Using Grouped Features to Improve Explainable AI Results for Atmospheric AI Models that use Gridded Spatial Data and Complex Machine Learning Technique

Evan Krell[1], Philippe Tissot[1], Antonios Mamalakis[1], Waylon Collins[1], Imme Ebert-Uphoff[1], and Scott King[1]

[1]Affiliation not available

February 15, 2024

Abstract

Atmospheric AI modeling is increasingly reliant on complex machine learning (ML) techniques and high-dimensional gridded inputs to develop models that achieve high predictive skill. Complex deep learning architectures such as convolutional neural networks and transformers are trained to model highly non-linear atmospheric phenomena such as coastal fog [1], tornadoes [2], and severe hail [3]. The input data is typically in the form of gridded spatial data composed of multiple channels of satellite imagery, numerical weather prediction output, reanalysis products, etc. In many studies, the use of complex architectures and high-dimensional inputs were shown to substantially outperform simpler alternatives.

A major challenge when using complex ML techniques is that it is very difficult to understand how the trained model works. The complexity of the model obfuscates the relationship between the input and prediction. It is often of interest to understand a model's decision-making process. By exposing the model's behavior, users could verify that the model has learned physically realistic predictive patterns. This information can be used to calibrate trust in the model. The model may have also learned novel patterns within the data that could be used to gain new insights into the atmospheric process. Extracting learned patterns could be used to generate hypotheses for scientific discovery. The rapid adoption of complex ML models and the need to understand how they work has led to the development of a broad class of techniques called eXplainable Artificial Intelligence (XAI). These methods probe the models in various ways to reveal insights into how they work.

Correlations among input features can make it challenging to produce meaningful explanations. The gridded spatial data common in atmospheric modeling applications typically have extensive correlation. Spatial autocorrelation is present among the cells of each spatial grid, but autocorrelation may exist across the gridded data volume due to spatial or temporal relationships between adjacent channels. In addition, there may be correlations between distant locations due to teleconnections between them.

Correlated input features may cause high variance among the trained models. If grid cells are highly correlated, then the target function that the network is attempting to learn is ill-defined and an infinite number of models can be generated that achieve approximately equal performance. Even assuming a perfect XAI method exists, the attribution reflects only the patterns learned for a given model. It is arbitrary which of the correlated features are used by a given model. This can lead to a misleading understanding of the actual relationship between the input features and target.

A potential solution is to group the correlated features before applying XAI. Attribution can be assigned to each group rather than to individual cells. In this case, all the correlated cells will be permuted at the same time to analyze their collective impact on the output. The purpose is to reveal the contribution of each group of related cells toward the model output. Ideally, the explanations are insensitive to the random choice among correlated features learned by the model. Without grouping, the user can be misled to consider a feature as not being related to the target because of the presence of correlated features. With grouping, the explanations should better reveal the learned patterns.

Grouping features based on correlation can be challenging. The correlation rarely equals one and the strength of the correlation influences the variance among trained models. Calculating the correlation can be difficult because of partial correlations and fuzzy, continuous boundaries. The choice of groups can greatly influence the explanations. Another challenge is that it is not straight-forward to assess the quantitative accuracy of an XAI technique. This is because there is rarely a ground truth explanation to compare to. If we knew the attribution, we would not need XAI methods.

Synthetic benchmarks for analyzing XAI have been proposed as a solution [4]. It is possible to define a non-linear function such that the contribution of each grid cell's value to the function output can be derived. This attribution map represents the

ground truth for comparison the the output of XAI methods that are applied to a model that very closely approximates the hand-crafted function. In this research, we develop a set of benchmarks to investigate the influence of correlated features on the variation in XAI outputs for a set of trained models. We then explore how features can be grouped to reduce the explanation variance so that users have improved insight into the learned patterns.

First, we create a set of very simple mathematical demonstrations that precisely demonstrate the influence of correlated features and how grouping features provides a solution. Using insights from these experiments, we develop a tool for detecting when correlated features are likely to cause misleading explanations. We then create a set of more realistic benchmarks that are based on atmospheric modeling problems such as sea surface temperature and coastal fog prediction. By defining benchmarks with known ground truth explanations, we can analyze various techniques for grouping the grid cells based on their correlations. Based on our findings, we offer recommendations for strategies to group correlated data so that users can better leverage XAI results toward model development and scientific insights.

[1] Kamangir, H., Collins, W., Tissot, P., King, S. A., Dinh, H. T. H., Durham, N., & Rizzo, J. (2021). FogNet: A multiscale 3D CNN with double-branch dense block and attention mechanism for fog prediction. Machine Learning with Applications, 5, 100038.

[2] Lagerquist, R. (2020). Using Deep Learning to Improve Prediction and Understanding of High-impact Weather.

[3] Gagne II, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. Monthly Weather Review, 147(8), 2827-2845.

[4] Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2022). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. Environmental Data Science, 1, e8.

Abstract content goes here

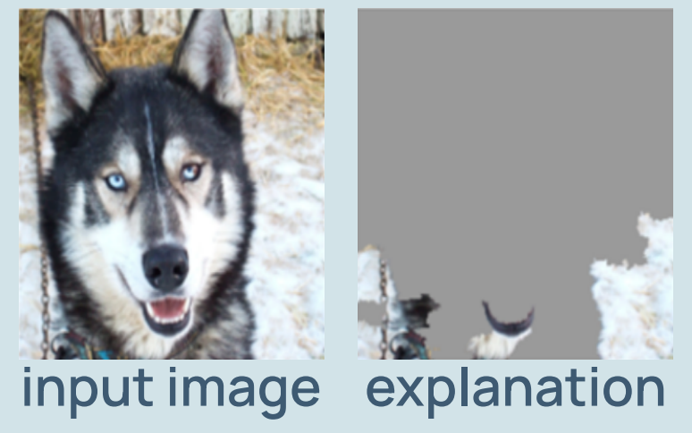
Using Grouped Features to Improve Explainable AI Results for Atmospheric AI Models that use Gridded Spatial Data and Complex Machine Learning Techniques

Evan Krell, Hamid Kamangir, Waylon G. Collins, Scott A. King, Philippe Tissot, Antonios Mamalakis & Imme Ebert-Uphoff

