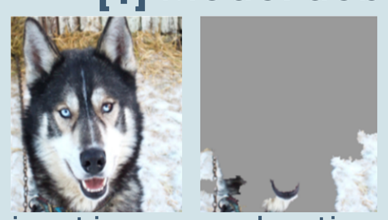
Using Grouped Features to Improve Explainable AI Results for Atmospheric AI Models that use Gridded Spatial Data and Complex Machine Learning Techniques

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:



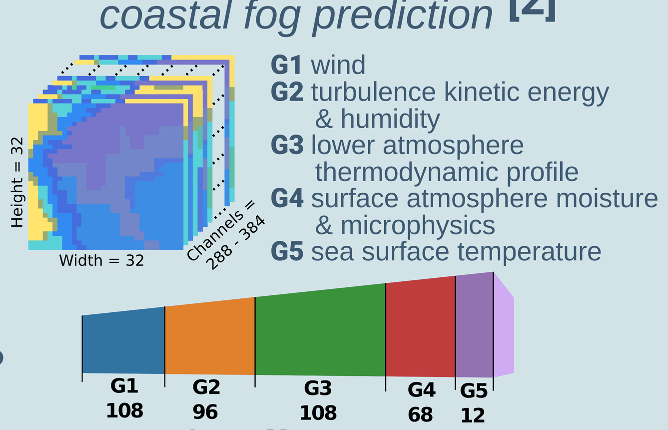
The model has high accuracy for task *wolf or husky?*, but actually looking at snow pixels... many wolf photos have a snowy background.

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

Challenge:
XAI techniques struggle with correlated features

FogNet Model

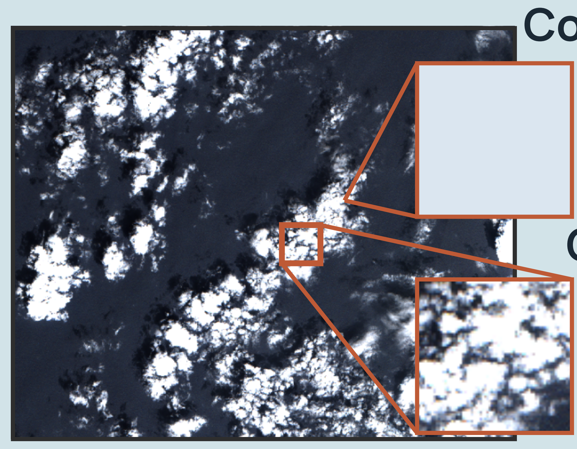
3D CNN for coastal fog prediction [2]



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?

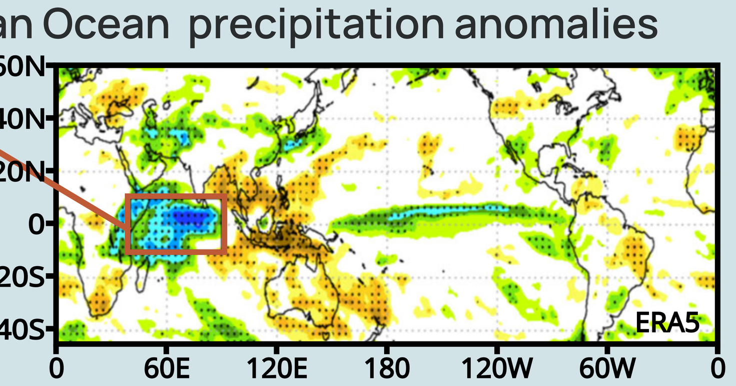
Consider evaluating a superpixel:
Captures a cloud feature that might trigger a change in probability

