The Influence of Feature Aggregation for Explainable AI for High-Dimensional Geoscience Applications

Evan Krell[a,b,c], Hamid Kamangir[a,b], Waylon Collins[a,d], Scott A. King[a,c], Philippe Tissot[a,b]

(a) NSF AI Institute for Research on Trustworthy AI in Weather, Climate and Coastal Oceanography
(b) Conrad Blucher Institute for Surveying and Science, Texas A&M University - Corpus Christi
(c) innovation in COmputing REsearch (iCORE)
(d) National Weather Service

ai2es.org

This material is based upon work supported by the National Science Foundation under Grant No. ICER-2019758

Manuscript submitted to Environmental Data Science. Working on revisions.

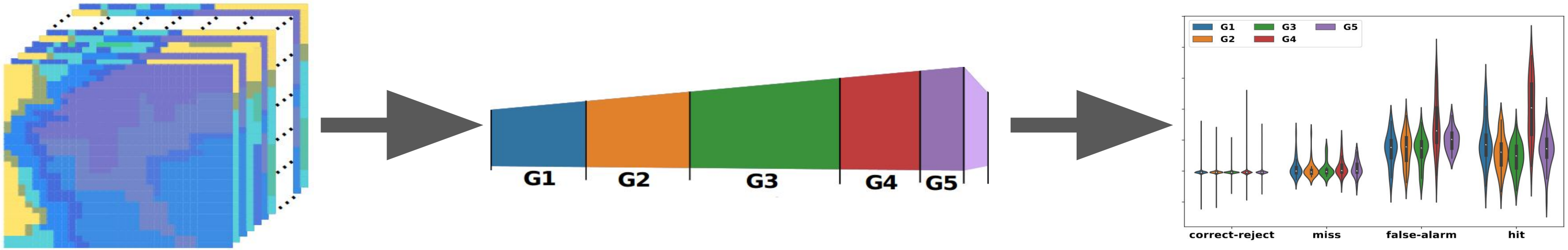
Krell, E., Kamangir, H., Collins, W., King, S. A., & Tissot, P. (2023). Aggregating XAI methods for insights into geoscience models with correlated and high-dimensional rasters.

Outline

1. Explainable Artificial Intelligence for Geoscience Models

2. Case Study: FogNet, 3D CNN for Forecasting Coastal Fog

3. Toward Synthetic Benchmarks for XAI Evaluation

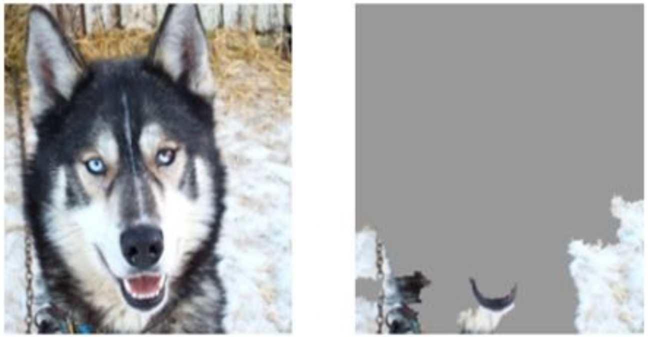


Geospatial data | **One of many aggregation schemes** | **XAI results for each group**

Research question:
How does the choice of grouping raster elements into features influence the explanations generated from XAI methods?

Explainable Artificial Intelligence (XAI)