Motivation: Explain Geoscience Models

Explainable AI

[1] Model debugging:



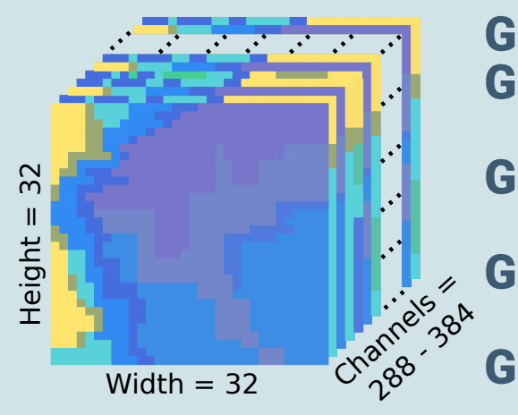
The model has high accuracy for task *wolf or husky?*, but actually looking at snow pixels... many wolf photos have a snowy background.

Scientific insights:
If the model performs well, has it learned something interesting?

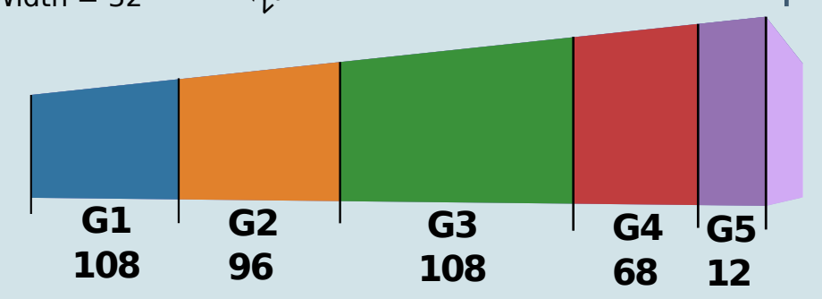
Challenge:
XAI techniques struggle with correlated features

FogNet Model

3D CNN for coastal fog prediction [2]



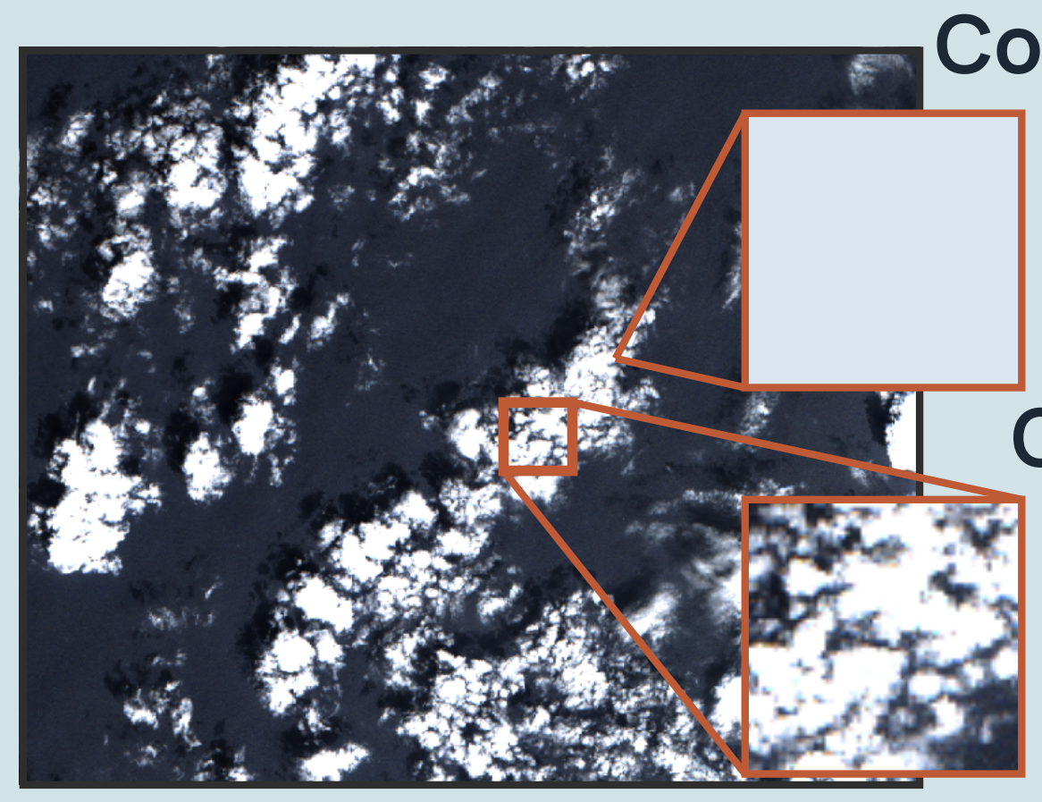
G1 wind
G2 turbulence kinetic energy & humidity
G3 lower atmosphere thermodynamic profile
G4 surface atmosphere moisture & microphysics
G5 sea surface temperature



Challenge:
Gridded spatial data typically has substantial correlation

Challenge: Correlations in Spatial Data

Autocorrelation



Consider evaluating individual pixels:
Expect minimal change in output → so the model uses nothing?