[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

x_1 $x_2 = x_1$ x_3 x_4

complete correlation

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

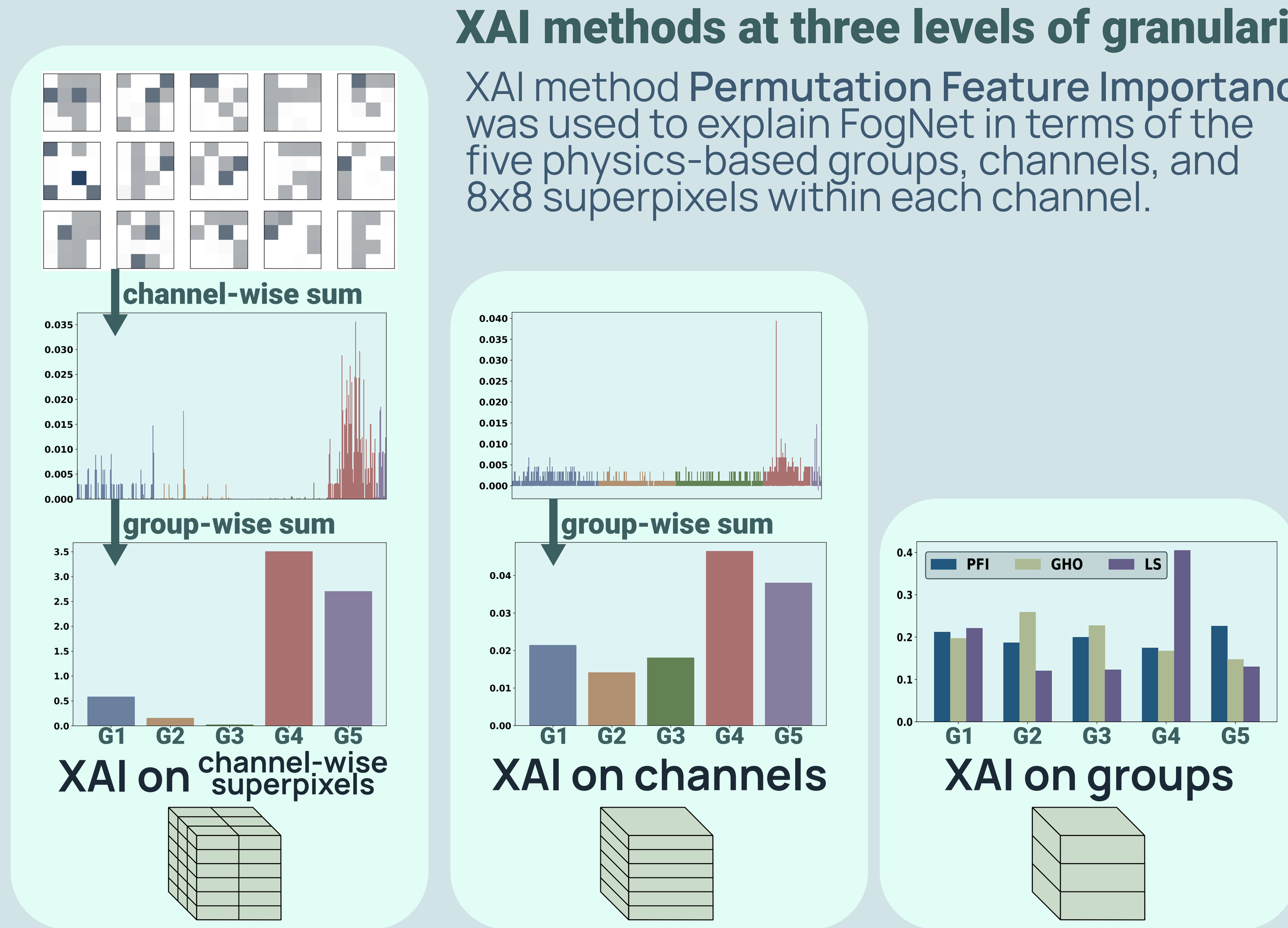
xai(y_3)

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.



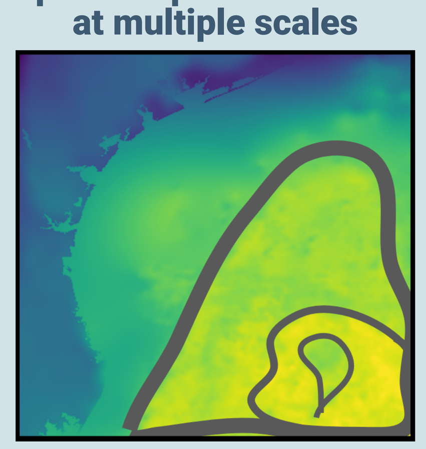
Observation 1:
Explanations are highly sensative to choice of grouping scheme.
Groups suggests that G3 provided ~20% of the predictive skill.
but **Channel-wise superpixels** suggests we could throw G3 out.