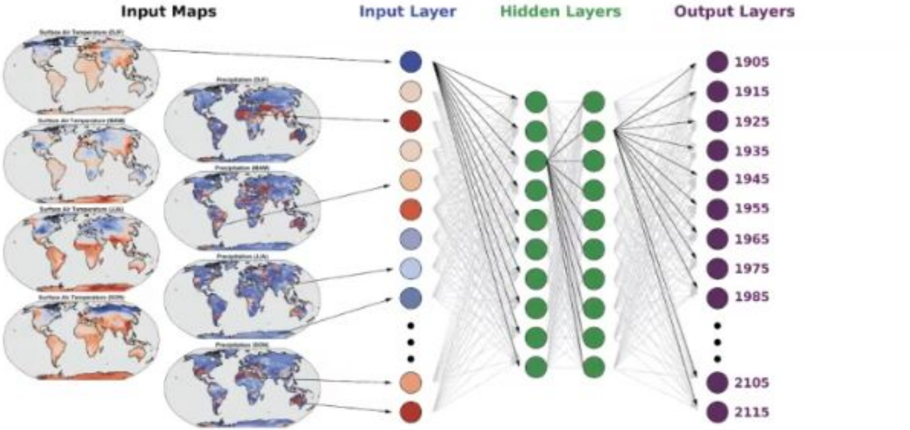
Model verification

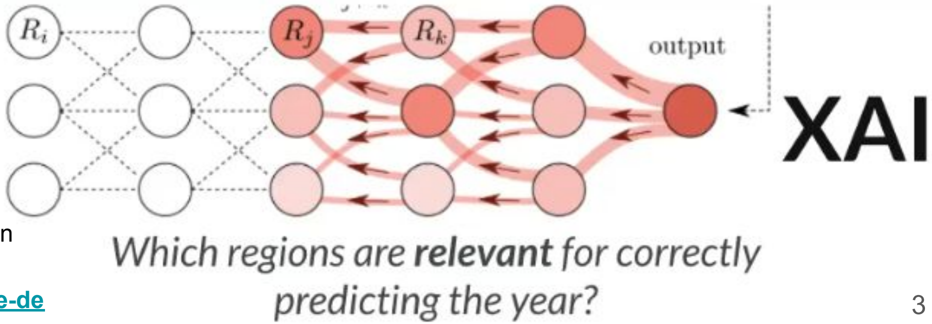


(a) Husky classified as wolf (b) Explanation

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).

Scientific insights

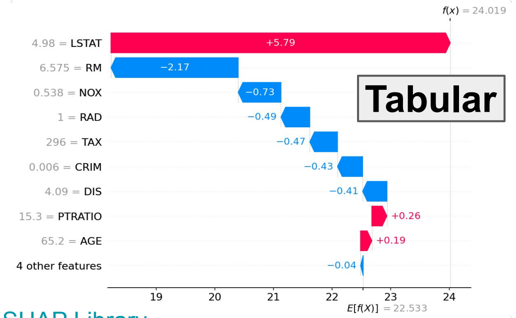




Presentation: Explainable AI (XAI) for Climate Science: Detection, Prediction and Discovery. Elizabeth Barnes. 2022. https://www.imsi.institute/videos/explainable-ai-xai-for-climate-science-detection-prediction-and-discovery/

XAI Approaches

Local Explanation: instance explanation based on a single sample

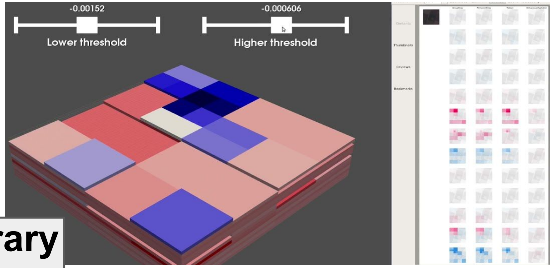


Tabular

[SHAP Library]

RGB

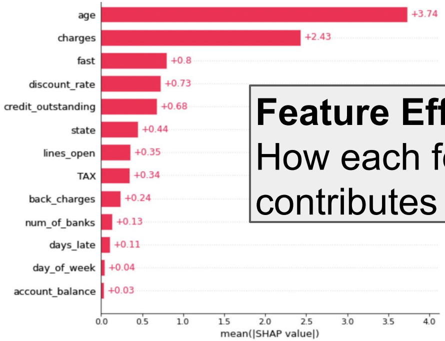
[Gradient-weighted Class Activation Mapping - Grad-CAM- | by Mohamed Chetoui | Medium]



Arbitrary raster

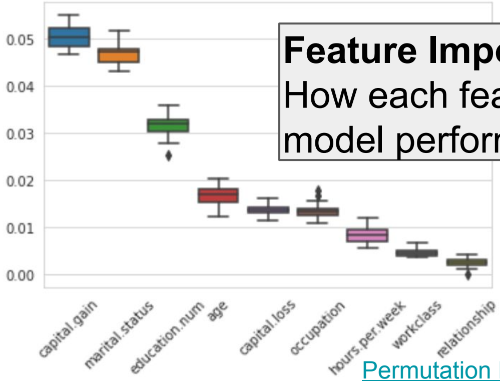
[PartitionShap: viewing multi-channel explanations in 3D]