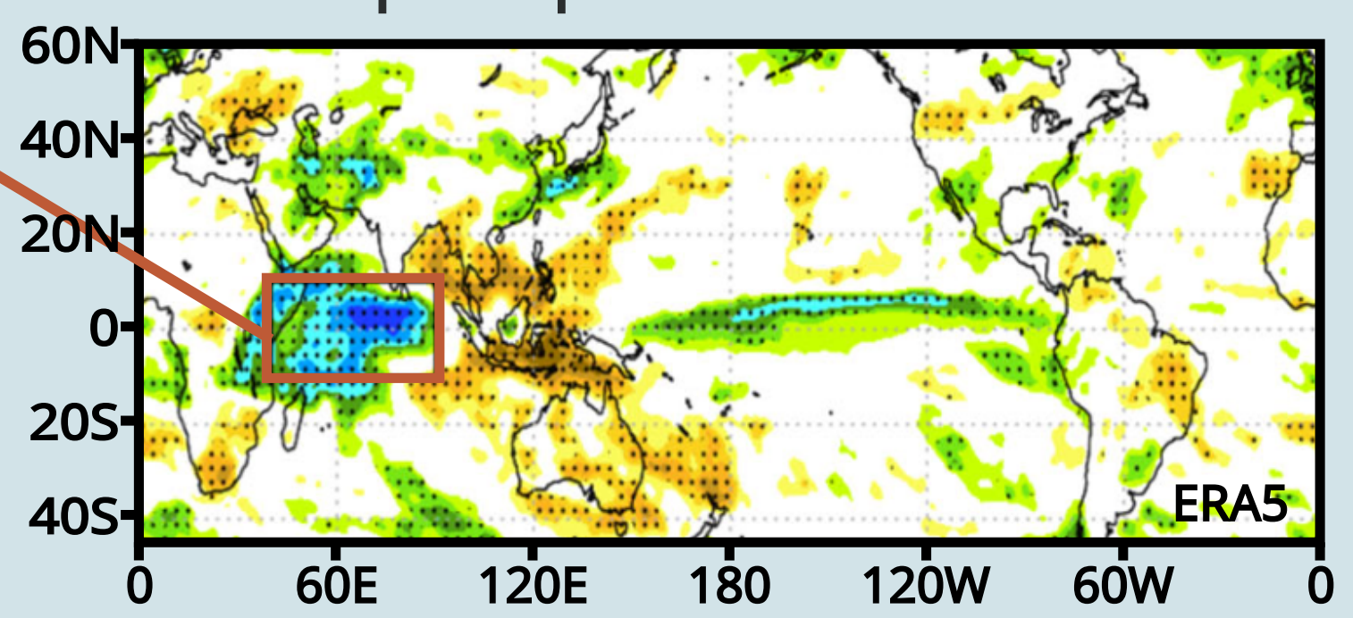
Consider evaluating a superpixel:
Captures a cloud feature that might trigger a change in probability

[3]

Teleconnections

[4] Western-Central Indian Ocean precipitation anomalies

The map shows how global precipitation anomalies are correlated with this region of the earth. Teleconnections are long-range relationships among spatial phenomena.



Long-range dependencies:
There are correlations between grid cells that could be captured by calculating pairwise dependency using a large dataset

Grouping Correlated Features for XAI

Data Relationships

$$\underbrace{x_1 \quad x_2 = x_1}_{\text{complete correlation}} \quad x_3 \quad x_4$$

X

x_1	x_2
x_3	x_4

Actual Function
$y = 0.25*x_1 + 0*x_2 + x_3 + x_4$

XAI from 3 learned functions

data sample

2	4
12	3

xai(y_1)

0.5	0
12	3

xai(y_2)

0	0.5
12	3

xai(y_3)

0.25	0.25
12	3

Some Valid Learned Functions
$y_1 = 0.25*x_1 + 0*x_2 + x_3 + x_4$
$y_2 = 0*x_1 + 0.125*x_2 + x_3 + x_4$
$y_3 = 0.125*x_1 + 0.0625*x_2 + x_3 + x_4$

Grouped XAI Results

grouped sample

2	4
12	3

xai(y_1)

0.5	
12	3

xai(y_2)

0.5	
12	3

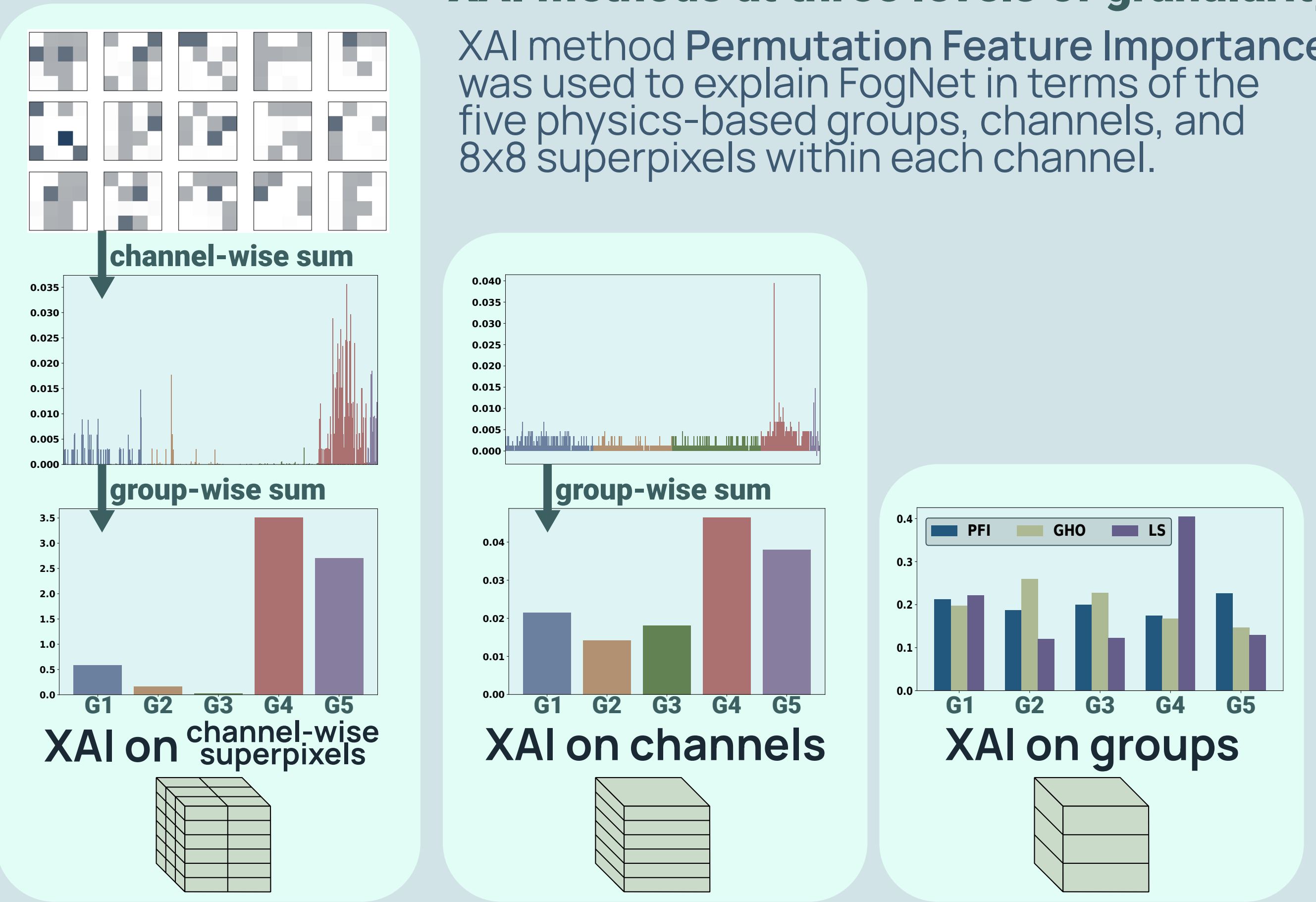
xai(y_3)