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

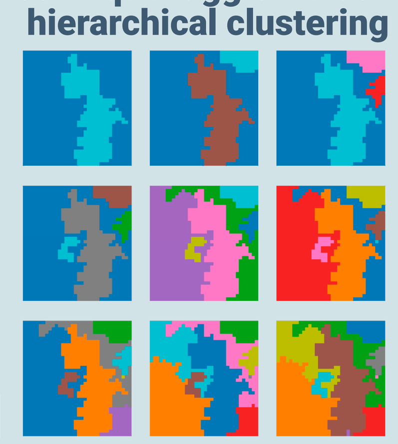
Goal 1: Group features in a data-driven fashion, not arbitrary geometry.
Goal 2: Explain a hierarchy to learn about features across scales.
Goal 3: Strategically select groups since infeasible to explain everything.

Sketch: nested clusters to capture important features at multiple scales



[5]

Technique: agglomerative hierarchical clustering



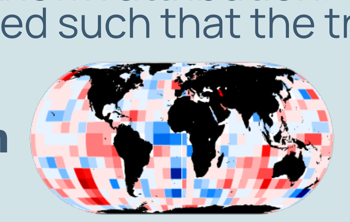
[6]

Challenge: How to Evaluate XAI?

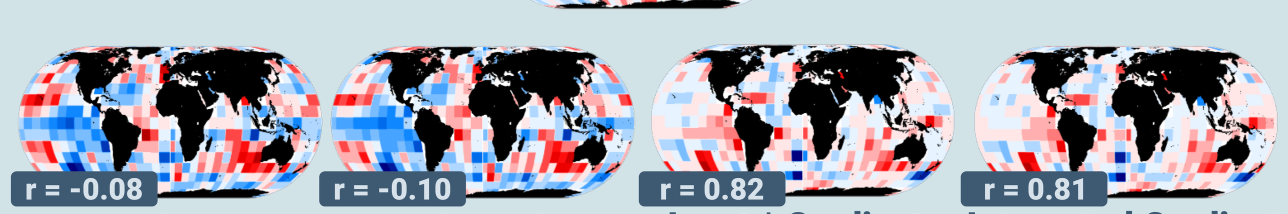
There are many XAI methods, but hard to quantitatively assess explanations: no ground truth explanation to compare against

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

Ground truth



Pearson's r: measure correlation between explanations

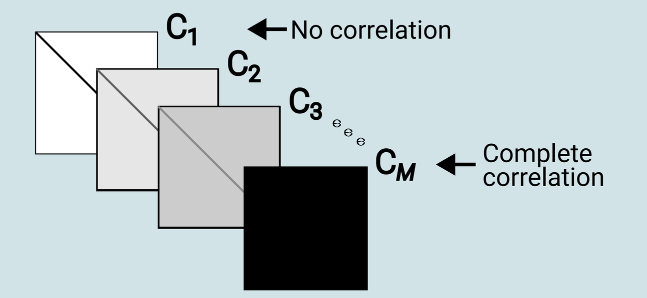


Gradient **Smooth Gradient** **Input * Gradient** **Integrated Gradients**

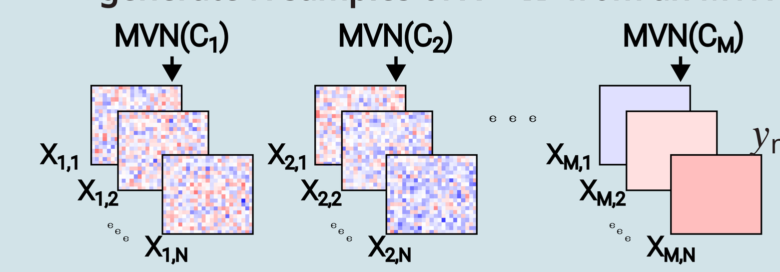
By training models that achieve near-perfect performance, assume that differences between XAI results and ground truth is due to characteristics of the XAI algorithm.

Benchmark Development Pipeline

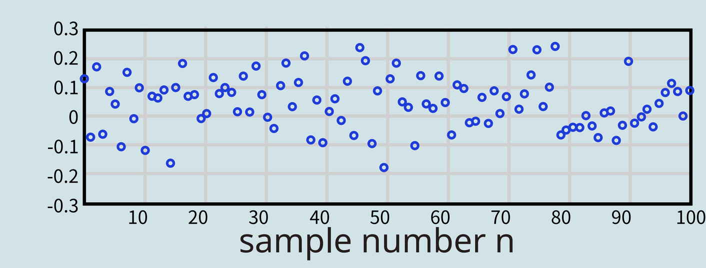
Step 1: Generate M covariance matrices to induce correlation in synthetic samples