Global Explanation: summary explanation over a set of samples



Feature Effect:
How each feature contributes to output

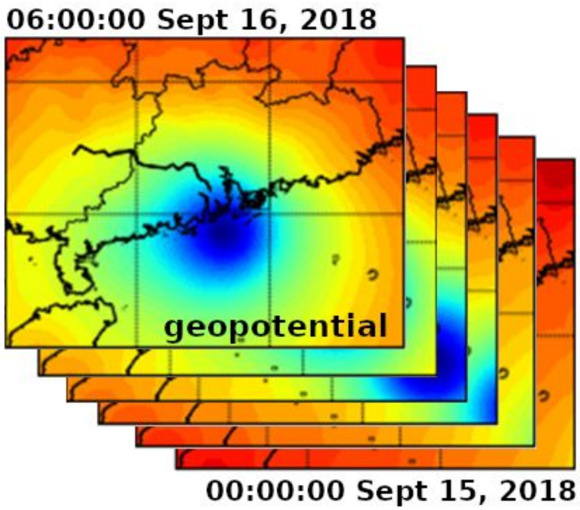
[Feature Importance - Arize AI]

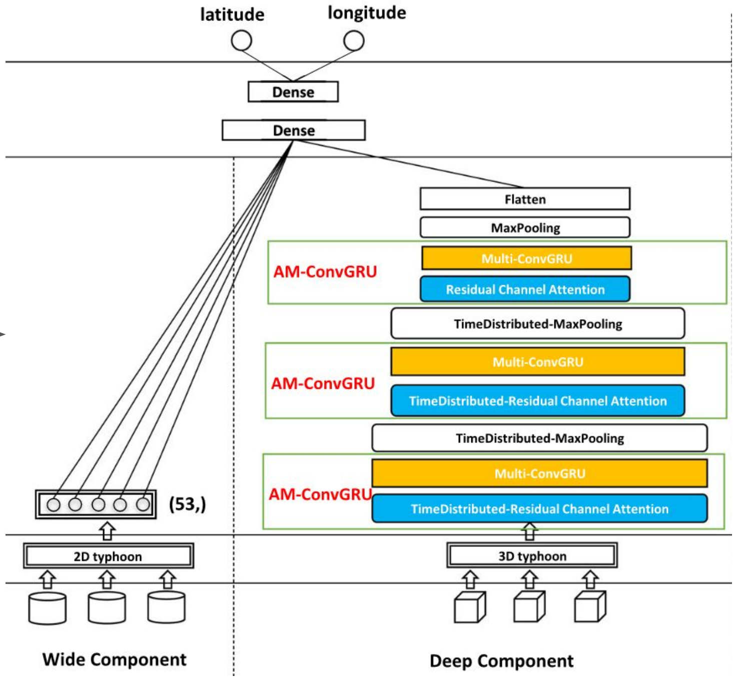


Feature Importance:
How each feature influences model performance

[Permutation Feature Importance]

Geoscience AI Models





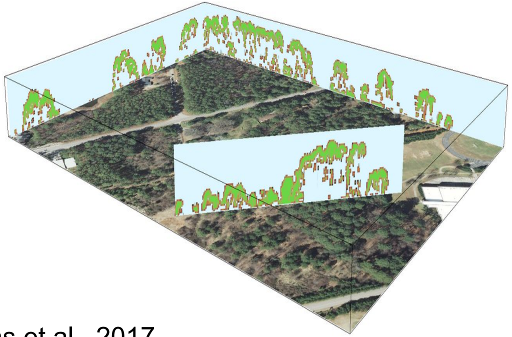
Xu, Guangning, et al. "AM-ConvGRU: a spatio-temporal model for typhoon path prediction." *Neural Computing and Applications* 34.8 (2022): 5905-5921.

- High-dimensional geospatial raster (gridded) data is used to train complex machine learning models.
- Often complex models (e.g. Deep Neural Net) greatly outperform simpler alternatives (e.g. Random Forest).
- These models are hard to interpret: what are the model's decision-making strategies?

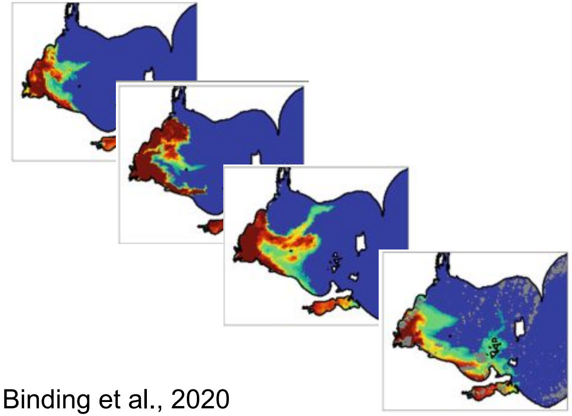
Autocorrelation in Geospatial Data

Harmful algal bloom

[NASA Earth Observatory]