0.5	
12	3

Case Study: Explaining FogNet

XAI methods at three levels of granularity

XAI method **Permutation Feature Importance** was used to explain FogNet in terms of the five physics-based groups, channels, and 8x8 superpixels within each channel.



Observation 1:
Explanations are highly sensative to choice of grouping scheme.
Groups suggests that G3 provided ~20% of the predictive skill.
but **Channel-wise superpixels** suggests we could throw G3 out.

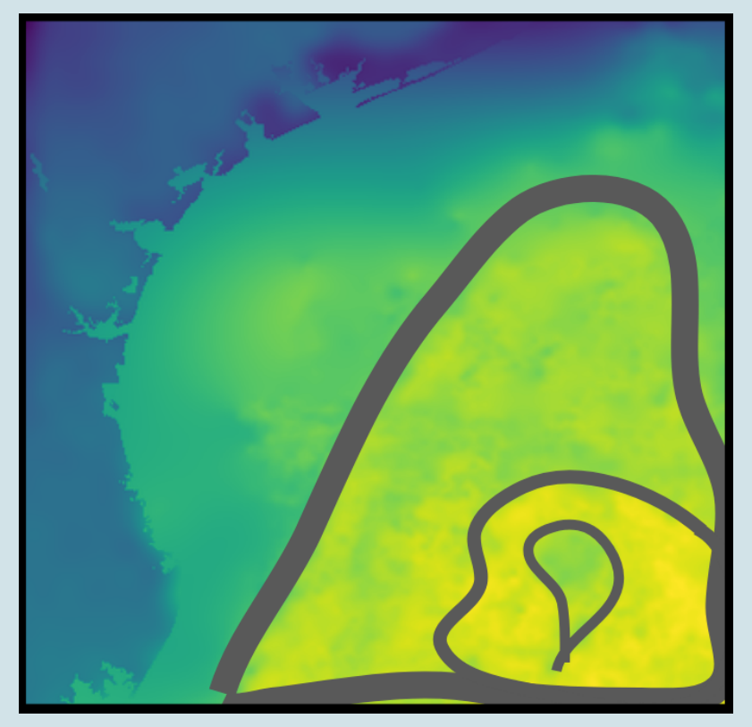
Observation 2:
These disagreements seem to reflect the nature of the data.
G3 contains a 3D atmospheric profile, so small-scale perturbations do not break the large-scale patterns learned using dilated 3D convolution.

Proposal: Hierarchical Clustering

Goal 1: Group features in a data-driven fashion, not arbitrary geometry.
Goal 2: Explain a hierarchy to learn about features across scales.
Goal 3: Strategically select groups since infeasible to explain everything.

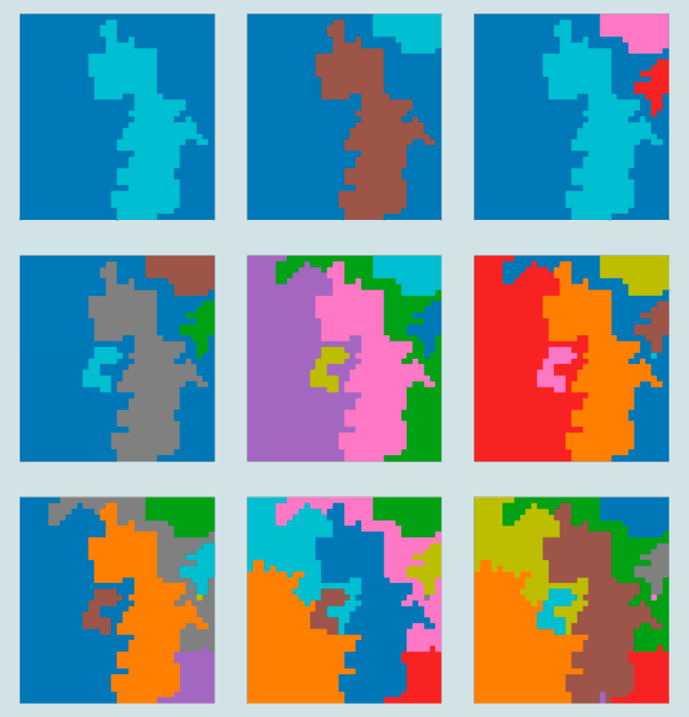
Sketch: nested clusters to capture important features at multiple scales

[5]



