

A comparison of explainable AI solutions for a climate change prediction task

Philine Lou Bommer¹ Marlene Kretschmer² Dilyara Bareeva¹ Kadircan Aksoy¹ Marina M.-C. Höhne¹

¹Institute for Machine Learning, Technical University Berlin, Berlin, Germany

²Department of Meteorology, University of Reading, Reading, UK



The University of Reading

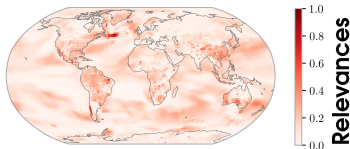
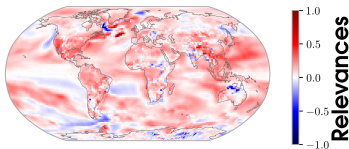
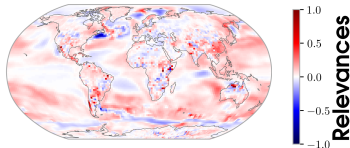
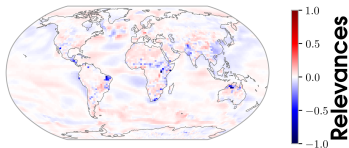
Motivation

Data:

- complex data $\rightarrow$ complex prediction behavior

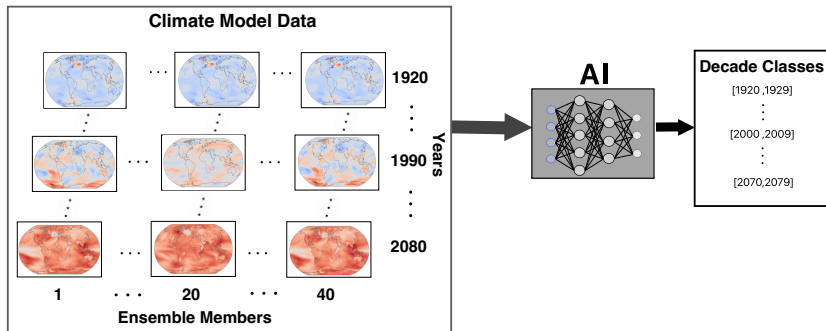
Explanation:

- lack of ground truth



Experimental Setup

[Labe and Barnes, 2021]



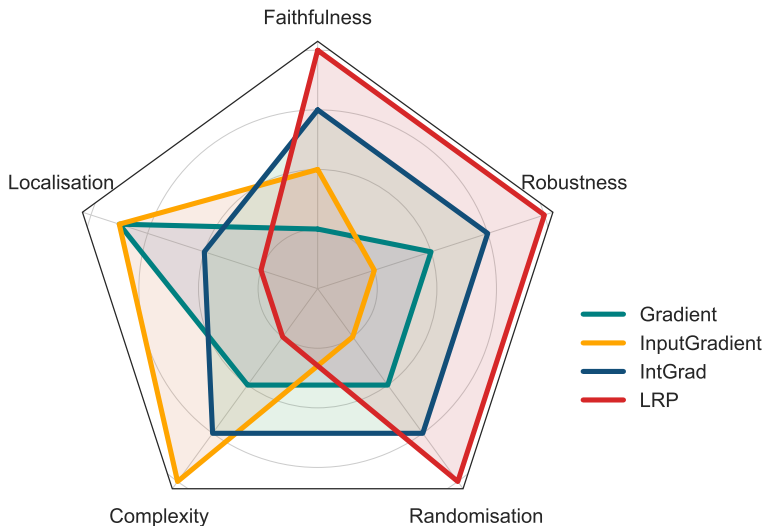
Explanation Methods

- Gradients [Baehrens et al., 2010]
- InputGradients (InputGradients)
- Integrated Gradients (IntGrad) [Sundararajan et al., 2017]
- Layer-wise Relevance Propagation (LRP) [Bach et al., 2015]

