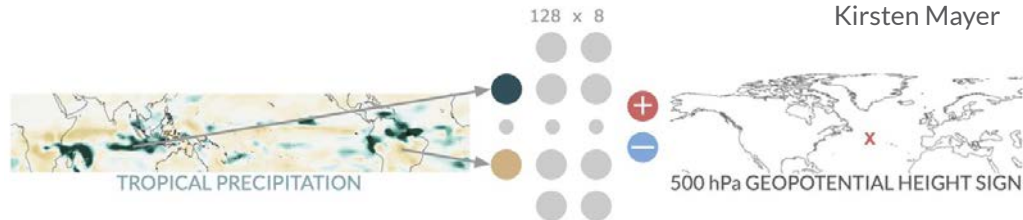
$\Pr(\text{cat})=.8$

XAI Attribution Methods

Attribution methods produce a heatmap of the most relevant regions of the input for each prediction

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

Prediction
of 1 sample

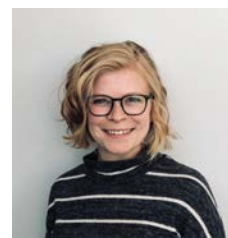


Kirsten Mayer

XAI Attribution Methods

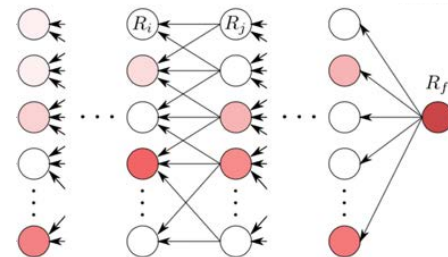
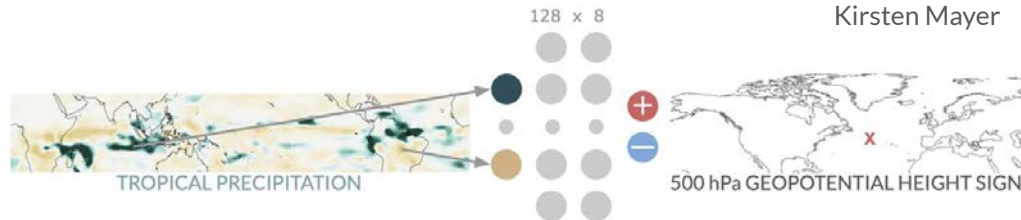
Attribution methods produce a heatmap of the most relevant regions of the input for each prediction

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



Kirsten Mayer

Prediction
of 1 sample



XAI

XAI Attribution Methods

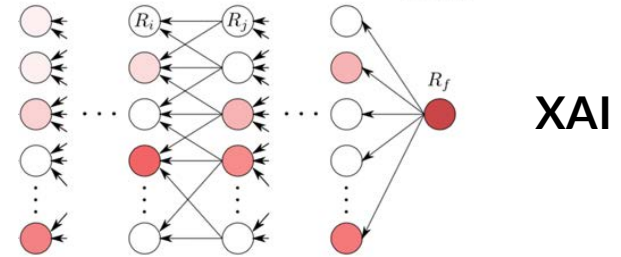
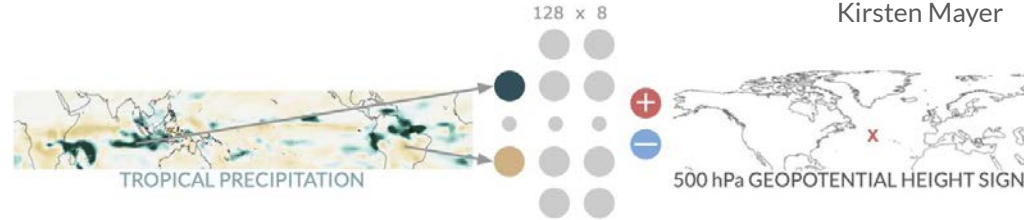
Attribution methods produce a heatmap of the most relevant regions of the input for each prediction