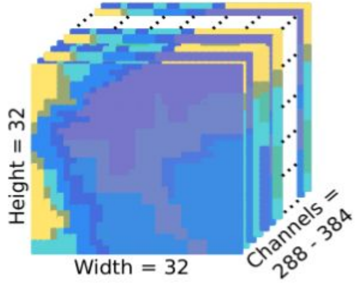


Petras et al., 2017
https://opengeospatialdata.springeropen.com/articles/10.1186/s40965-017-0021-8



Binding et al., 2020
https://link.springer.com/chapter/10.1007/698_2020_589

2D spatial | **3D spatial** | **3D temporal**



FogNet: 4D data (spatio-temporal) packaged as 3D

VVel 850mb t0 | *VVel 850mb t1* | *VVel 850mb t2* | *VVel 850mb t3* | *VVel 875mb t0* | …

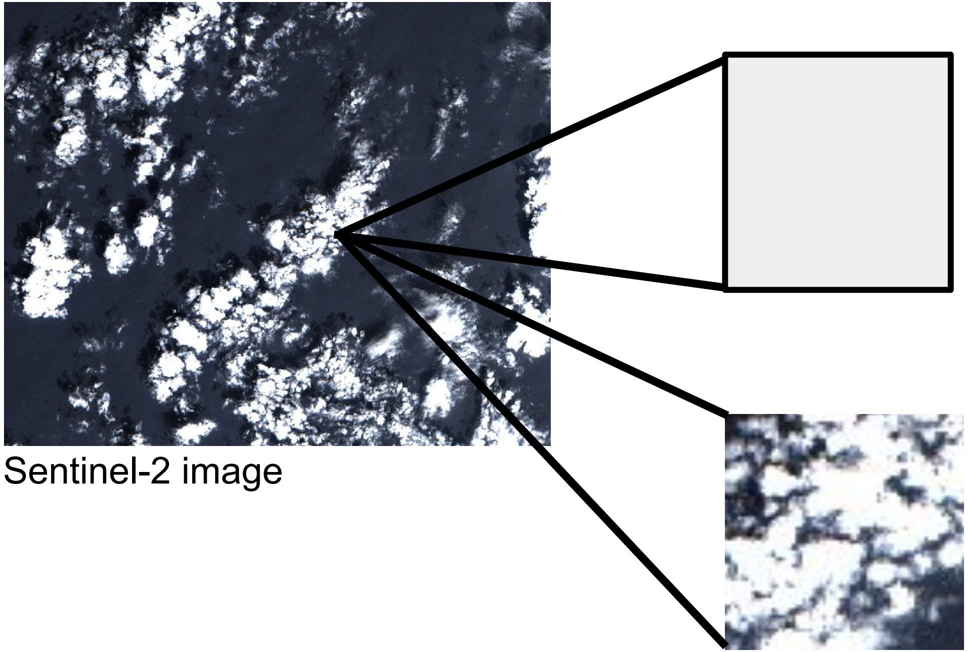
4 adjacent bands → time sequence

followed by next altitude

Spatial Autocorrelation & XAI

XAI: how much does each pixel contribute to detection of clouds?

Cloud detection model



Sentinel-2 image

Consider evaluating individual pixels:

If you change this pixel, does model output change?

Hopefully, robust to noise →no significant change

No pixels are important… but model detects clouds!

Consider evaluating superpixels:

Changing this superpixel, does model output change?

Clearly a cloud feature that could have been learned

Removing it could lower model’s detection confidence

For meaningful XAI results: need to group grid cells and explain those groups

Grouped Geospatial XAI Assumptions

Coarse groups:

- **More** reliable feature importance/effect ranking
- **Lower** resolution model insights

Granular groups:

- **Less** reliable feature importance/effect ranking
- **Higher** resolution model insights

When XAI highlights an influential feature:

- That feature is expected to actually be influential
- But the features **not** highlighted could be as or more influential

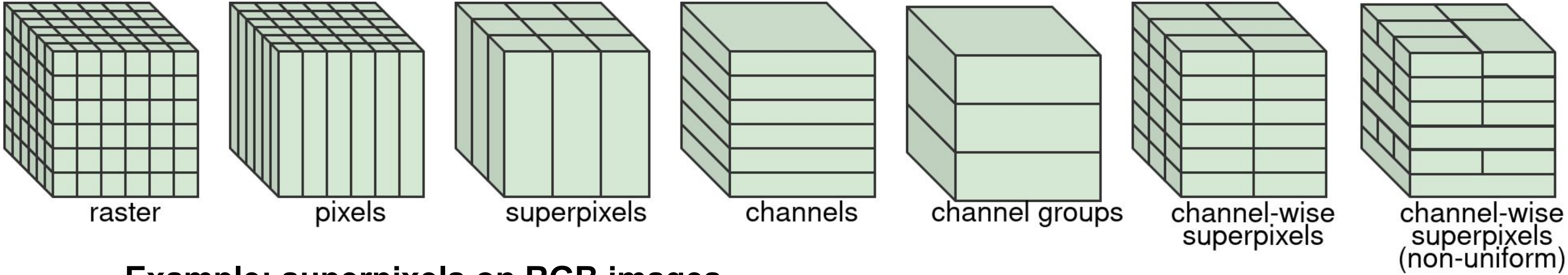
When grouping scheme granularities disagree:

