Evaluation of explainable AI solutions in climate science

Philine Lou Bommer*

Anna Hedström

Marlene Kretschmer

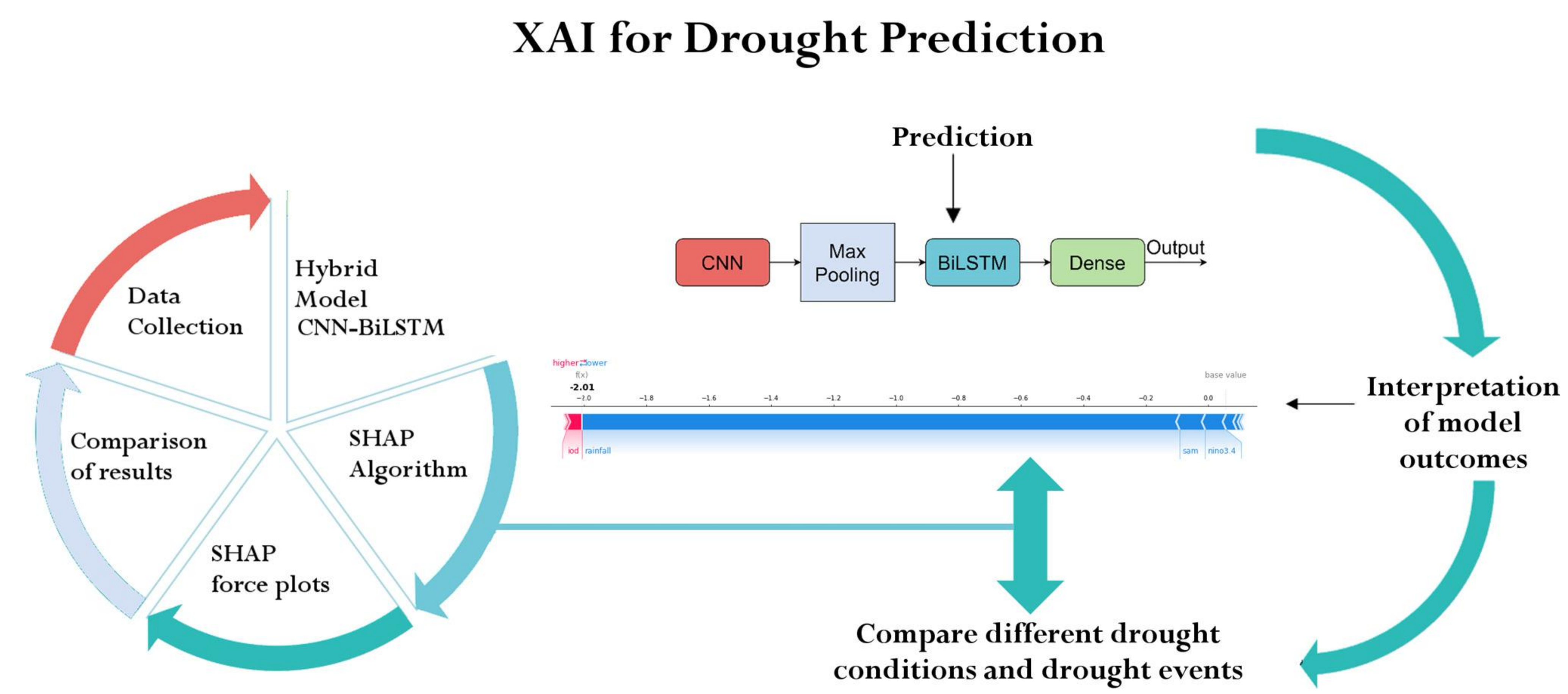
Dilyara Bareeva

Marina M.-C. Höhne

1. Introduction

❖ **Explainable AI (XAI)**: deeper understanding of network decision

➢ assessment of the model skill (trustworthiness and reliability)