Technique: agglomerative hierarchical clustering

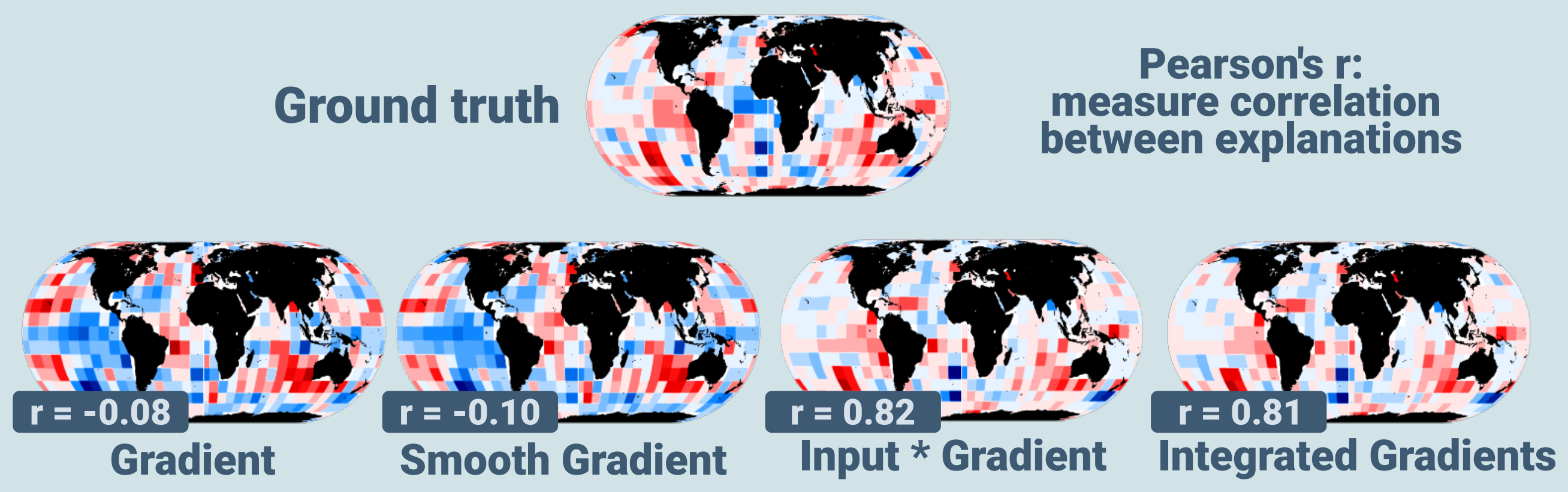
[6]



Challenge: How to Evaluate XAI?

There are many XAI methods, but hard to quantitatively assess explanations: no ground truth explanation to compare against

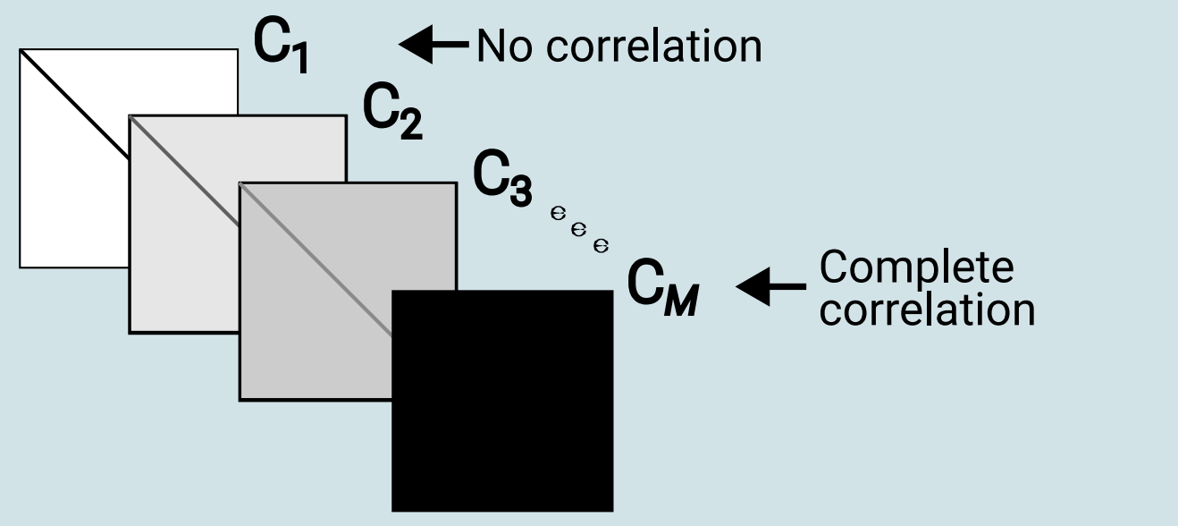
Mamalakis et al.:
[7] XAI benchmarks with known attribution.
The function is designed such that the true explanation is known.



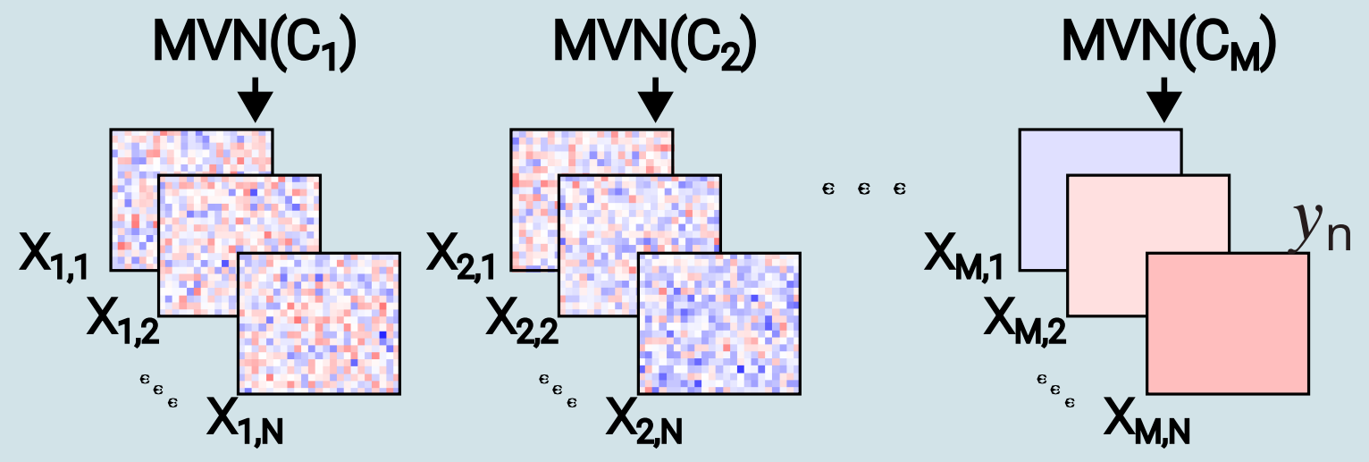
By training models that achieve near-perfect performance, assume that differences between XAI results and ground truth is due to characteristics of the XAI algorithm.