- Suggests something about the scale of the learned feature

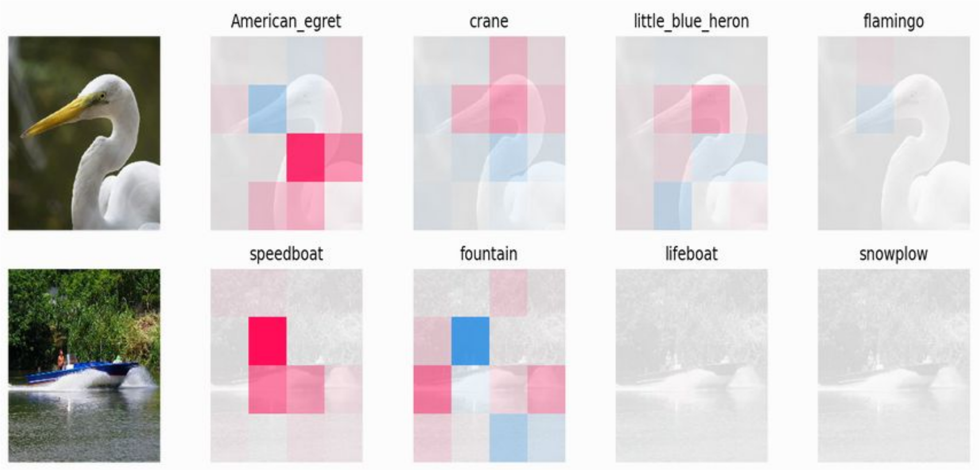
It is very easy to apply XAI methods and be greatly mislead by the results

Geometric Grouping Schemes

Several schemes for grouping raster elements

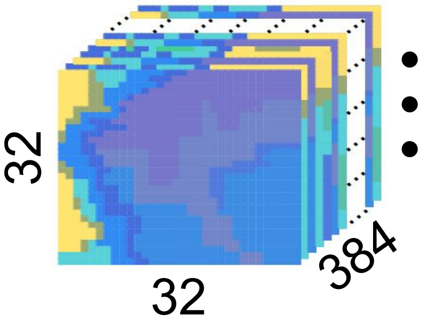


Example: superpixels on RGB images



- **PartitionSHAP:** recursively computes SHAP values by halving superpixels
  [shap.explainers.Partition — SHAP documentation]
- Recursion guided by change in SHAP value
- By default, only considers rows & cols
- Our fork: **Channel-wise PartitionSHAP**
  https://github.com/conrad-blucher-institute/shap

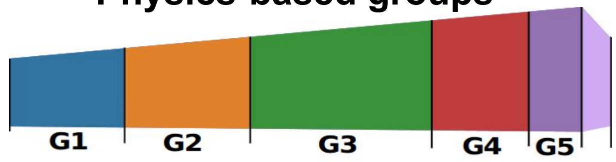
FogNet: 3D CNN for Forecasting Coastal Fog