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



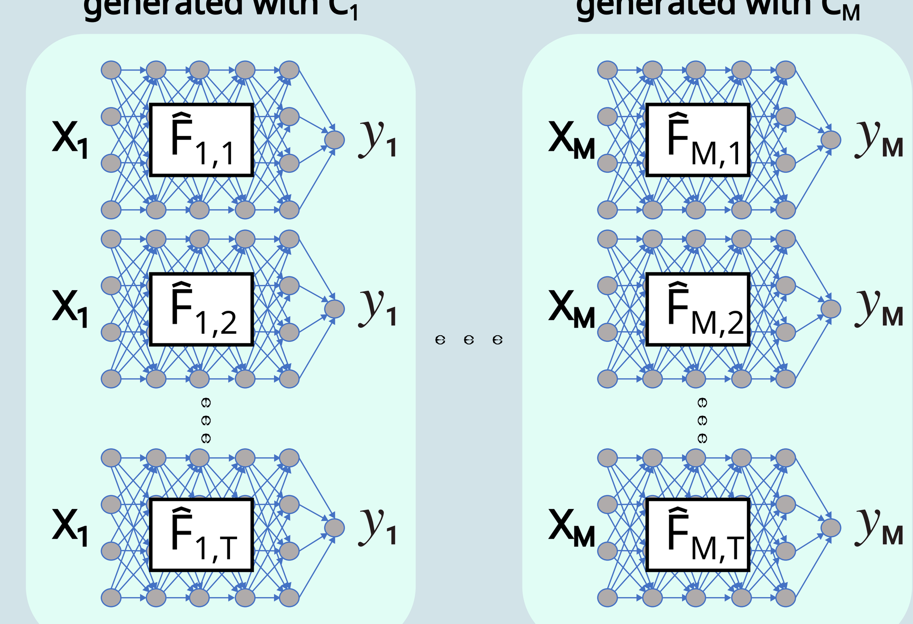
Step 3: Use P to define a known function F that maps each vector X_n into a scalar y_n



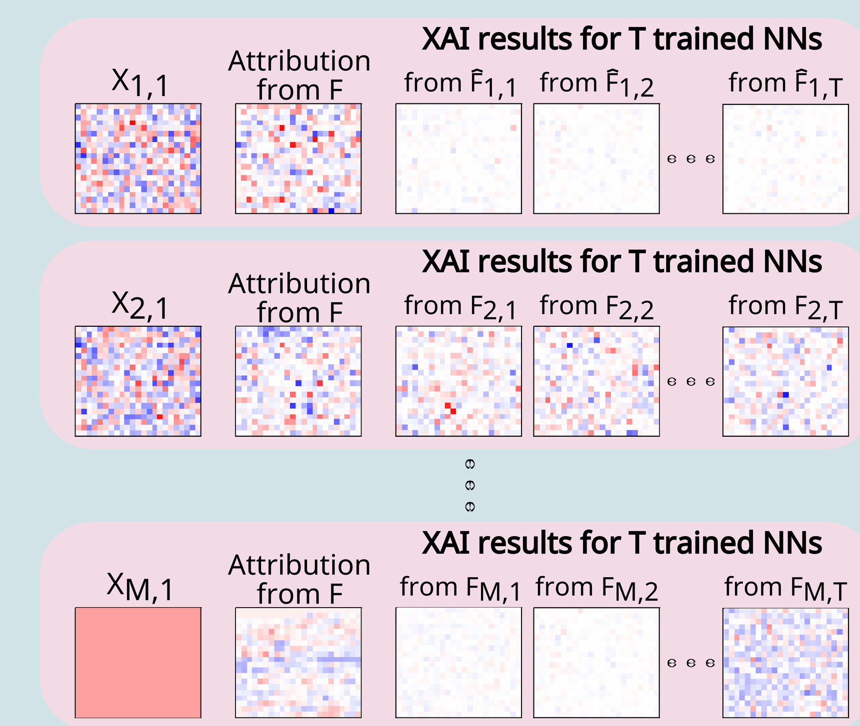
Step 4: For each C_i, pretend F is unknown and train T NNs with inputs X_n, outputs y_n

T trials for X_1 samples generated with C_1

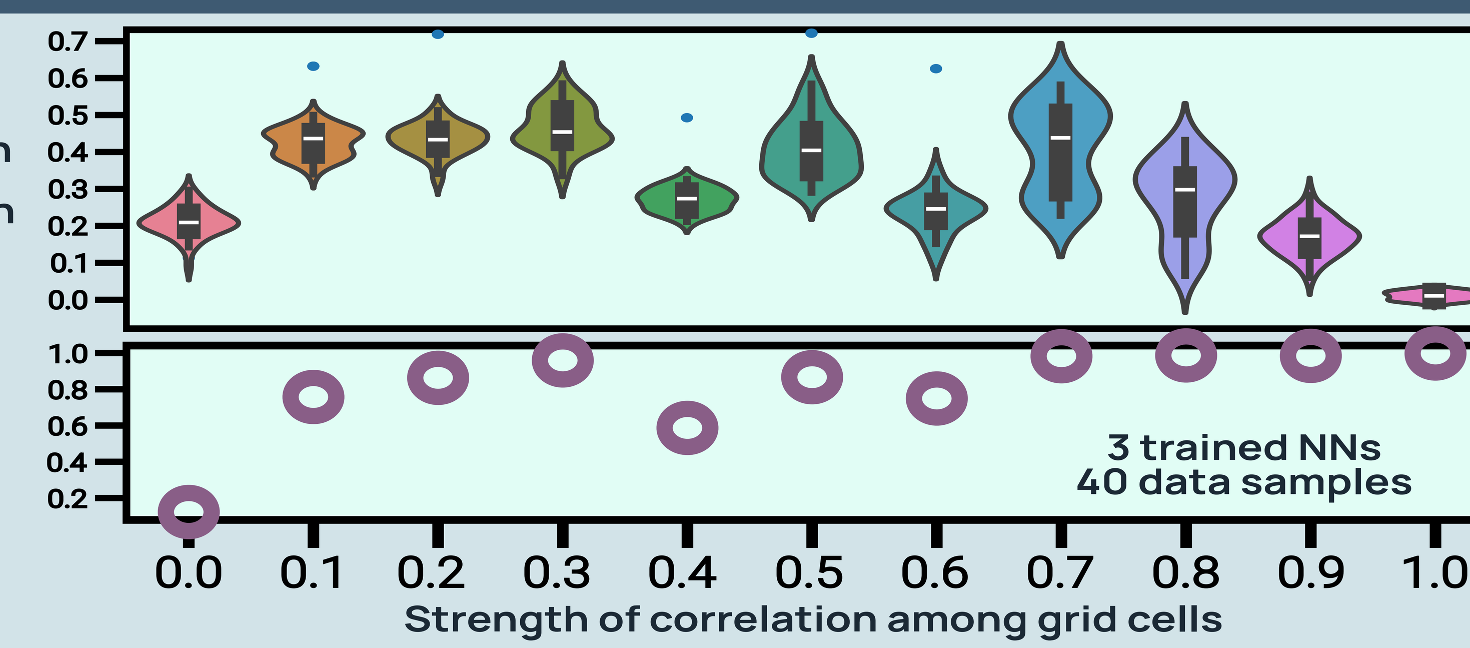
T trials for X_M samples generated with C_M



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



Benchmark Results

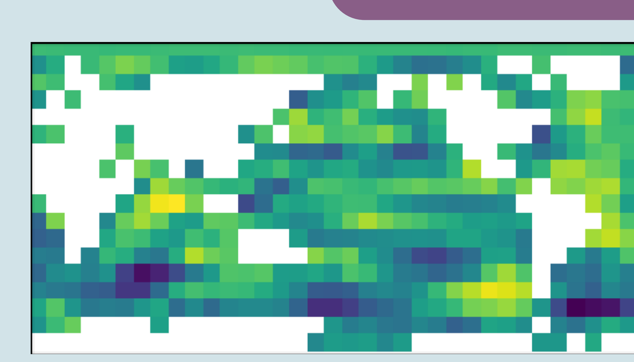


Initially, no patterns to learn from:
poor model & inconsistent XAI

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

Next Benchmarks

Teleconnections



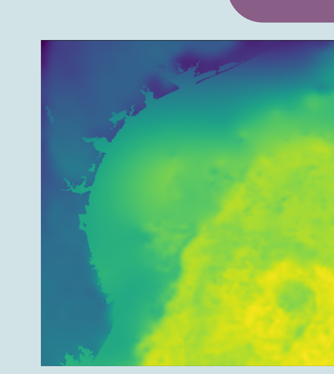
[7]

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