Benchmark Development Pipeline

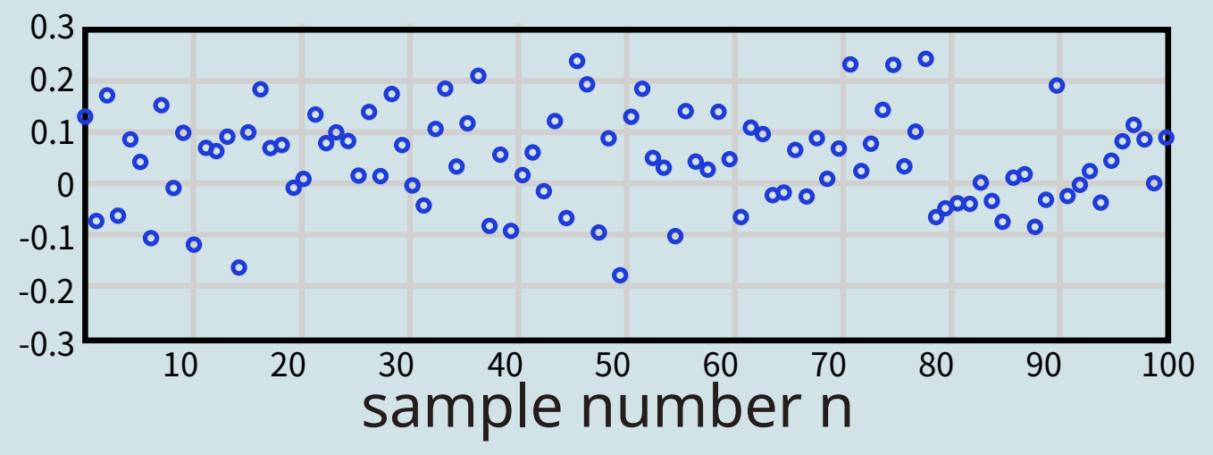
Step 1: Generate M covariance matrices to induce correlation in synthetic samples



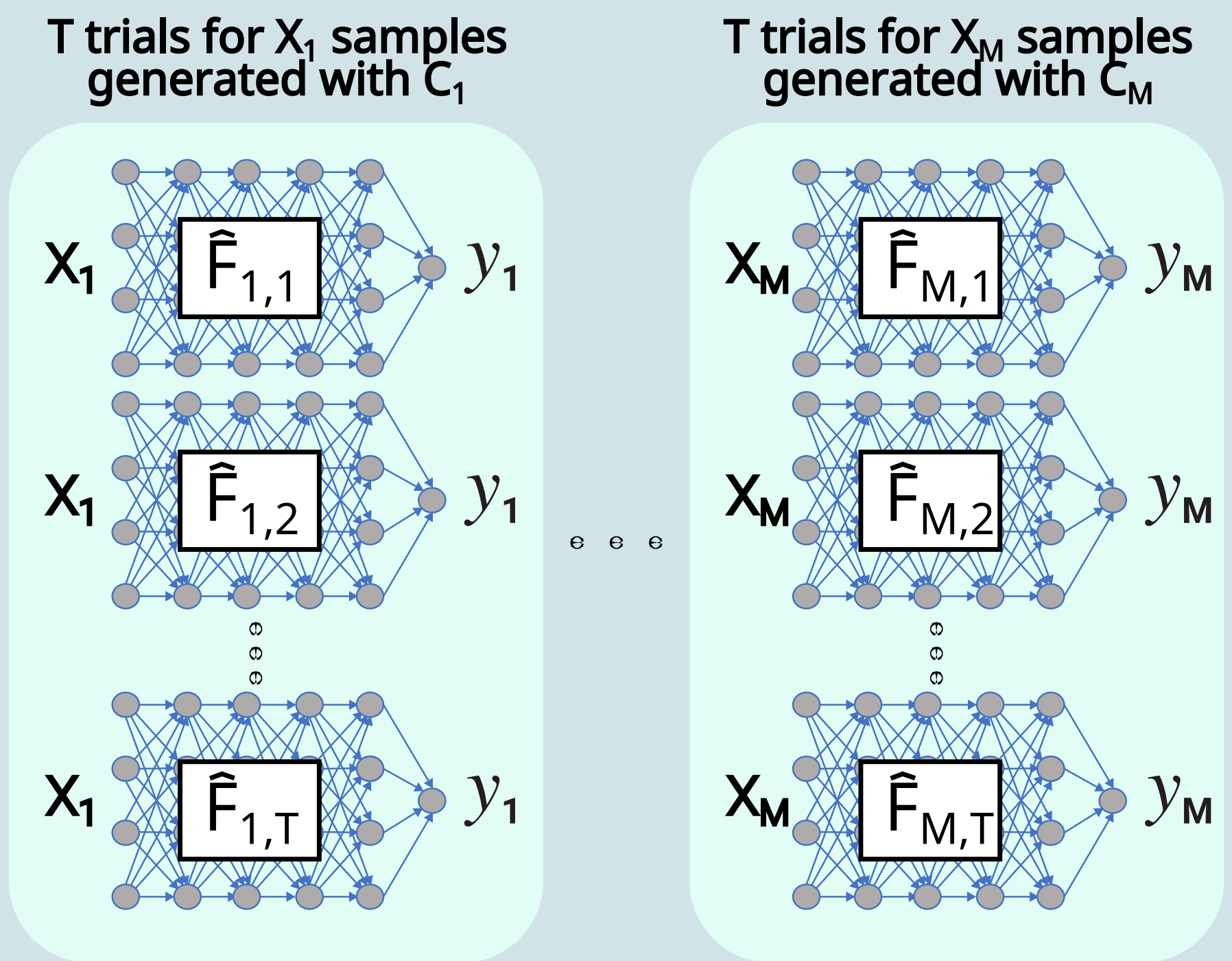
Step 2: For each covariance matrix C_i, generate N samples of $X \in \mathbb{R}^d$ from an MVN