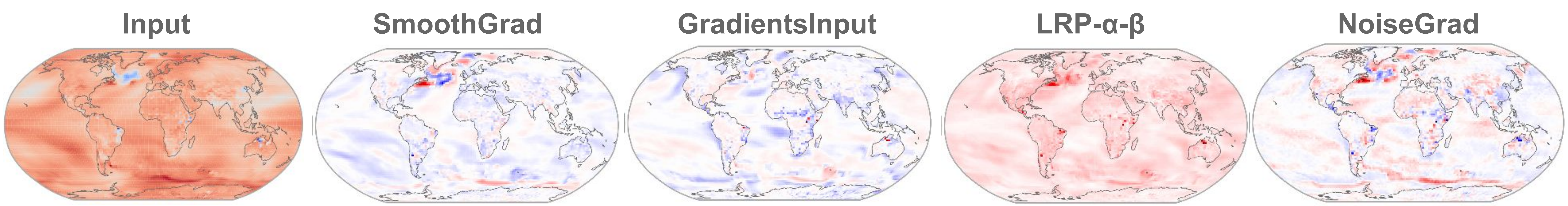
Dikshit et. al., 2021

The Challenge of XAI Method Selection

❖ Increasing number of methods with often **no ground-truth**

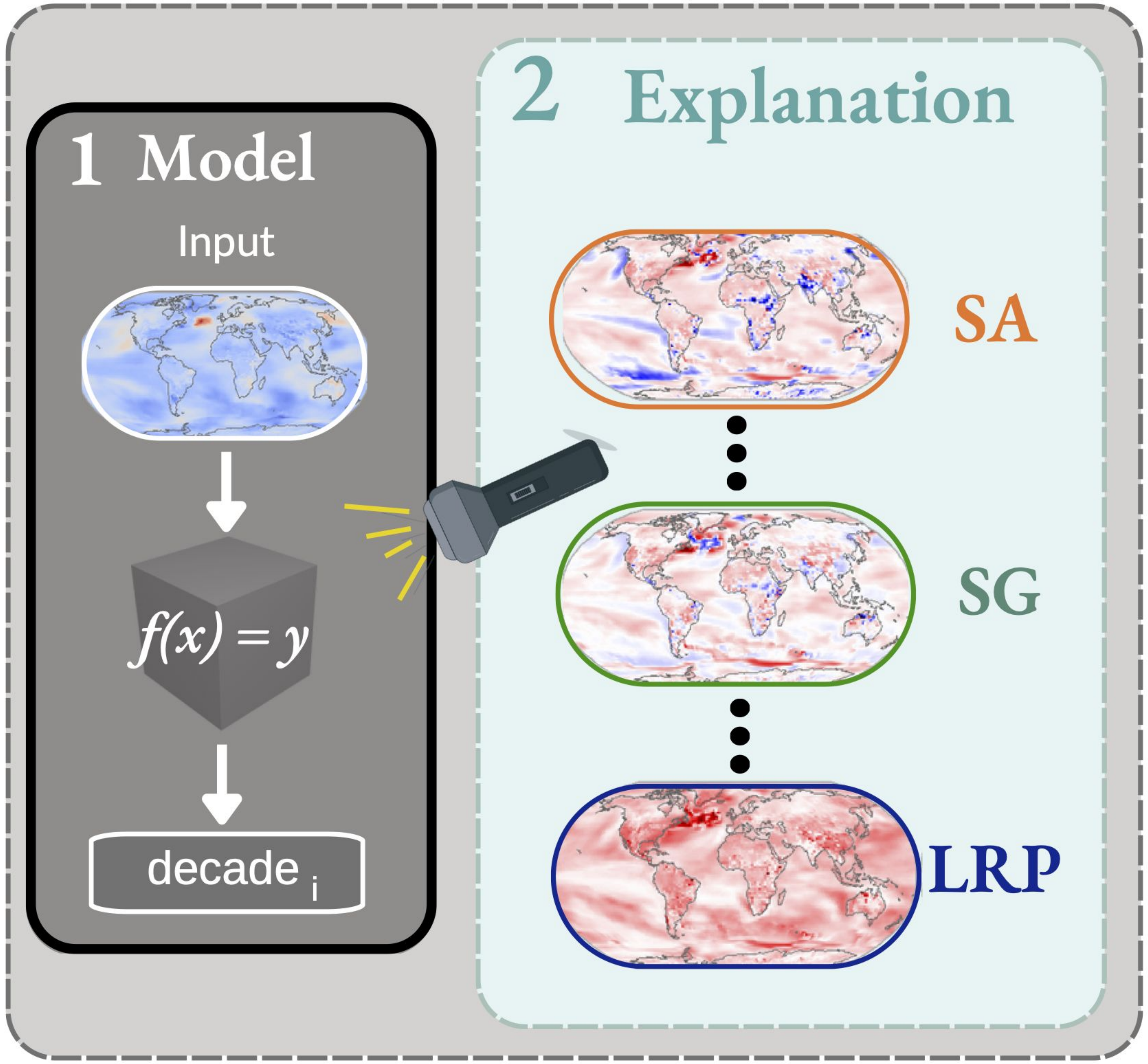
➢ Choice by popularity or easy access (Krishna et. al. (2022))

❖ different explanations for the same network decision lead to different conclusions → **lack of trust and reliability**



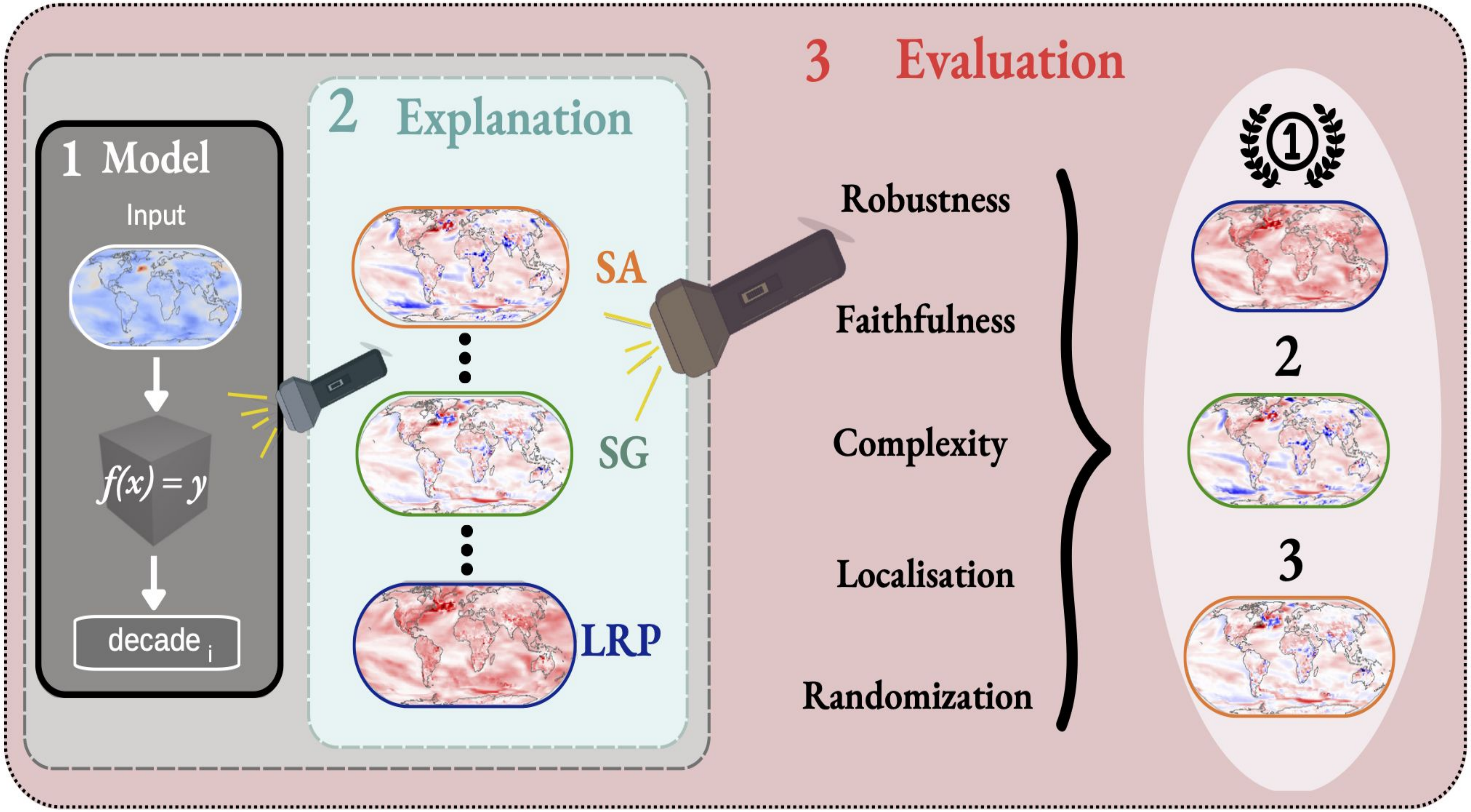
Bommer et. al. (2023)

2. Climate XAI task



Schematic of the XAI evaluation procedure (Bommer et. al., 2023)

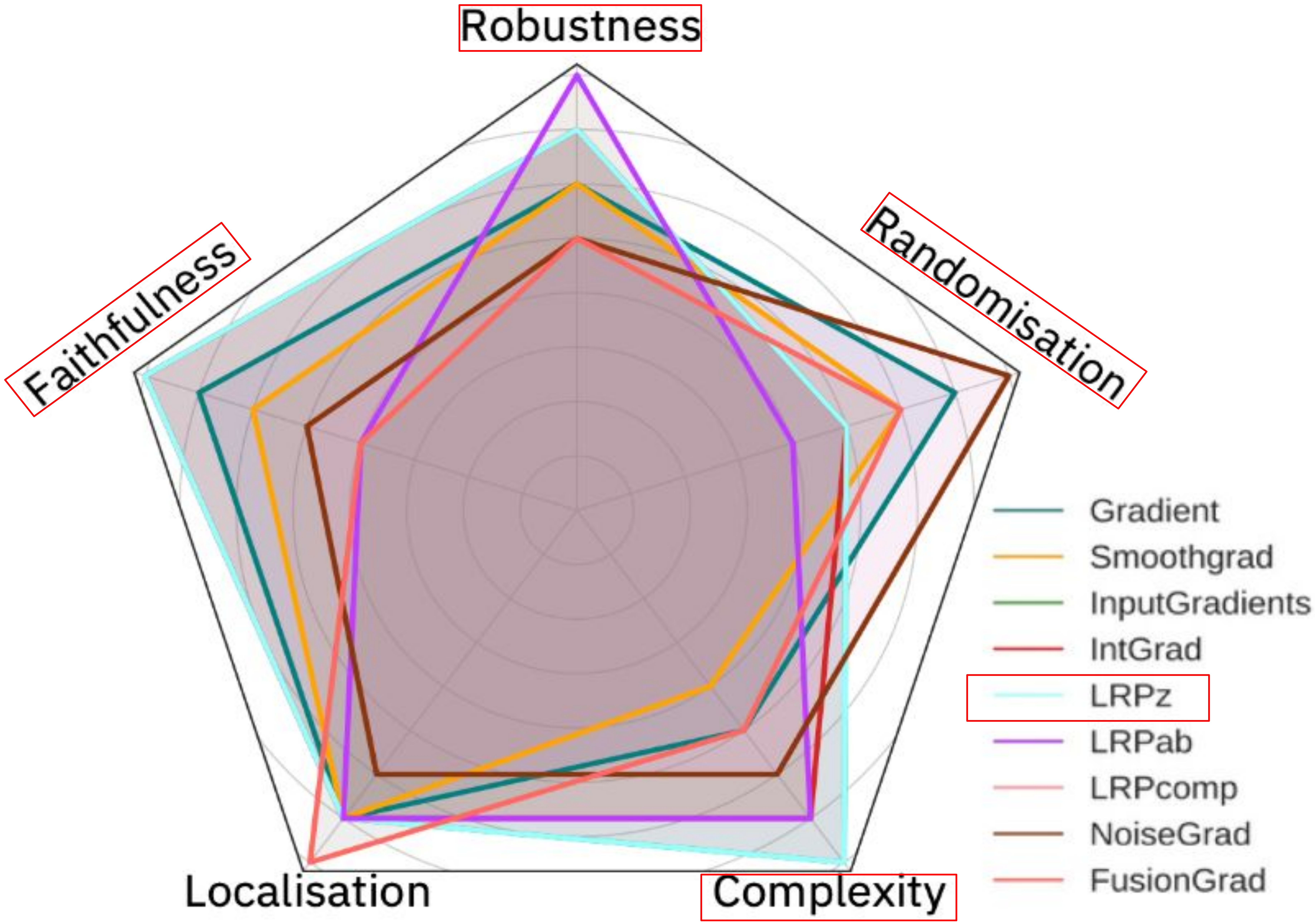
3. XAI evaluation



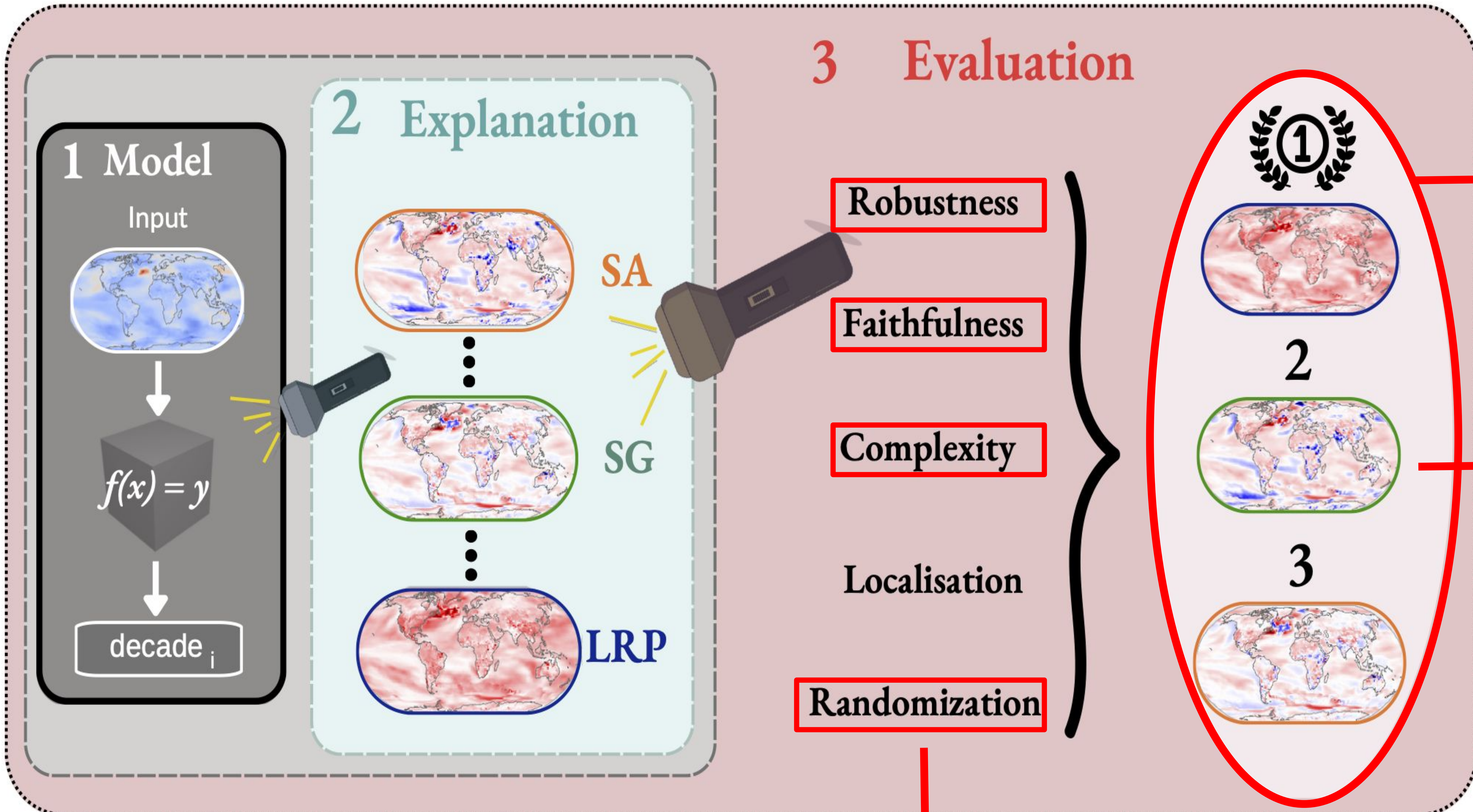
Schematic of the XAI evaluation procedure (Bommer et. al., 2023)