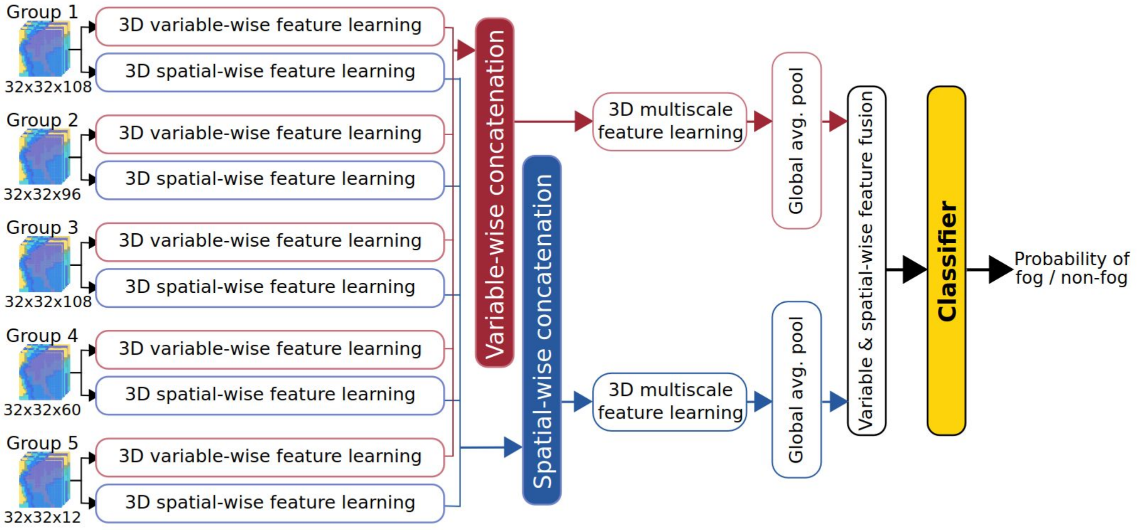
- 3D CNN with attention, dense block, & dilated convolution
- Beats NOAA's operational `High Resolution Ensemble Forecast (HREF)`
- Input data: spatio-temporal raster of metocean variables, divided into 5 related groups

https://gridftp.tamucc.edu/fognet/

Physics-based groups



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



XAI Methods Applied

Feature Importance

Global methods → how did feature influence model performance?

- **Permutation Feature Importance (PFI):** replace feature with permuted values
  McGovern, Amy, et al. "Making the black box more transparent: Understanding the physical implications of machine learning." Bulletin of the American Meteorological Society 100.11 (2019): 2175-2199.

- **LossSHAP (LS):** approximate Shapley values . . . combinatorial complexity
  Covert, Ian, Scott Lundberg, and Su-In Lee. "Feature removal is a unifying principle for model explanation methods." arXiv preprint arXiv:2011.03623 (2020).

- **Group-hold-out (GHO):** entirely remove feature & retrain model
  Au, Quay, et al. "Grouped feature importance and combined features effect plot." Data Mining and Knowledge Discovery 36.4 (2022): 1401-1450.

Feature Effect

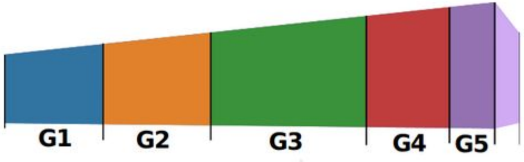
Local methods → how did feature influence specific model decision?

- **Channel-wise PartitionSHAP (CwPS):** approximate Shapley values for superpixels in each channel
  Kamangir, Hamid, et al. "Importance of 3D convolution and physics on a deep learning coastal fog model." Environmental Modelling & Software (2022): 105424.

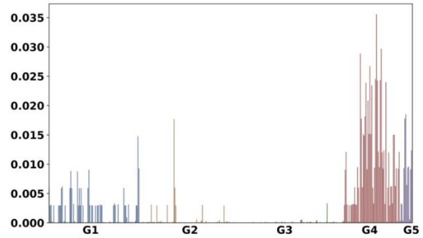
Feature Importance Results

(a) CwSP, top 15 channels

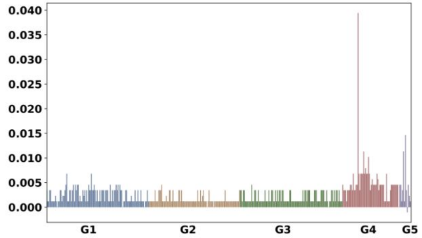
- 3D CNN with double-branch dense block & attention mechanism
- Applied geometric rather than data-driven groupings for XAI
- Compared 3 grouping schemes:
  - Physics-based channel groups
  - Channel-wise
  - Channel-wise SuperPixels (CwSP)



(b) CwSP, channel sums

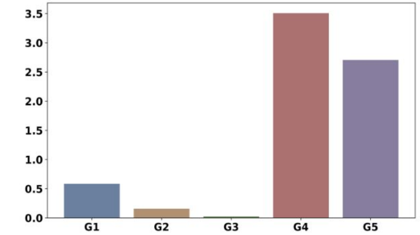


(c) Channel-wise

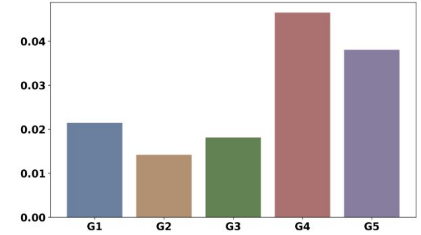


- Groups 1-3 dilute as we increase granularity
- Groups 1-3 contain vertical profiles where small-scale features have little predictive power
- Suggests that FogNet learns 3D features