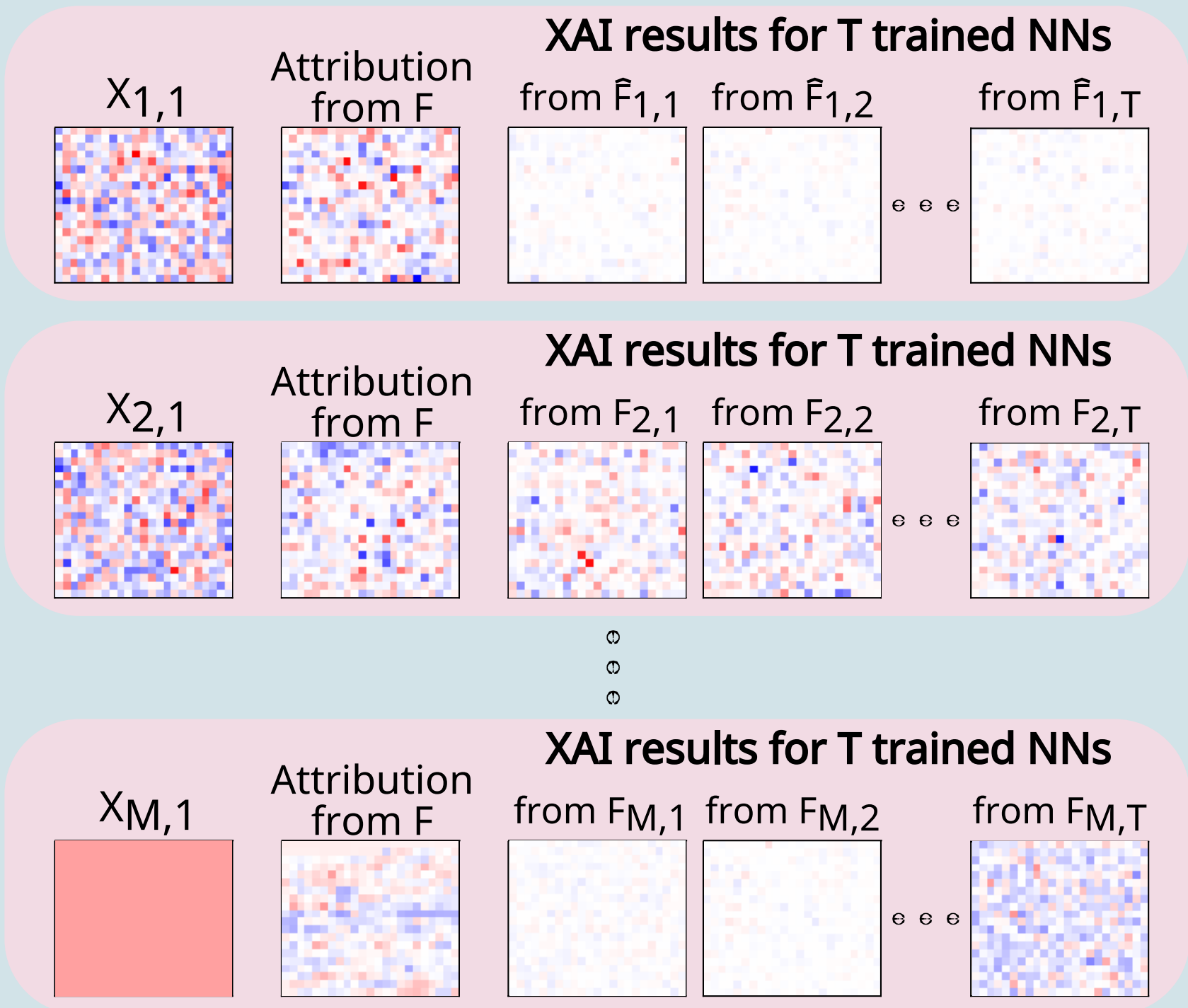
Step 3: Use P to define a known function F that maps each vector X_n into a scalar y_n



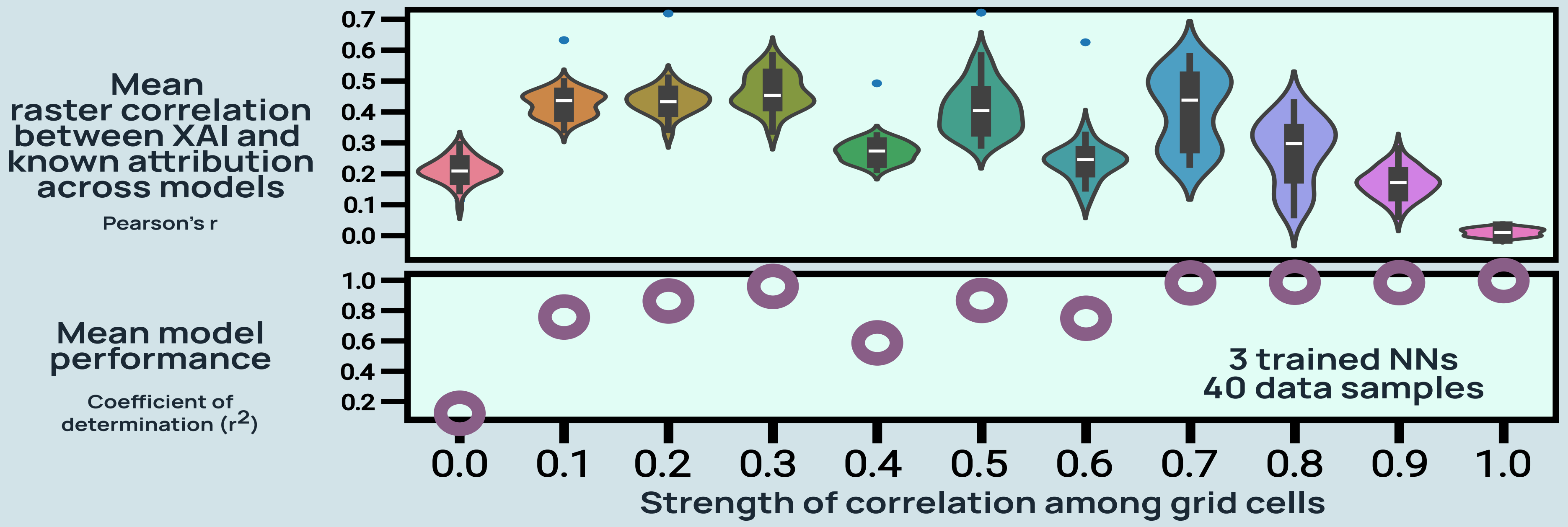
Step 4: For each Ci, pretend F is unknown and train T NNs with inputs X_n, outputs y_n