4. XAI Method Selection

1. Choose properties
2. Calculate scores
3. Rank scores
4. Choose XAI method



5. Summary



Determine reliable XAI method

Strengths & Weaknesses

AI task-specific

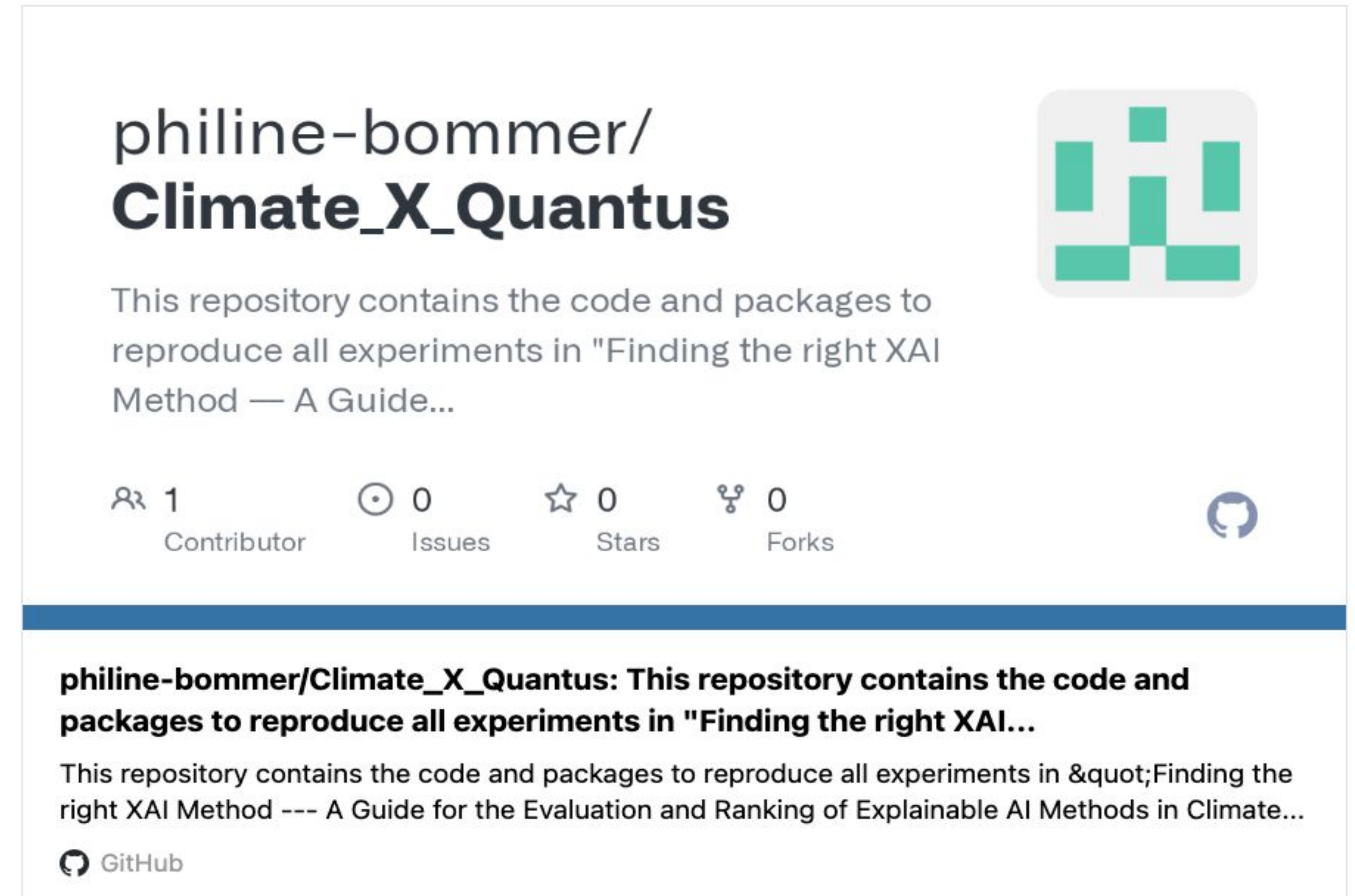
Resources

Tutorial

- https://colab.research.google.com/drive/1RW4jRCtjL1zx5Cm6cphtHmWFIw30gRM1

Github

- https://github.com/philine-bommer/Climate_X_Quantus
- https://github.com/understandable-machine-intelligence-lab/Quantus



References

- Bommer et al. (2023), https://arxiv.org/abs/2303.00652
- Hedström et al (2023a) https://jmlr.org/papers/v24/22-0142.html
- Labe and Barnes (2021) https://agupubs.onlinelibrary.wiley.co<m/doi/full/10.1029/2021MS002464

Bommer, P., Kretschmer, M., Hedstroem, A., Bareeva, D., and Hoehne, M. M.-C.: Evaluation of explainable AI solutions in climate science, EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-12528, https://doi.org/10.5194/egusphere-egu23-12528

A2. XAI Evaluation - Properties

Measure Explanation Quality

- **Faithfulness** (↑) quantifies to what extent explanations follow the predictive behaviour of the model, asserting that more important features affect model decisions more strongly e.g., (Bach et al., 2015; Dasgupta et al., 2022).
- **Robustness** (↓) measures to what extent explanations are stable/ similar when subjected to slight input perturbations, assuming an approximately constant model output e.g., (Alvarez-Melis et al., 2018; Yeh et al., 2019).
- **Randomisation** (↓) tests to what extent explanations deteriorate as labels or model parameters gets randomised e.g., (Adebayo et. al., 2018); Sixt et al., 2020).
- **Localisation** (↑) tests if the explainable evidence is centred around a region of interest, e.g., defined through a bounding box, a segmentation mask or a cell within a grid e.g., (Zhang et al., 2018; Arras et al., 2021).
- **Complexity** (↓) captures to what extent explanations are concise, i.e., that few features are used to explain a model prediction e.g., (Chalasani et al., 2020; Bhatt et al., 2020).

A2. XAI Evaluation - Quantus

Goals & Applications

- Quantus is an XAI toolkit for responsible **evaluation** of neural network explanations, for ML practitioners
- Quantus has been used for various healthcare applications [1,2,3,4], XAI optimisation [5], climate science [6, 7, 8]

Library Content

- Providing 30+ metrics in 6 categories for XAI evaluation with tutorials and API reference
- Supporting different data types (image, time-series, tabular/ NLP) and ML frameworks models (PyTorch and Tensorflow)
- Additional built-in XAI methods support

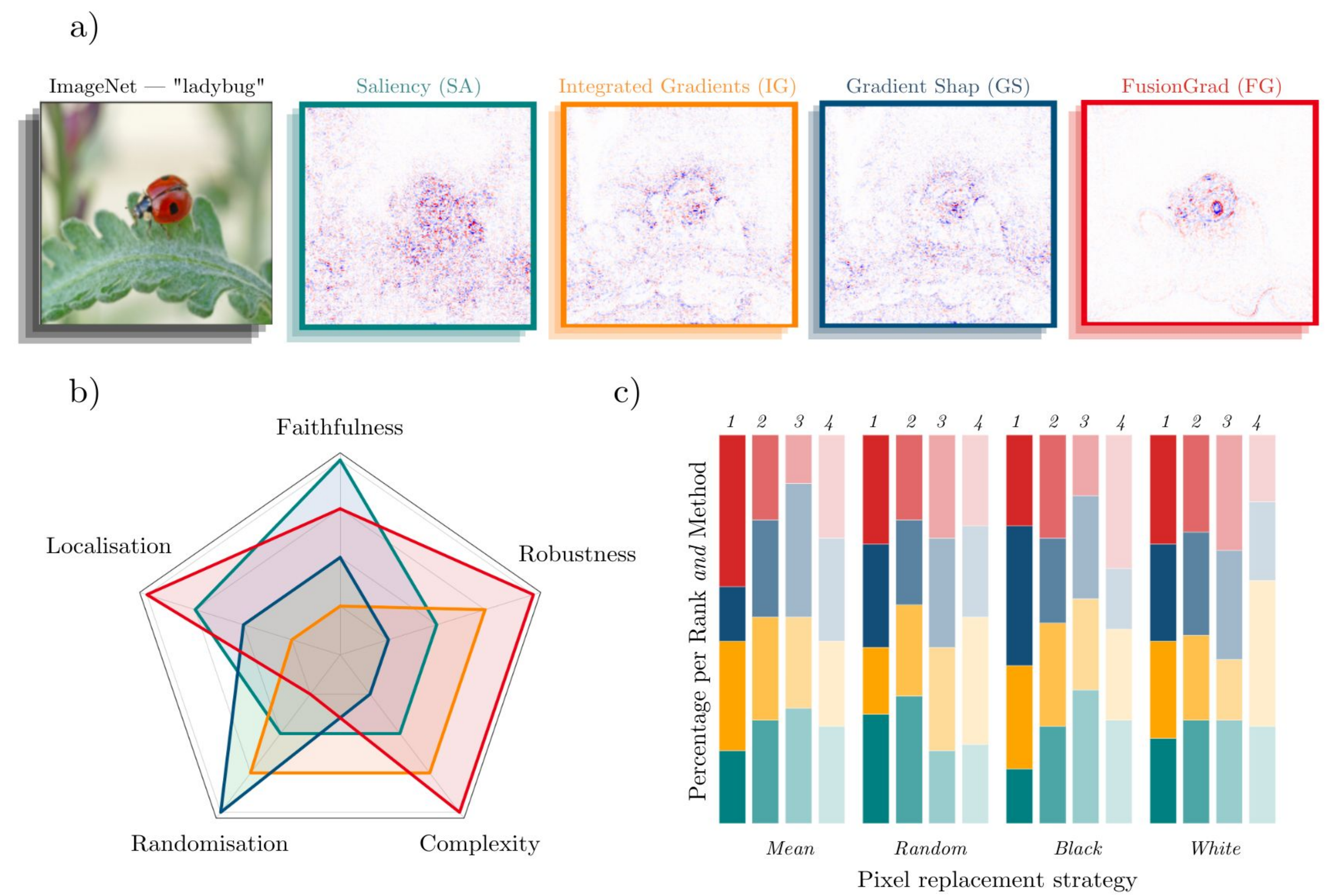


Figure: a) Simple qualitative comparison of XAI methods is often not sufficient to distinguish which gradient-based method — Saliency, Integrated Gradients, GradientShap or FusionGrad is preferred. With Quantus, we can obtain richer insights on how the methods compare b) by holistic quantification on several evaluation criteria and c) by providing sensitivity analysis of how a single parameter, e.g., pixel replacement strategy of a faithfulness test influences the ranking of XAI methods.

5) Comprehensive evaluation

- *Quantus* package (Hedström et al., 2023)
- metric functions for each category (Bommer et. al., 2023)
  - *Local Lipschitz Estimate* (Robustness), *Faithfulness Correlation* (Faithfulness), *Model Parameterization Test* (Randomisation), *Sparseness* (Complexity), *Relevance Rank Accuracy* (Localisation)
- **ranking**: normalized mean score and SEM across 50 random explanation samples (Bommer et. al., 2023)

	Robustness		*Faithfulness*		*Randomisation*		*Complexity*		*Localisation*	
	MLP	**CNN**	**MLP**	**CNN**	**MLP**	**CNN**	**MLP**	**CNN**	**MLP**	**CNN**
FusionGrad	*4.*	*5.*	*5.*	5.	3.	**1.**	4.	3.	**1.**	**1.**
InputGradients	2.	3.	**1.**	**1.**	4.	4.	**1.**	2.	2.	4.
Integrated Gradients	2.	3.	**1.**	**1.**	4.	4.	2.	2.	2.	2.
LRP-z	2.	3.	**1.**	**1.**	4.	4.	**1.**	2.	2.	4.
SmoothGrad	3.	3.	3.	3.	2.	2.	*5.*	3.	2.	2.
LRP-α-β	**1.**	2.	*5.*	*7.*	*5.*	*5.*	2.	*4.*	2.	3.
NoiseGrad	*4.*	4.	4.	4.	**1.**	2.	3.	3.	2.	5.
Gradient	3.	3.	2.	2.	2.	3.	4.	3.	2.	4.
LRP-composite	–	**1.**	–	6.	–	4.	–	**1.**	–	*6.*

Bommer et. al. (2023)

5) XAI Method Selection

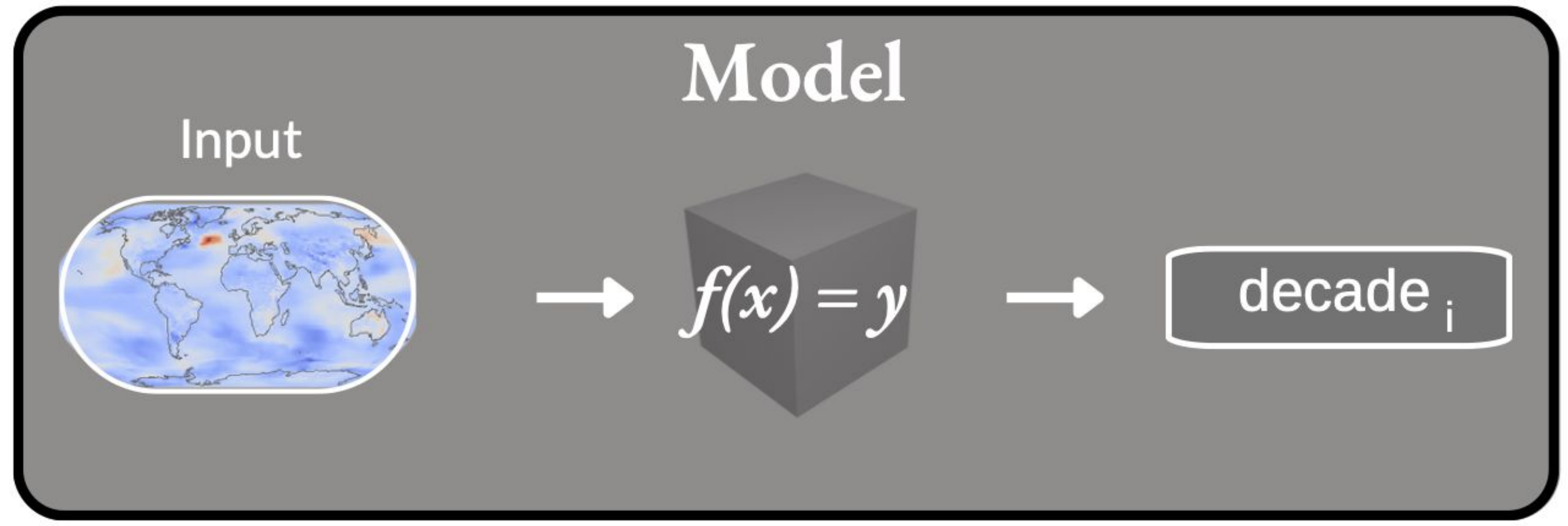
1. Choose evaluation properties for the task
2. Calculate scores for all methods and each chosen property (*Quantus*)
3. Rank explanation methods
4. Choose best ranked explanation method

	Robustness	*Faithfulness*	*Randomisation*	*Complexity*	*Localisation*
FusionGrad	*4.*	*5.*	3.	4.	**1.**
InputGradients	2.	**1.**	4.	**1.**	2.
Integrated Gradients	2.	**1.**	4.	2.	2.
LRP-z	2.	**1.**	4.	**1.**	2.
SmoothGrad	3.	3.	2.	*5.*	2.
LRP-α-β	**1.**	*5.*	*5.*	2.	2.
NoiseGrad	*4.*	4.	**1.**	3.	2.
Gradient	3.	2.	2.	4.	2.
LRP-composite	–	–	–	–	–

2. Climate XAI task

Task

❖ **Classification** of annual temperature maps based on their decade ([Bommer et. al. (2023)](#))



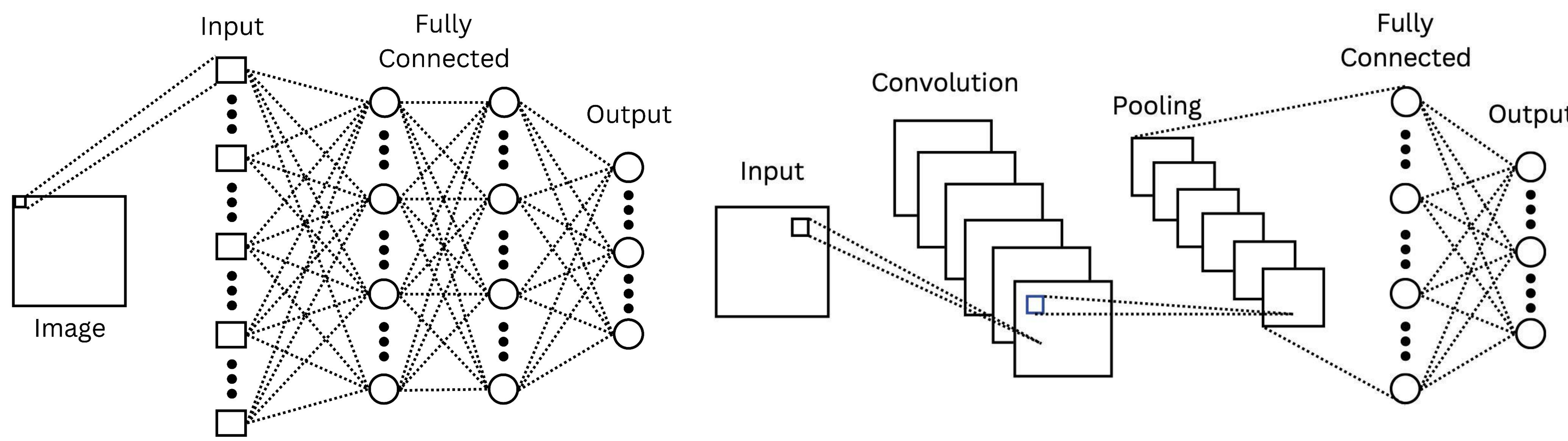
Data

❖ Standardized, annual, 2-m air temperature (T2m) temperature maps from 1920-2080 (Hurrell et al. (2013))

Network

❖ multi-layer perceptron (MLP) (Labe and Barnes (2021))

❖ Convolutional neural network (CNN)



MLP CNN

XAI Methods

❖ Several **local** explanation methods:

➢ *Gradients, SmoothGrad, InputGradients, Integrated Gradients, LRP-a-b, LRP-z, LRP-comp, NoiseGrad, FusionGrad*

European Geosciences Union