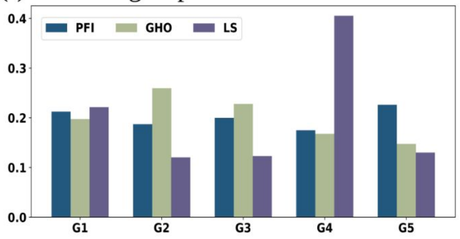
(d) CwSP, group sums



(e) Channel-wise, group sums



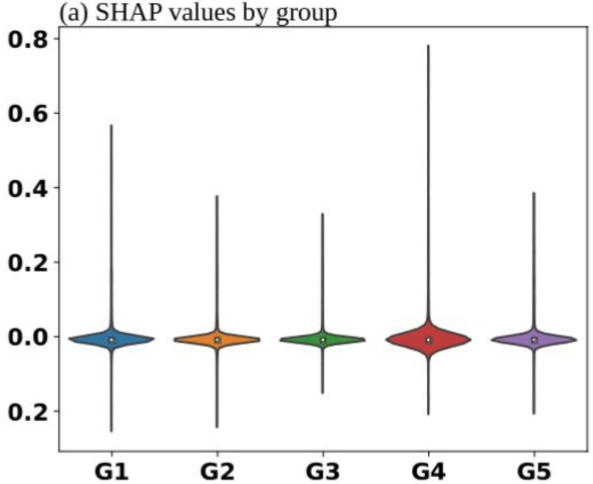
(f) Channel groups

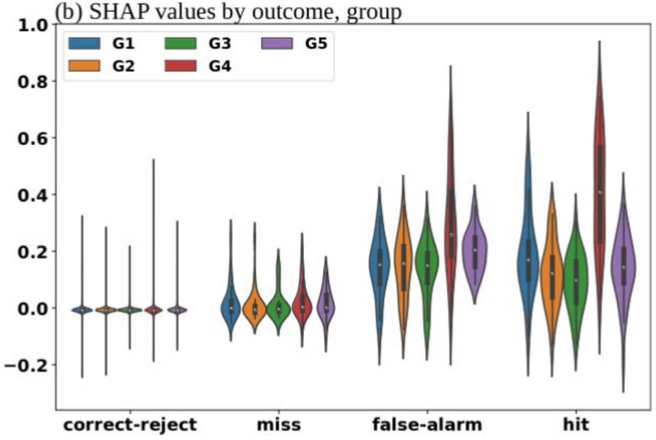


PFI: Permutation Feature Importance
GHO: Group Hold-Out
LS: LossSHAP

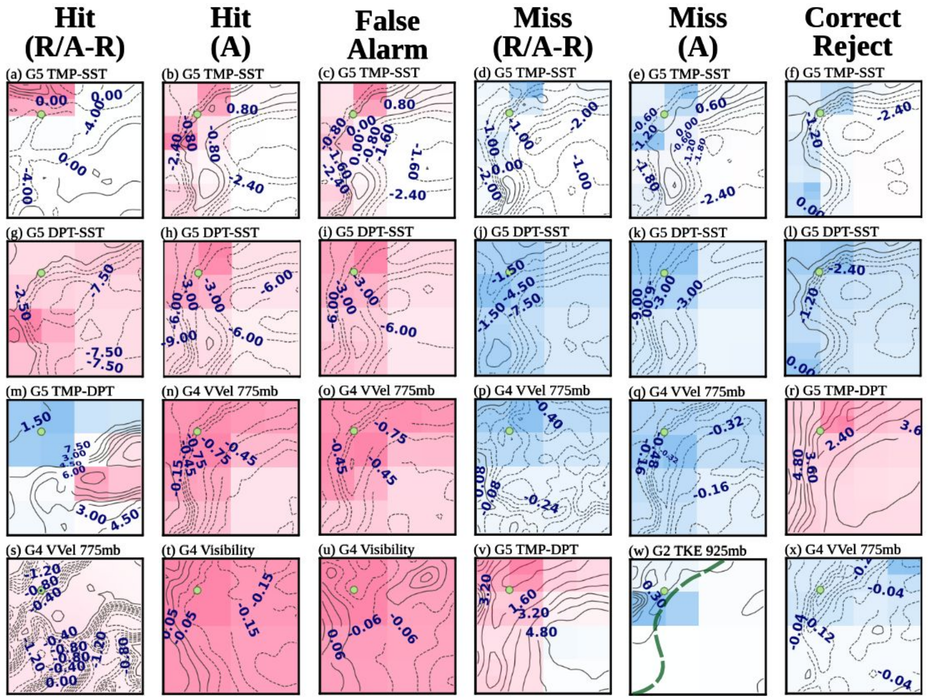
Feature Effect Results

SHAP on groups





Channel-wise PartitionSHAP
(channels aggregated & ranked)



XAI Insights for Geospatial Models

XAI Pitfalls

1. **FogNet does not use G3 (wind)**
   - Based on more granular XAI, appears that G3 has no influence of the model
   - But we know that G3 responsible for ~20% of the performance

2. **FogNet relies mostly on information around the target**
   - Based on CwPS, appears that FogNet is very focused on target region and ignores offshore
   - But we know that G1 - G3 are important even though they appear less using superpixels

XAI Insights

1. FogNet appears strongly influenced by SST near the target airport (KRAS)
2. FogNet appears to learn large-scale patterns for G1 - G3, such as in the vertical wind profile
3. FogNet appears to only learn strategies for the majority fog case: advection fog

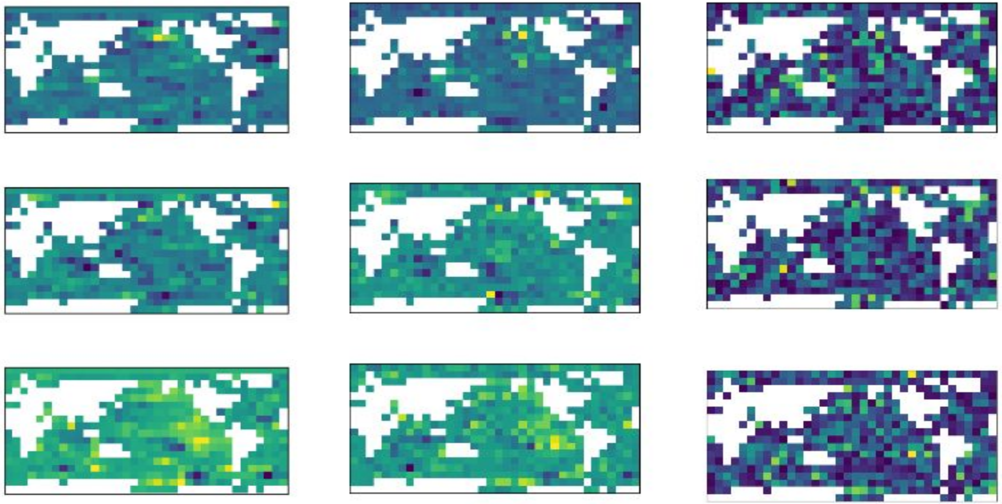
See upcoming manuscript for detailed meteorological interpretation of XAI results from Waylon Collins from the National Weather Service

Synthetic Benchmarks for XAI Assessment

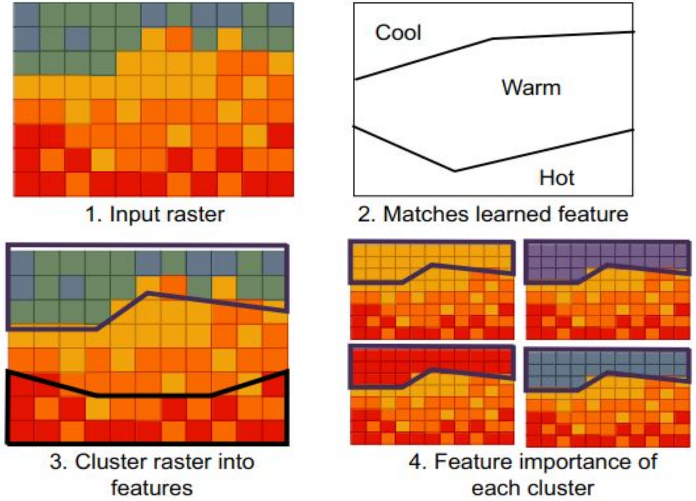
- Developing grouping strategies to improve XAI
- But hard to determine which produces best explanations
- Extending work by Mamalakis et al.: benchmark data & functions with known attribution

  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.

Ground truth explanation | **Integrated Gradients** | **Saliency Maps**



Then we can explore data-driven feature aggregation schemes



Key Conclusions

1. XAI outputs can be better interpreted by understanding what question the method asks

2. XAI should be analyzed in various ways to avoid major pitfalls

Questions?