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

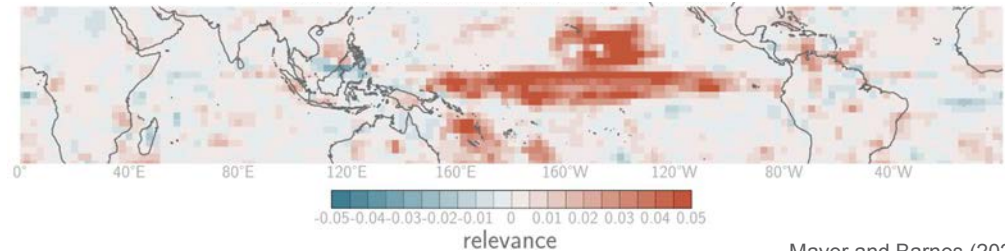


Kirsten Mayer

Prediction
of 1 sample



Attribution
of 1 sample

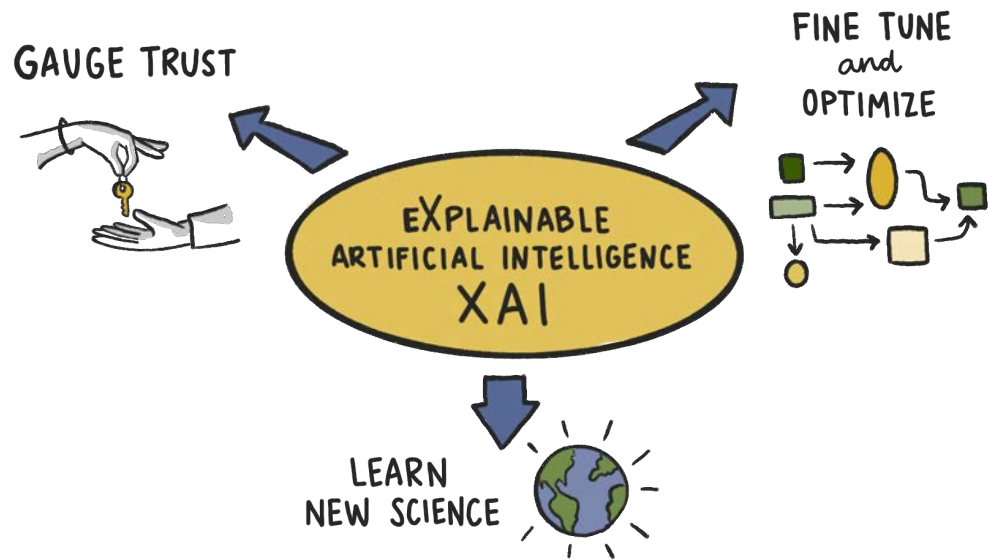


Why you should care about XAI

As scientists our ultimate goal is to understand “why?”. But even if you don’t care “why” you should still care about XAI.

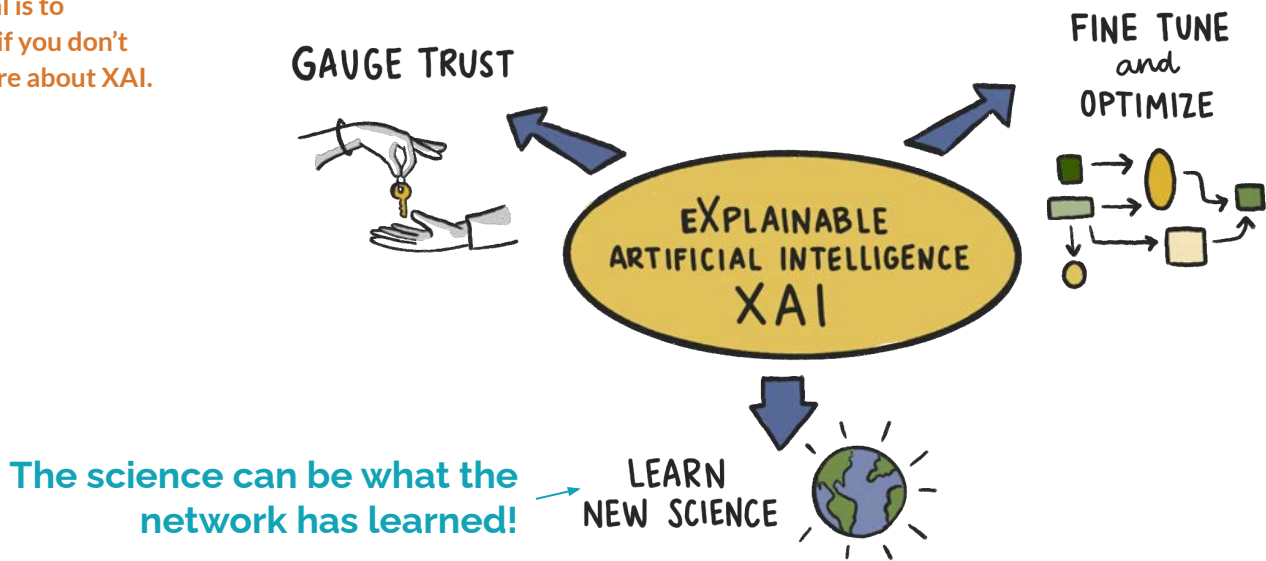
Why you should care about XAI

As scientists our ultimate goal is to understand “why?”. But even if you don’t care “why” you should still care about XAI.



Why you should care about XAI

As scientists our ultimate goal is to understand “why?”. But even if you don’t care “why” you should still care about XAI.



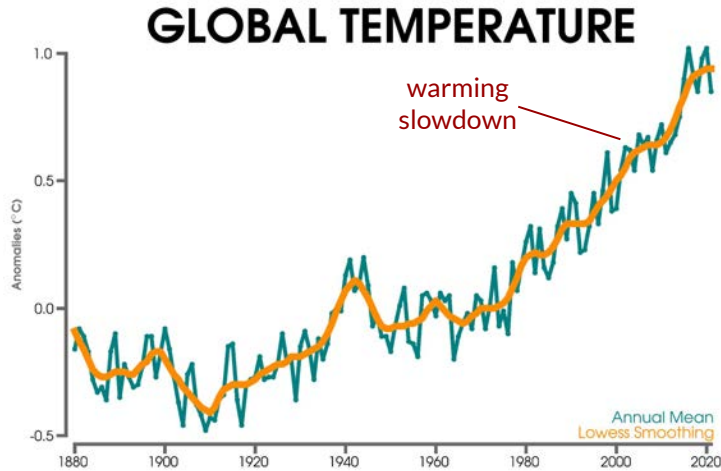


Dr. Zack Labe

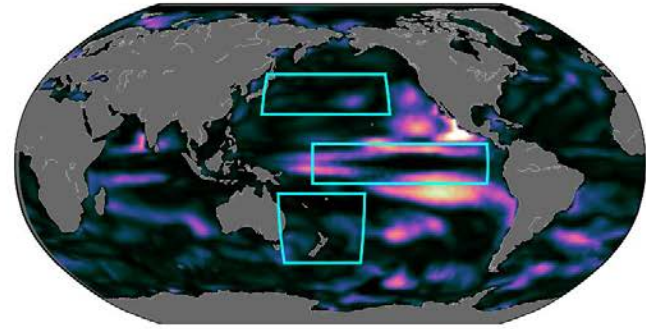
XAI for climate change & variability

Train a neural network to predict temporary slowdowns in global mean surface temp.

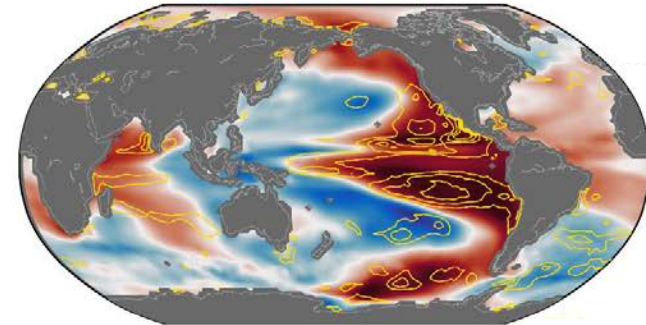
- ANN learns patterns of upper ocean heat content associated with decadal slowdowns in both climate model data and observations
- XAI reveals the ANN is learning off-equatorial patterns of anomalous ocean heat content that resemble transitions in the phase of the Interdecadal Pacific Oscillation



CORRECT SLOWDOWN PREDICTIONS



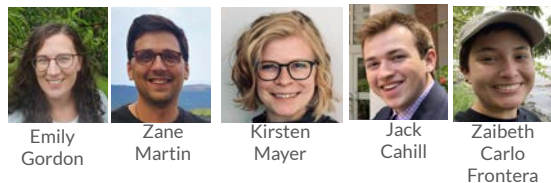
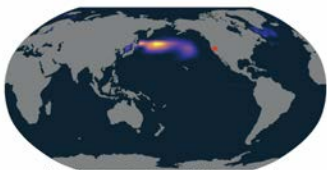
XAI



ocean heat content
(0-100 m)

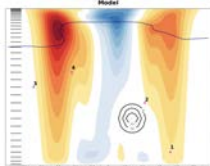
Other example uses of XAI @ CSU (only a subset)

Subseasonal-to-decadal predictability



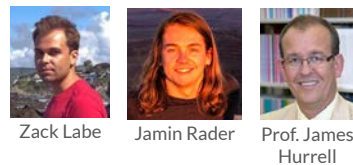
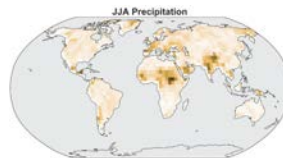
Exploring subseasonal-to-decadal climate dynamics with implications for prediction, scientific mechanisms, and basic theory

Forced response of midlatitude dynamics



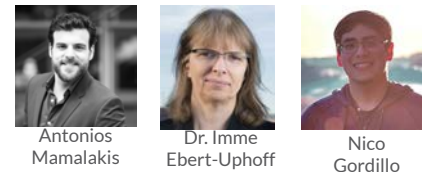
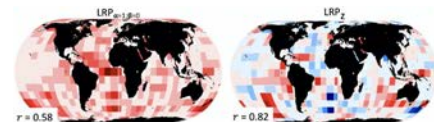
Understanding basic general circulation responses to climate forcings *[figure from Baker et al. 2017]*

Indicator patterns of forced change



Learn non-linear, time-evolving patterns of forced change in climate simulations and observations

XAI benchmarking & robust predictions



Develop robust AI methods and benchmarks for XAI method evaluation and comparison

Exciting Frontiers

#1 Knowledge-guided machine learning

Continue fusing scientific knowledge and
AI for climate science

- the availability of extensive existing knowledge
- the desire of Earth scientists to gain scientific insights rather than just “get numbers” from an algorithm
- the high complexity of the Earth system
- the limited sample size and lack of reliable labels in many Earth science applications
- improves transparency and trustworthiness

Make physics and ML work together



<https://www.pxfuel.com/en/free-photo-oadru>

#2 Transfer learning

Leverage imperfect climate model output through a transfer learning framework

For many earth science applications we have very little observational data

...BUT...

We have thousands of years of imperfect earth system model simulations from which ML tools can learn.

Step 1: Train ML model with climate model simulations

Step 2: Update the ML model with data from observations

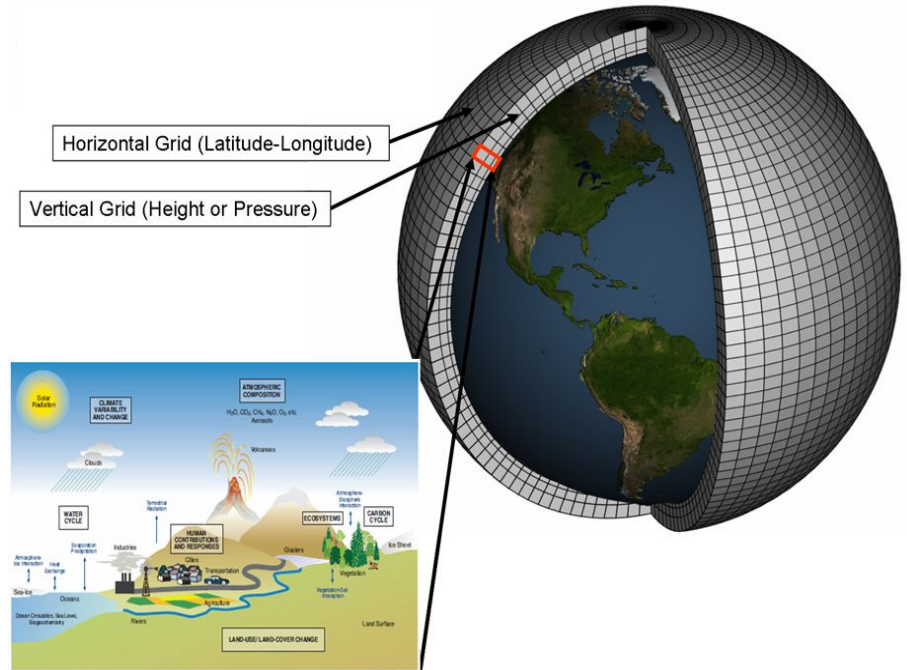


End Result: A trained prediction model that leverages dynamical simulations but applies better to the real world

#3 Improve climate projections

Bring ML methods into the building, evaluation and use of climate model projections

- There is great promise for improving climate models through ML-developed convective/radiative parameterizations
- ML for model comparisons and evaluation against observations.
- ML to explore bias correcting / transforming climate model projections to narrow uncertainties



Climate science requires the **mixing of knowledge** from many fields. And ultimately we want more than just a prediction - we want to know **“why?”**



Explainable / interpretable ML is a game changer for climate research.