XAI Evaluation

- Quantus [Hedström et al., 2022]¹

¹<https://github.com/understandable-machine-intelligence-lab/Quantus>



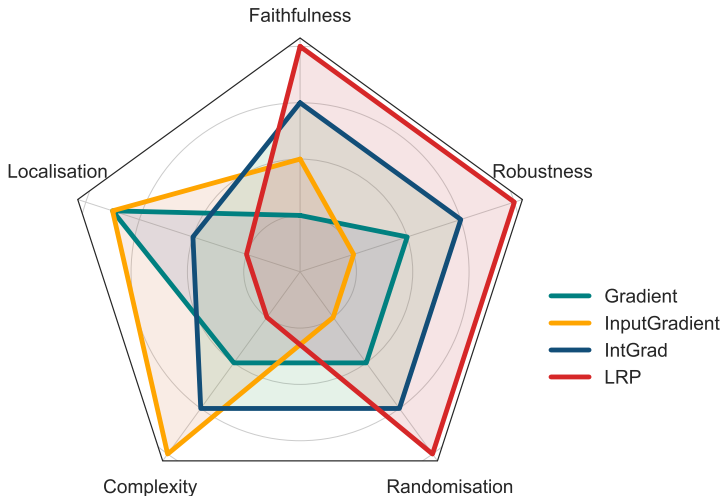
References I

-  Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015).
On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.
[PLOS ONE](#), 10(7):e0130140.
-  Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. (2010).
How to explain individual classification decisions.
[Journal of Machine Learning Research](#), 11(61):1803–1831.
-  Hedström, A., Weber, L., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., and Höhne, M. M. C. (2022).
Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations.

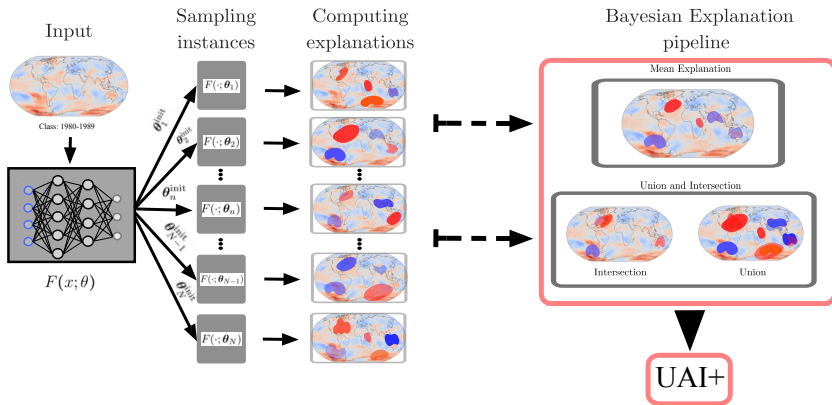
References II

-  Labe, Z. M. and Barnes, E. A. (2021).
Detecting climate signals using explainable AI with
single-forcing large ensembles.
[Journal of Advances in Modeling Earth Systems](#), 13(6).
-  Sundararajan, M., Taly, A., and Yan, Q. (2017).
Axiomatic attribution for deep networks.
In [International conference on machine learning](#), pages
3319–3328. PMLR.

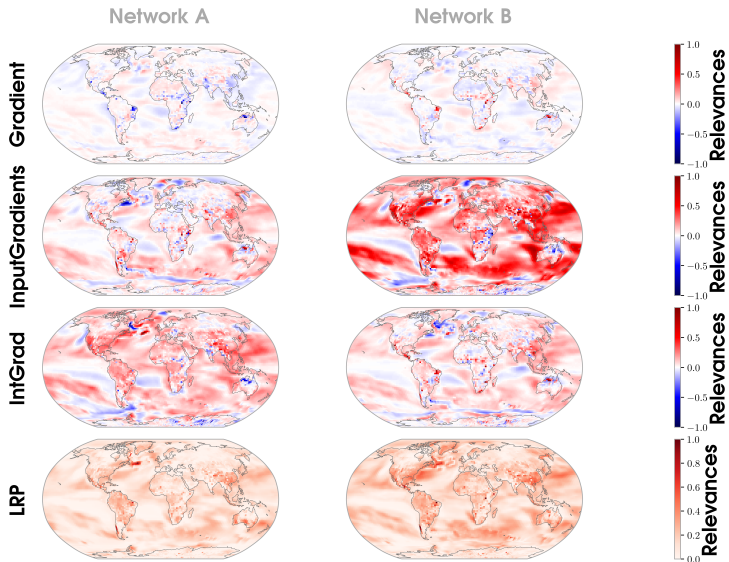
XAI Evaluation



Network Uncertainty

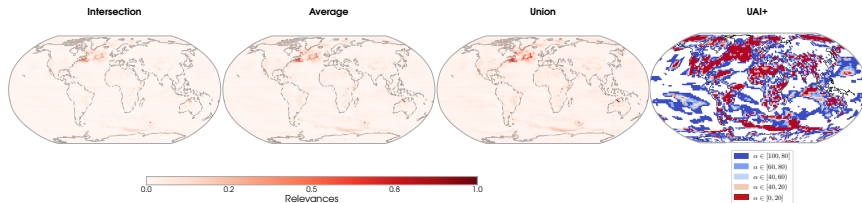


Network Uncertainty



Network Uncertainty

2040



2079