T trials for X1 samples generated with C1
T trials for XM samples generated with CM
X1
F̂1,1
y1
F̂1,2
F̂1,T
XM
F̂M,1
yM
F̂M,2
F̂M,T

Step 5: Use XAI methods to explain each NN and compare explanation consistency



Benchmark Results

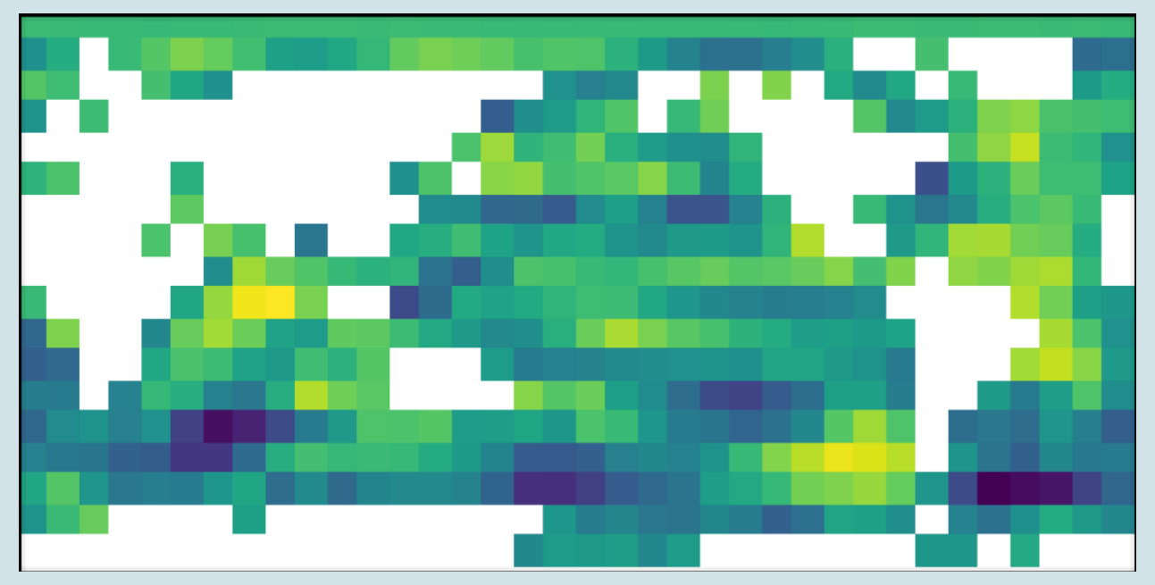


Initially, no patterns to learn from:
poor model & inconsistent XAI

With high correlation, many equivalent models:
accurate model but inconsistent XAI

Next Benchmarks

Teleconnections



[7]

Global, Low-Res SST

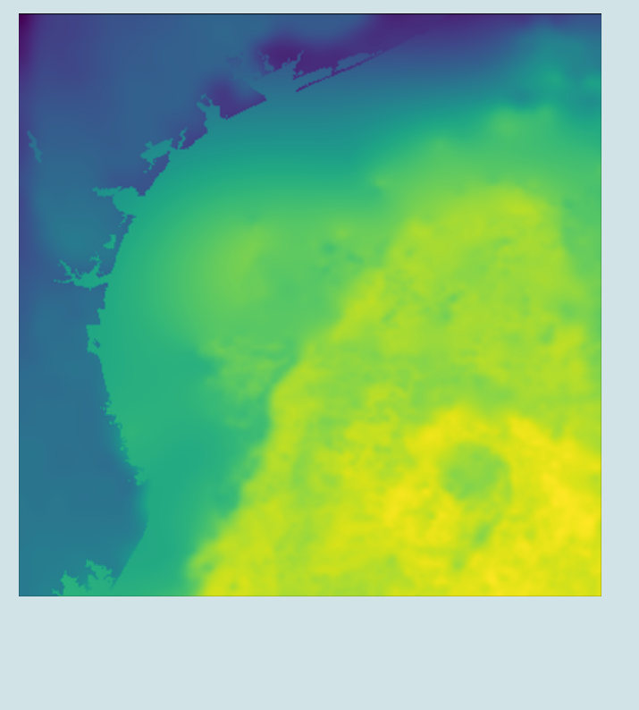
Averages out local values

Discontinuity between cells

Long-range dependencies

Idea:
Clustering based on correlation matrix

Autocorrelation



Local, High-Res SST

Zoom in on a smaller region

Huge autocorrelation influence

Long-range is less important

Idea:
Clustering based on similar values in a single sample

References

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[2] Kamangir, H., Collins, W., Tissot, P., King, S. A., Dinh, H. T. H., Durham, N., & Rizzo, J. (2021). FogNet: A multiscale 3D CNN with double-branch dense block and attention mechanism for fog prediction. Machine Learning with Applications, 5, 100038.

[3] Sentinel 2 image. https://www.azavea.com/blog/2021/02/08/cloud-detection-in-satellite-imagery/

[4] Abid, M. A., Kucharski, F., Molteni, F., & Almazroui, M. (2023). Predictability of Indian Ocean precipitation and its North Atlantic teleconnections during early winter. npj Climate and Atmospheric Science, 6(1), 17.

[5] Chin, T. M., Vazquez-Cuervo, J., & Armstrong, E. M. (2017). A multi-scale high-resolution analysis of global sea surface temperature. Remote sensing of environment, 200, 154-169.

[6] Adamiak, Maciej, Krzysztof Będkowski, and Anna Majchrowska. "Aerial imagery feature engineering using bidirectional generative adversarial networks: a case study of the pilica river region, poland." Remote Sensing 13.2 (2021): 306.

[7] Mamalakis, Antonios, Imme Ebert-Uphoff, and Elizabeth A. Barnes. "Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset." Environmental Data Science 1 (2022): e8.

This material is based upon work supported by the National Science Foundation under award 2019758. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation