Finding the right XAI method — a guide for the evaluation and ranking of explainable AI methods in climate science

Article

Accepted Version

Finding the right XAI Method — A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science

Philine Lou Bommer[a,b], Marlene Kretschmer[c,d], Anna Hedström[a,b], Dilyara Bareeva[a,b], Marina M.-C. Höhne[a,b,e,f,g]

[a] *Understandable Machine Intelligence Lab, TU Berlin, Berlin, Germany*

[b] *Department of Data Science, ATB, Potsdam, Germany*

[c] *Leipzig Institute for Meteorology, University of Leipzig, Leipzig, Germany*

[d] *Department of Meteorology, University of Reading, Reading, UK*

[e] *Institute of Computer Science - University of Potsdam, Potsdam, Germany*

[f] *BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany*

[g] *Machine Learning Group, UiT the Arctic University of Norway, Tromsø, Norway*

Corresponding author: Philine Bommer, philine.l.bommer@tu-berlin.de

Early Online Release: This preliminary version has been accepted for publication in *Artificial Intelligence for the Earth Systems*, may be fully cited, and has been assigned DOI 10.1175/AIES-D-23-0074.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: Explainable artificial intelligence (XAI) methods shed light on the predictions of machine learning algorithms. Several different approaches exist and have already been applied in climate science. However, usually missing ground truth explanations complicate their evaluation and comparison, subsequently impeding the choice of the XAI method. Therefore, in this work, we introduce XAI evaluation in the climate context and discuss different desired explanation properties, namely robustness, faithfulness, randomization, complexity, and localization. To this end, we chose previous work as a case study where the decade of annual-mean temperature maps is predicted. After training both a multi-layer perceptron (MLP) and a convolutional neural network (CNN), multiple XAI methods are applied and their skill scores in reference to a random uniform explanation are calculated for each property. Independent of the network, we find that XAI methods Integrated Gradients, layer-wise relevance propagation, and input times gradients exhibit considerable robustness, faithfulness, and complexity while sacrificing randomization performance. Sensitivity methods – gradient, SmoothGrad, NoiseGrad, and FusionGrad, match the robustness skill but sacrifice faithfulness and complexity for randomization skill. We find architecture-dependent performance differences regarding robustness, complexity and localization skills of different XAI methods, highlighting the necessity for research task-specific evaluation. Overall, our work offers an overview of different evaluation properties in the climate science context and shows how to compare and benchmark different explanation methods, assessing their suitability based on strengths and weaknesses, for the specific research problem at hand. By that, we aim to support climate researchers in the selection of a suitable XAI method.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

SIGNIFICANCE STATEMENT: Explainable artificial intelligence (XAI) helps to understand the reasoning behind the prediction of a neural network. XAI methods have been applied in climate science to validate networks and provide new insight into physical processes. However, the increasing number of XAI methods' can overwhelm practitioners making it difficult to choose an explanation method. Since XAI methods results can vary, uninformed choices might cause misleading conclusions about the network decision. In this work, we introduce XAI evaluation to compare and assess the performance of explanation methods based on five desirable properties. We demonstrate that XAI evaluation reveals the strengths and weaknesses of different XAI methods. Thus, our work provides climate researchers with the tools to compare, analyze, and subsequently choose explanation methods.

1. Introduction

Deep learning (DL) has become a widely used tool in climate science and assists various tasks, such as nowcasting (Shi et al. 2015; Han et al. 2017; Bromberg et al. 2019), climate or weather monitoring (Hengl et al. 2017; Anantrasirichai et al. 2019) and forecasting (Ham et al. 2019; Chen et al. 2020; Scher and Messori 2021), numerical model enhancement (Yuval and O'Gorman 2020; Harder et al. 2021), and up-sampling of satellite data (Wang et al. 2021; Leinonen et al. 2021). However, a deep neural network (DNN) is mostly considered a black box due to its inaccessible decision-making process. This lack of interpretability limits their trustworthiness and application in climate research, as DNNs should not only have high predictive performance but also provide accessible and consistent predictive reasoning aligned with existing theory (McGovern et al. 2019; Mamalakis et al. 2020; Camps-Valls et al. 2020; Sonnewald and Lguensat 2021; Clare et al. 2022; Flora et al. 2022). Explainable artificial intelligence (XAI) aims to address the lack of interpretability by explaining potential reasons behind the predictions of a network. In the climate context, XAI can help to validate DNNs and on a well-performing model provide researchers with new insights into physical processes (Ebert-Uphoff and Hilburn 2020; Hilburn et al. 2021). For example, Gibson et al. (2021) demonstrated using XAI that DNNs produce skillful seasonal precipitation forecasts based on known relevant physical processes. Moreover, XAI was used to improve the forecasting of droughts (Dikshit and Pradhan 2021), teleconnections (Mayer and Barnes 2021), and regional precipitation (Pegion et al. 2022), to assess external drivers of global climate change (Labe and

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

TABLE 1: Overview and categorization of research on the transparency and understandability of neural networks. For this categorization we follow works like Samek et al. (2019); Ancona et al. (2019); Mamalakis et al. (2022b); Letzgus et al. (2022); Flora et al. (2022)

post-hoc	explanation target	**local** (e.g. Shapley values (Lundberg and Lee 2017) or LRP (Bach et al. 2015))
		global (e.g. activation-maximization (Simonyan et al. 2014) or DORA (Bykov et al. 2022a))
	components	**model-aware** (e.g. gradient (Baehrens et al. 2010) ,LRP (Montavon et al. 2019) or Grad-CAM (Selvaraju et al. 2017))
		model-agnostic (e.g. LIME (Ribeiro et al. 2016) or Shapley values (Lundberg and Lee 2017))
	explanation output	**sensitivity** (e.g. gradient (Baehrens et al. 2010) and GradCAM (Selvaraju et al. 2017))
		feature contribution— salience —attribution (e.g. Integrated Gradients (Sundararajan et al. 2017) or LRP (Bach et al. 2015))
		examples (e.g. RISE (Petsiuk et al. 2018))
ante-hoc	self-explaining network	prototype network (Chen et al. 2019; Gautam et al. 2022, 2023)
		concept networks (Alvarez Melis and Jaakkola 2018)
		contrastive networks (Sawada and Nakamura 2022)

Barnes 2021) and to understand sub-seasonal drivers of high-temperature summers (Van Straaten et al. 2022). Additionally, Labe and Barnes (2022) showed that XAI applications can aid in the comparative assessment of climate models.

Generally, explainability methods can be divided into ante-hoc and post-hoc approaches (Samek et al. 2019) (see Table 1). Ante-hoc approaches modify the DNN architecture to improve interpretability, like adding an interpretable prototype layer to learn humanly understandable representations for different classes (see e.g. Chen et al. (2019) and Gautam et al. (2022, 2023)) or constructing mathematically similar but interpretable models (Hilburn 2023). Such approaches are also called self-explaining neural networks and link to the field of interpretability (Flora et al. 2022). Post-hoc

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

XAI methods, on the other hand, can be applied to any neural network architecture (Samek et al. 2019) and here we focus on three characterizing aspects (Samek et al. 2019; Letzgus et al. 2022; Mamalakis et al. 2022b), as shown in Table 1. The first considers the explanation target (i.e. what is explained) which can differ between local and global decision-making. While local explanations provide explanations of the network decision for a single data point (Baehrens et al. 2010; Bach et al. 2015; Vidovic et al. 2016; Ribeiro et al. 2016), e.g., by assessing the contribution of each pixel in a given image based on the predicted class, global explanations reveal the overall decision strategy, e.g. by showing a map of important features or image patterns, learned by the model for the whole class (Simonyan et al. 2014; Vidovic et al. 2015; Nguyen et al. 2016; Lapuschkin et al. 2019; Grinwald et al. 2022; Bykov et al. 2022a). The second aspect concerns the components used to calculate the explanation, differentiating between model-aware and model-agnostic methods. Model-aware methods use components of the trained model for the explanation calculation, such as network weights, while model-agnostic methods consider the model as a black box and only assess the change in the output caused by a perturbation in the input (Strumbelj and Kononenko 2010; Ribeiro et al. 2016). The third aspect considers the DNN explanation output. Here we can differentiate between methods where the assigned value of a pixel indicates the sensitivity of the network regarding that pixel also called *sensitivity methods*, such as absolute gradient, as well as methods, that display the positive or negative contribution of a pixel to predict the class, such as layer-wise Relevance Propagation (see Section 3c) also called *salience methods*, and methods presenting input examples leading to the same prediction. Beyond these three characteristics, recent efforts (Flora et al. 2022) also differentiate between feature importance methods encompassing mostly global methods, which calculate feature contribution based on the network performance (e.g. accuracy), and feature relevance methods describing mostly local methods which calculate contributions to the model prediction. In climate research, the decision patterns learned by DNNs have been analyzed with local explanation methods such as LRP or Shapley values (Gibson et al. 2021; Dikshit and Pradhan 2021; Mayer and Barnes 2021; Labe and Barnes 2021; He et al. 2021; Felsche and Ludwig 2021; Labe and Barnes 2022). However, different local explanation methods can identify different input features as being important to the network decision, subsequently leading to different scientific conclusions (Lundberg and Lee 2017; Han et al. 2022; Flora et al. 2022). Thus, with the increasing number of XAI methods available, selecting the most suitable

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

method for a specific task poses a challenge and the practitioner's choice of a method is often based upon popularity or upon easy-access (Krishna et al. 2022). To navigate the field of XAI, recent climate science publications have compared and assessed different explanation techniques using benchmark datasets, where the XAI method was assessed by comparing its predictions with a defined target, considered as ground truth (Mamalakis et al. 2022b,a). While benchmark datasets (Yang and Kim 2019; Arras et al. 2020; Agarwal et al. 2022) certainly contribute to the understanding of local XAI methods, the existence of a ground truth explanation is highly debated (e.g., Janzing et al. (2020); Sturmfels et al. (2020)). In the case of DNNs, ground truth explanation labels can only be considered approximations and are not guaranteed to align precisely with the model's decision process or the features it utilizes (Ancona et al. 2019; Hedström et al. 2023a). For exact ground truth, either perfect knowledge of how the model handles the available information or a carefully engineered model would be required, which is usually not the case. Additionally, post-hoc explanation methods are generally only approximations of a model's behavior (Lundberg and Lee 2017; Han et al. 2022), and the distinct mathematical concepts of the different methods would consequently lead to distinct ground truth explanations.

Here, we address these challenges, by introducing *XAI evaluation* in the context of climate science to compare different local explanation methods. The field of XAI evaluation has emerged recently and refers to the development of metrics to compare, benchmark, and rank explanation methods, in different explainability contexts (e.g. Adebayo et al. (2018); Hedström et al. (2023b,a)). As discussed below in more detail, using evaluation metrics we are able to quantitatively assess the robustness, complexity, localization, randomization, and faithfulness of explanation methods, making them comparable regarding their suitability, their strengths, and weaknesses (Hoffman et al. 2018; Arrieta et al. 2020; Mohseni et al. 2021; Hedström et al. 2023b).

In this work, we discuss these properties in an exemplary manner and build upon work from (Labe and Barnes 2021). In their work, an MLP was trained with global annual temperature anomaly maps and the network´s task was to assign the respective year or decade of occurrence. The MLP achieves the assignment, as global-mean warming progresses. Using layer-wise relevance propagation (LRP) they then identified the signals relevant to the network´s decision and found the North Atlantic, Southern Ocean, and Southeast Asia as key regions. Here, we use their work

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

as a case study and train an MLP as well as a CNN for the same prediction tasks (see step 1 in Fig. 1). Then, we apply several explanation methods and show the variation in their explanation maps, potentially leading to different scientific insights (step 2 in Figure 1). We therefore introduce XAI evaluation metrics and quantify the skill of the different XAI methods against a random baseline in different properties to compare their performance with respect to the underlying task.

This paper is structured as follows. In Section 2 we discuss the used dataset and network types, and briefly describe the different analysed explanation methods. Section 3 introduces XAI evaluation and describe five evaluation properties. Then, in Section 4, we first discuss the performance of both network types and provide a motivational example highlighting the risks of an uninformed choice of an explanation method. Next, we evaluate different XAI methods applied to the MLP, using two different metrics for each evaluation property, and then compare the XAI evaluation results for the different networks (see Section 4b and c). Finally, in Section 4d, we present a guideline on using XAI evaluation to choose a suitable XAI method. The discussion of our results and our conclusion are presented in Section 5.

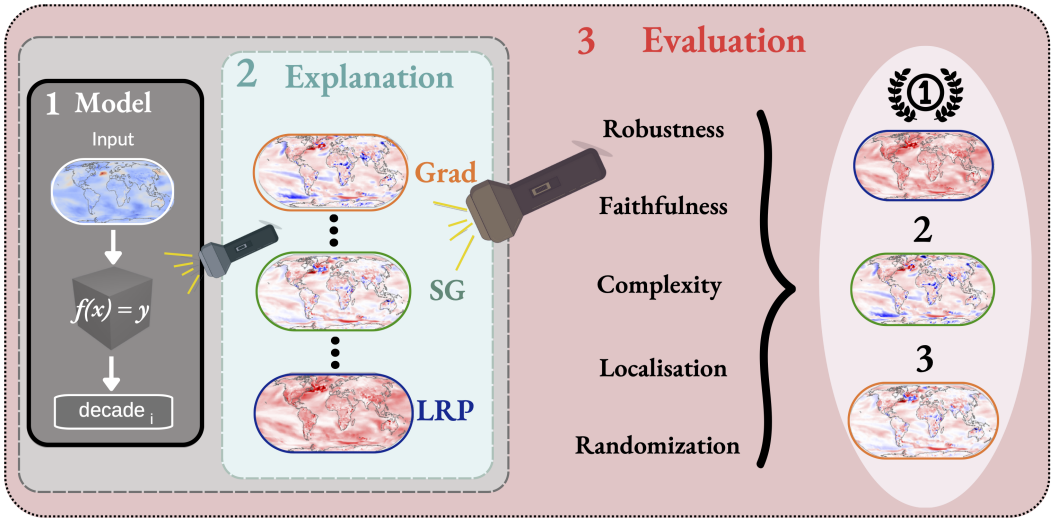

FIG. 1: Schematic of the XAI evaluation procedure. Based on an annual temperature anomaly map as input, the network predicts the respective decade (box 1). The explanation methods (Grad - gradient, SG - SmoothGrad applied to gradient, LRP - layer-wise relevance propagation) then provide insights (i.e., "shine a light", see box 2) into the specific network's decision. The different explanation maps (marked in orange - Grad, green - SG, and blue - LRP) highlight different areas as positively (red) and negatively (blue) contributing to the network decision. Here XAI evaluation can 'shine a light' on the explanation methods and help choose a suitable method (here indicated by the first rank) since evaluation explores the explanation maps regarding their robustness, faithfulness, localization, complexity, and randomization properties.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

2. Data and Methods

a. Data

We analyze data simulated by the general climate model, CESM1 (Hurrell et al. 2013), focusing on the "ALL" configuration (Kay et al. 2015), which is discussed in detail in Labe and Barnes (2021). We use the global 2-m air temperature (T2m) maps from 1920 to 2080. The data $\mathbf{\Omega}$ consist of I = 40 ensemble members $\Omega_{i\in\{1,\ldots,I\}}$, and each member is generated by varying the atmospheric initial conditions z_i with fixed external forcing θ_{clima}. Following Labe and Barnes (2021), we compute annual averages and apply a bilinear interpolation. This results in $T = 161$ temperature maps for each member $\Omega_i \in \mathbb{R}^{T\times v\times h}$, with $v = 144$ and $h = 95$ denoting the number of longitudes and latitudes, with 1.9° sampling in latitude and 2.5° sampling in longitude. Accordingly, the whole dataset $\mathbf{X} \in \mathbb{R}^{\mathrm{I}\times T\times v\times h}$ contains $\mathrm{I}\times T$ samples. The data is split into a training Ω_{tr} and a test set Ω_{test}. More precisely, we sample 20% of the ensemble members (i.e., in total 8 ensemble members) as a test set $\mathbf{X}_{\text{test}} \in \mathbb{R}^{0.2\mathrm{I}\times T\times v\times h}$, and use the remaining 80% (i.e., 32 ensemble members) for training and validation. Of these 32 ensemble members all temperature maps are split into a training (80% of the data points, i.e. 64% of all ensemble members) and validation (20% of the temperature maps, i.e.16% of all ensemble members) set. All temperature maps $x \in \mathbb{R}^{vxh}$ are standardized by subtracting the mean and subsequently dividing by the corresponding standard deviation at each grid-point individually, whereby the mean $\mathbf{x}_{\text{mean}} \in \mathbb{R}^{v\times h}$ and standard deviation $\mathbf{x}_{\text{std}} \in \mathbb{R}^{v\times h}$ are computed over the training set only.

b. Networks

Following Labe and Barnes (2021), we train an MLP, $f_{\text{MLP}} : \mathbb{R}^d \rightarrow \mathbb{R}^c$ with network weights $W \in \mathcal{W}$, to solve a fuzzy classification problem by combining classification and regression. As input $\mathbf{x} \in \Omega$, the network considers the flattened temperature maps with dimensionality $d = v \times h$. Given the goal of fuzzy classification, first, the network assigns each map to one of the $C = 20$ different classes, where each class corresponds to one decade between 1900 and 2100 (see Figure 1 in Labe and Barnes (2021)). The network output $f(x)$, thus, is a probability vector $\mathbf{y} \in \mathbb{R}^{1\times C}$ across $C = 20$ classes. Afterward, since the network can assign a nonzero probability to more than

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

one class, regression is used to predict the year $\hat{y}$ of the input as:

$$\hat{y} = \sum_{i=1}^{C} y_i \bar{Y}_i, \tag{1}$$

where y_i is the probability of class i, predicted by the network $\mathbf{y} = f(\mathbf{x})$ in the classification step, and $\bar{Y}_i$ denotes the central year of the corresponding decade class i (e.g. for class $i = 1$, $\bar{Y}_1 = 1925$ represents the decade $1920 - 1929$). Accordingly, the task ensures the association of temperature patterns to the respective year or decade. Here we train using the binary cross-entropy loss, considering Eq. (1) only for performance evaluation.

Additionally, we construct a CNN $f_{\text{CNN}} : \mathbb{R}^{v \times h} \rightarrow \mathbb{R}^c$ that maintains the longitude-latitude grid of the data $\mathbf{x}_{\text{img}} \in \mathbb{R}^{v \times h}$ for each input sample (see Section 2a), unlike the flattened input used for the MLP. The CNN consists of a 2D-convolutional layer (2dConv) with 6×6 window size and a stride of 2. The second layer includes a max-pooling layer with a 2×2 window size, followed by a dense layer with L^2-regularization and a softmax output layer.

c. Explainable Artificial Intelligence (XAI)

In this work, we focus on local model-aware explanation methods belonging to the group of feature-attribution methods (Ancona et al. 2019; Das and Rad 2020; Zhou et al. 2022). For the mathematical details, we refer to Appendix A-a.

(i) Gradient (Baehrens et al. 2010) explains the network decision by computing the first partial derivative of the network output $f(\mathbf{x})$ with respect to the input. This explanation method feeds backward the network's prediction to the features in the input $\mathbf{x}$, indicating the change in network prediction given a change in the respective features. The explanation values correspond to the network's *sensitivity* to each feature, thus belonging to the group of sensitivity methods. The absolute gradient, often referred to as Saliency map, can also be used as an explanation (Simonyan et al. 2014).

(ii) Input times gradient is an extension of the gradient method and computes the product of the gradient and the input. In the explanation map, a high relevance is assigned to an input feature if it has a high value and the model gradient is sensitive to it. Therefore, contrary to the gradient as a sensitivity method, input times gradient and other methods including the input pixel value are

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

considered salience methods Ancona et al. (2019) (or attribution methods, e.g. Mamalakis et al. (2022a)).

(iii) Integrated Gradients (Sundararajan et al. 2017) extends input times gradient, by integrating a gradient along a line path from a baseline (generally a reference vector for which the network's output is zero, e.g. all zeros for standardized data) to the explained sample **x**. In practice, the gradient explanations of a set of images lying between the baseline and **x** are averaged and multiplied by the difference between the baseline and the explained input (see Eq. (A3)). Hence, the Integrated Gradients method is a salience method and highlights the difference between the features important to the prediction of **x** and features important to the prediction of the baseline value.

(iv) Layerwise Relevance Propagation(LRP) (Bach et al. 2015; Montavon et al. 2019) computes the relevance for each input feature by feeding the network's prediction backward through the model, layer by layer, until the prediction score is distributed over the input features and is a salience method. Different propagation rules can be used, all resembling the energy conservation rule, i.e., the sum of all relevances within one layer is equal to the original prediction score. In case of the α-β**-rule** relevance is assigned at each layer to each neuron. All positively contributing activations of connected neurons in the previous layer are weighted by α, while β is used to weight the contribution of the negative activations. The default values are $\alpha = 1$ and $\beta = 0$, where only positively contributing activations are considered. Contrary to that, the **z-rule** calculates the explanation by including both negative and positive neuron activations. Hence, the corresponding explanations, visualized as heatmaps, display both positive and negative evidence. The **composite rule** combines various rules for different layer types. The method accounts for layer structure variety in CNNs, such as fully connected, convolutional, and pooling layers.

(v) SmoothGrad (Smilkov et al. 2017) aims to filter out the background noise (i.e., the gradient shattering effect, where gradients resemble white noise with increasing layer number (Balduzzi et al. 2017)) to enhance the interpretability of the explanation. To this end, multiple noisy samples are generated by adding random noise to the input, then the explanations of the noisy samples are computed and averaged, such that the most important features are enhanced and the less important features are "canceled out".

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

(vi) NoiseGrad (Bykov et al. 2022b) perturbs the weights of the model, instead of the input feature as done by SmoothGrad. The explanations, resulting from explaining the predictions made by the noisy versions of the model on the same image, are averaged to suppress the background noise of the image in the final explanation.

(vii) FusionGrad (Bykov et al. 2022b) combines SmoothGrad and NoiseGrad by perturbing both the input features and the network weights. The purpose of the method is to account for uncertainties within the network and the input space (Bykov et al. 2021).

(viii) Deep SHapley Additive exPlanations (DeepSHAP) (Lundberg and Lee 2017) estimates Shapley values for the full DNN by dividing it into small network components, calculating the Shapley values, and averaging them across all components. The idea behind SHAP (SHapley Additive exPlanations) values is to fairly distribute the contribution of each feature to the prediction of a specific instance considering all possible feature combinations. Following the game-theoretic concept of Shapley values (Shapley 1951), DeepSHAP explanations satisfy properties such as local accuracy, missingness, and consistency (Lundberg and Lee 2017) and is a salience method.

In this work, we maintain literature values for most hyperparameters of the explanation methods. Exceptions are hyperparameters of explanation methods NoiseGrad, and FusionGrad. We adjust the perturbation levels of the parameters, as discussed in Bykov et al. (2022b) to ensure at most 5% loss in accuracy. All hyperparameters are presented in Table B1 (see Appendix B-a). Additionally, both Integrated Gradients and DeepSHAP require background images as reference points to calculate the explanations (see also Lundberg and Lee (2017) and Appendix A-a). To allow for a fair performance comparison, for both methods we sample 100 maps containing all zero values. We note that there are other possible reference values, e.g., natural images from training, or all-one-maps, and this choice can affect the explanation performance. Lastly, the baseline for SmoothGrad, NoiseGrad, and FusionGrad can be any local explanation method, and here we use the gradient explanations. Accordingly, gradient, SmoothGrad, NoiseGrad, and FusionGrad are sensitivity methods.

3. Evaluation techniques

Due to the lack of a ground truth explanation, XAI research developed alternative metrics to assess the reliability of an explanation method. These evaluation metrics analyze different properties an

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

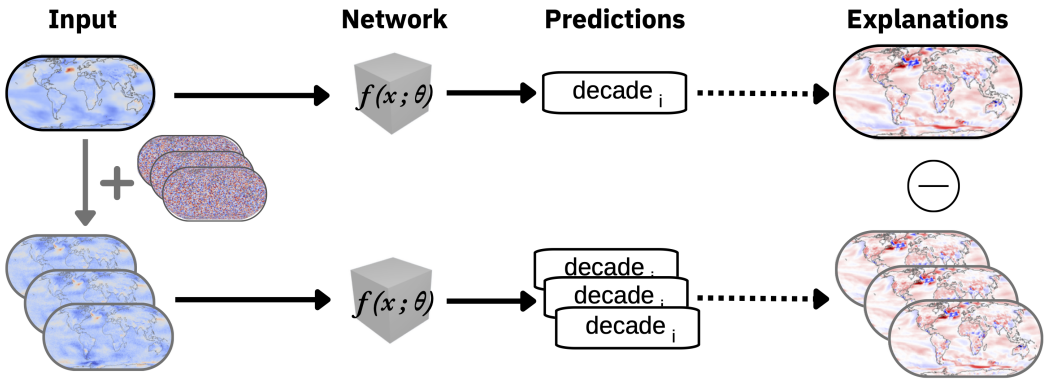

FIG. 2: Diagram of the concept behind the *robustness* property. Perturbed input images are created by adding uniform noise maps of small magnitude to the original temperature map (left part of Figure). The perturbed maps are passed to the network, resulting in an explanation map for each prediction. The explanation maps of the perturbed inputs (explanation maps with grey outlines) are then compared to (indicated by a minus sign) the explanation of the unperturbed input (explanation map with black outline). A robust XAI method is expected to produce similar explanations for the perturbed input and unperturbed inputs.

explanation method should fulfill and can serve to evaluate different explanation methods (Hoffman et al. 2018; Arrieta et al. 2020; Mohseni et al. 2021; Hedström et al. 2023b). Following Hedström et al. (2023b), we describe five different evaluation properties and based on the classification task from Labe and Barnes (2021) we illustrate each property in a schematic diagram (See Figures 2-4).

a. Robustness

Robustness measures the stability of an explanation regarding small changes in the input image $\mathbf{x}+\delta$ (Alvarez-Melis and Jaakkola 2018; Yeh et al. 2019; Montavon et al. 2018). Ideally, these small changes ($\delta < \epsilon$) in the input sample should produce only small changes in the model prediction and successively only small changes in the explanation (see Figure 2).

To measure robustness, we choose the Local Lipschitz Estimate $q_{\mathrm{LLE},m}$ (Alvarez-Melis and Jaakkola 2018) and the Average Sensitivity $q_{\mathrm{AS},m}$ (Yeh et al. 2019) as representative metrics. Both use Monte Carlo sampling-based approximation to measure the Lipschitz constant or the average sensitivity of an explanation. For an explanation $\Phi^m(f, c, \mathbf{x}) \in \mathbb{R}^d$ of a XAI method m and a given input $\mathbf{x}$, the scores are defined by:

$$q_{\mathrm{LLE},m} = \max_{\mathbf{x}+\delta \in \mathcal{N}_\epsilon(\mathbf{x}) \leq \epsilon} \frac{\|\Phi^m(f, c, \mathbf{x}) - \Phi^m(f, c, \mathbf{x}+\delta)\|_2}{\|\mathbf{x} - (\mathbf{x}+\delta)\|_2}, \tag{2}$$

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

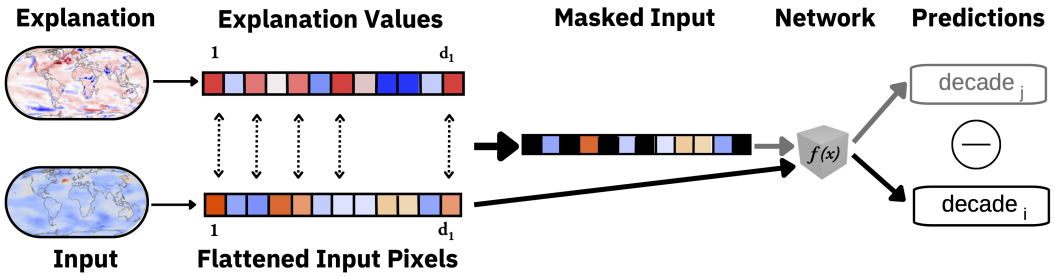

FIG. 3: Diagram of the concept behind the *faithfulness* property. Faithfulness assesses the impact of highly relevant pixels in the explanation map on the network decision. First, the explanation values are sorted to identify the highest relevance values (here shown in red). Next, the corresponding pixel positions in the flattened input temperature map are identified (see dotted arrows) and masked (marked in black); i.e., their value is set to a chosen masking value, such as 0 or 1. Both the masked and the original input maps are passed through the network and their predictions are compared. If the masking is based on a *faithful* explanation, the prediction of the masked input (j, grey) is expected to change compared to (indicated by a minus sign) the unmasked input (i, black), e.g., a different decade is predicted.

$$q_{\mathrm{AS},m} = \mathop{\mathbb{E}}_{\mathbf{x}+\delta \in \mathcal{N}_\epsilon(\mathbf{x}) \leq \epsilon} \left[\frac{\| (\Phi^m(f,c,\mathbf{x}) - \Phi^m(f,c,\mathbf{x}+\delta)) \|}{\|\mathbf{x}\|} \right], \tag{3}$$

where ϵ defines the discrete, finite-sample neighborhood radius $\mathcal{N}_\epsilon$ for every input $\mathbf{x} \in \mathbf{X}$, $\mathcal{N}_\epsilon(\mathbf{x}) = \{\mathbf{x}+\delta \in X \mid \|\mathbf{x} - (\mathbf{x}+\delta)\| \leq \epsilon\}$, and c denotes the true class of the input sample (for more details on intuition and calculation we also suggest the primary publications (Alvarez-Melis and Jaakkola 2018; Yeh et al. 2019)).

The robustness metrics assess the difference between the explanation of a true and perturbed image as can be seen in Eq. (2) and (3). Accordingly, the lowest score represents the highest robustness.

b. Faithfulness

Faithfulness measures whether changing a feature that an explanation method assigned high relevance to, changes the network prediction (see Figure 3). This can be examined through the iterative perturbation of an increasing number of input pixels corresponding to high-relevance values and subsequent comparison of each resulting model prediction to the original model prediction. Since explanation methods assign relevance to features based on their contribution to the network's prediction, changing high-relevance features should have a larger impact on the model prediction compared to features of lesser relevance (Bach et al. 2015; Samek et al. 2017; Montavon et al. 2018; Bhatt et al. 2020; Nguyen and Martínez 2020).

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

To measure this property, we apply RemOve And Debias (also called ROAD) (Rong et al. 2022a) which returns a curve of scores $\hat{\mathbf{q}} =: (\hat{q}_1, \ldots, \hat{q}_I)$ for a chosen percentage range $\mathbf{p} \in \mathbb{R}^{1\times I}$, with $I \in \mathbb{N}$ being the number of percentage steps (curve visualizations can also be found in Rong et al. (2022a)). For each curve value $\hat{q}_i$, a percentage $p_i \in \mathbf{p}, p_i \in [0, 1]$ of the pixels in the input $\mathbf{x}_n$ is perturbed, according to their value in the explanation $\mathbf{\Phi}^m(f, c_n, \mathbf{x}_n)$ (starting with the highest relevance). The predictions based on the input $\mathbf{x}_n$ and corresponding perturbed input $\hat{\mathbf{x}}_n^i$ are compared, resulting in 1 for equal predictions and 0 otherwise. The procedure is repeated for several inputs n. Accordingly, the ROAD score $\hat{q}^m_{ROAD,i}$ for each percentage i corresponds to the average and is defined as:

$$\hat{q}_{\text{ROAD},m,i} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_{c_n}(c_{pred,n}) \quad \text{with} \quad \mathbf{1}_{c_n}(c_{pred,n}) = \begin{cases} 1 & c_n = c_{pred,n} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbf{1}_{c_n} : \mathbf{C} \rightarrow [0, 1]$ is an indicator function comparing the predicted class $c_{pred,n} = f(\hat{\mathbf{x}}_n^i)$ of $\hat{\mathbf{x}}_n^i$ to $c_n = f(\mathbf{x}_n)$ the predicted class of the unperturbed input $\mathbf{x}_n$. We calculate the score values for up to 50 % of pixel replacements $\mathbf{p}$ of the highest relevant pixel, calculated in steps of 1%; resulting in a curve $\hat{\mathbf{q}}^m_{ROAD}$. For faithful explanations, this curve should degrade faster towards increasing percentages of perturbed pixels (see Eq. (5)). The area under the curve (AUC) is then used as the final ROAD score $q^{ROAD,m}$:

$$q_{\text{ROAD},m} = \text{AUC}(\mathbf{p}, \hat{\mathbf{q}}^m_{ROAD}) \tag{5}$$

Accordingly, a lower ROAD score corresponds to higher faithfulness.

Furthermore, to measure faithfulness, we consider the Faithfulness Correlation $q_{\text{FC},m}$ (Bhatt et al. 2020), defined as:

$$q_{\text{FC},m} = \underset{S\in|S|\subseteq d}{\text{corr}} \left(\bar{\phi}^m_S \,,\; f(\mathbf{x}) - f\left(\mathbf{x}_{[\mathbf{x}_s=\bar{\mathbf{x}}_s]}\right)\right) \tag{6}$$

where $S \in |S| \subseteq d$ is a set of $|S|$ random indices drawn from all pixel indices d in sample $\mathbf{x}$ and $\bar{\phi}^m_S := \sum_{i\in S} \mathbf{\Phi}^m_i(f, c, \mathbf{x})$ is the sum across explanation map values i that are part of the random subset $i \in S$. This set of random indices S is masked (i.e. perturbed) in the input $\mathbf{x}_{[\mathbf{x}_S=\bar{\mathbf{x}}_S]}$, with $\bar{\mathbf{x}} \in \mathbb{R}^d$ being an array filled with the perturbation values (e.g. 0 or 1), which are used to replace all indices i in the perturbed input $\mathbf{x}_{[\mathbf{x}_S=\bar{\mathbf{x}}_S]}$. Accordingly, the correlation of the prediction difference between perturbed and unperturbed input $f(\mathbf{x}) - f\left(\mathbf{x}_{[\mathbf{x}_s=\bar{\mathbf{x}}_s]}\right)$, and the sum across the explanation values of

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

the perturbed pixels $\bar{\phi}_S^m$ is calculated (see Bhatt et al. (2020) for more details and visualizations). Unlike ROAD, the Faithfulness Correlation score increases as the faithfulness improves.

c. Complexity

Complexity is a measure of conciseness, indicating an explanation should consist of a few highly important features (Chalasani et al. 2020; Bhatt et al. 2020) (See Figure 4). The assumption is that concise explanations, characterized by prominent features, facilitate researcher interpretation and potentially include higher informational value with reduced noise.

Here, we use Complexity $q_{\mathrm{COM},m}$ (Bhatt et al. 2020) and Sparseness $q_{\mathrm{SPA},m}$ (Chalasani et al. 2020) as representative metric functions, which can be formulated as follows:

$$q_{\mathrm{COM},m} = \mathrm{H}\left(\mathcal{P}(\Phi^m)\right), \quad \text{with } \mathcal{P}(\Phi^m) := \frac{\Phi(f,c,\mathbf{x})}{\sum_{j\in[d]} \left|\Phi(f,c,\mathbf{x})_j\right|} \tag{7}$$

$$q_{\mathrm{SPA},m} = \frac{\sum_{i=1}^{d}(2i-d-1)\Phi_m(f,\mathbf{x})}{d\sum_{i=1}^{d}\Phi(f,\mathbf{x})}, \tag{8}$$

where $\mathrm{H}(\cdot)$ is the Shannon entropy, $\mathcal{P}(\Phi^m)$ is a valid probability distribution across the frac-

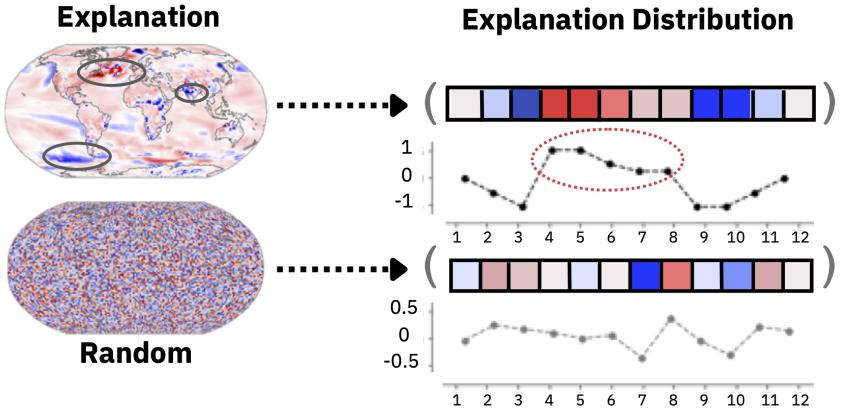

FIG. 4: Diagram of the concept behind the *complexity* property. Complexity assesses how the evidence values are distributed across the explanation map. For this, the distribution of the relevance values from the original explanation is compared to a "random" explanation drawn from a random uniform distribution. Here, shown in a 1-D example, the evidence distribution of the explanation exhibits clear maxima and minima (see maxima in red oval), which is considered desirable and linked to increased scores. The noisy features show a uniform distribution linked to a low complexity score.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

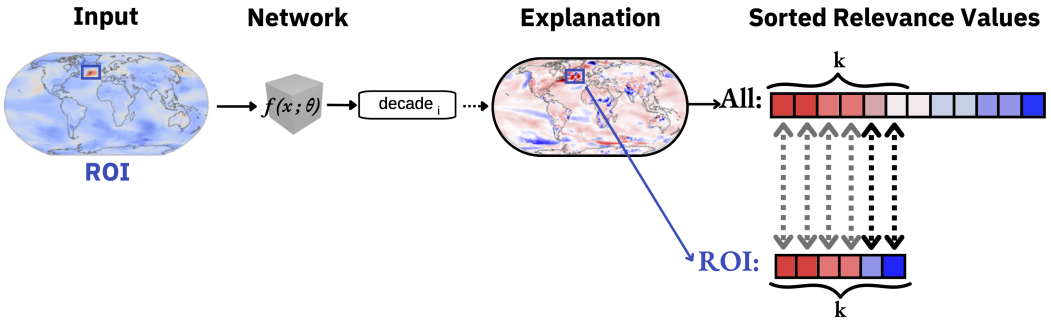

FIG. 5: Diagram of the concept behind the *localization* property. First, an expected region of high relevance for the network decision, the region of interest (ROI), is defined in the input temperature map (blue box). Here, the North Atlantic is chosen, as this region has been discussed to affect the prediction (see Labe and Barnes (2021)). Next, the sorted explanation values of the ROI, encompassing k pixels, are compared to the k highest values of the sorted explanation values across all pixels. An explanation method with strong localization should assign the highest relevance values to the ROI.

tional contribution of all features $\mathbf{x}_i$ of $\mathbf{x}$ to the total magnitude of the explanation values $\sum_{j\in[d]} \left|\Phi(f,c,\mathbf{x})_j\right|$, d is the total number of pixels in $\mathbf{x}$, f is the network function and c is the explained class. Sparseness is based on the Gini index (Hurley and Rickard 2009), while Complexity is calculated using the entropy (see also Bhatt et al. (2020) and Chalasani et al. (2020), where both metric functions are discussed in more detail). While the lower the entropy, the less complex the explanation, a high Gini index indicates less complexity.

d. Localization

For localization, the quality of an explanation is measured based on its agreement with a user-defined region of interest (ROI, see Figure 5). Accordingly, the position of pixels with the highest relevance values (given by the XAI explanation) is compared to the labeled areas, e.g. bounding boxes or segmentation masks. Based on the assumption that the ROI should be mainly responsible for the network decision (ground truth) (Zhang et al. 2018; Arras et al. 2020; Theiner et al. 2022; Arias-Duart et al. 2022), an explanation map yields high localization if high relevance values are assigned to the ROI.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

As localization metrics we use the Top-K-pixel metric (also referred to as Top-K) (Theiner et al. 2022) which is computed as follows:

$$q_{\text{Top-}K,m} = \frac{|\,\mathbf{K} \cap \mathbf{s}\,|}{|\,\mathbf{K}\,|}, \tag{9}$$

where $\mathbf{K} := \Phi^m \cap \mathbf{r}_{|K|}$ denotes the vector of indices of explanation Φ corresponding to the $|\,K\,|$ highest ranked features with $\mathbf{r} = Rank\,(\Phi^m(f, c, \mathbf{x}))$, and $\mathbf{s}$ refers to the indices of ROI (see Theiner et al. (2022) for more details). Furthermore, we consider the Relevance Rank Accuracy $q_{\text{RRA},m}$ (Arras et al. 2020):

$$q_{\text{RRA},m} = \frac{|\,\Phi^m_{|\mathbf{s}|} \cap \mathbf{s}\,|}{|\,\mathbf{s}\,|}, \tag{10}$$

where $\Phi^m_{|\mathbf{s}|} := \Phi^m \cap \mathbf{r}_{|\mathbf{s}|}$ denotes the vector of indices of the explanation Φ corresponding to the highest ranked features $\mathbf{r}_{|\mathbf{s}|} \in \mathbb{R}^{1\times|\mathbf{s}|}$ and $|\,\mathbf{s}\,|$ is the number of pixels in the ROI (details on the calculation and intuition can also be found in Arras et al. (2020)). Thus, Top-K and Relevance Rank Accuracy are the same for $|\,K\,|$ chosen such that it is equal to the number of pixels in the ROI $|\,\mathbf{s}\,|$. Both corresponding scores are high for well-performing methods and low for explanations with low localization.

e. Randomization

Randomization assesses how a random perturbation scenario changes the explanation (See Figure 6). Either the network weights (Adebayo et al. 2018) are randomized or a random class that was not predicted by the network for the input sample $\mathbf{x}$ is explained (Sixt et al. 2020). In both cases, a change in the explanation is expected, since the explanation of an input $\mathbf{x}$ should change if the model changes or if a different class is explained.

Here, we evaluate randomization based on the Model Parameter Randomization Test (Adebayo et al. 2018). The score $q_{\text{MPT},m}$ is defined as the average correlation coefficient between the explanation of the original model f and the randomized model $f_{\mathbf{W}}$ over all layers L:

$$q_{\text{MPT},m} = \frac{1}{L}\sum_{l=1}^{L} \rho(\Phi^m(f, c, \mathbf{x}), \Phi^m(f_l, c, \mathbf{x})) \tag{11}$$

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

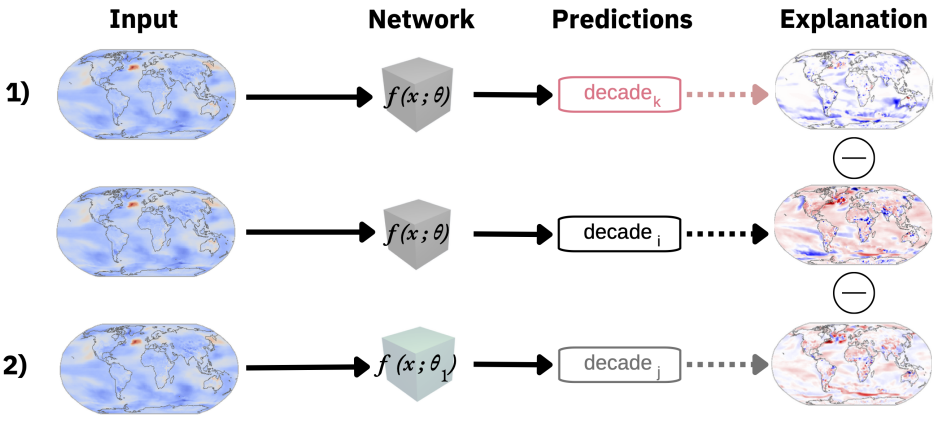

FIG. 6: Diagram of the concept behind the *randomization* property. In the middle row, the original input temperature map is passed through the network, and the explanation map is calculated based on the predicted (grey background) decade. For the Random Logit metric (first row - **1**), the input temperature map and the network remain unchanged but the decade k used to calculate the explanation is randomly chosen (pink font). The resulting explanation map is then compared to the original explanation (indicated by a minus sign) to test its dependence on the class. For the Model Parametrization Randomization Test (bottom row - **2**), the network is perturbed (see green box) with noisy parameters ($\theta_1 = \theta + \text{noise}$), potentially altering the predicted decade (j, grey). The explanation map of the perturbed model should differ from the original explanation map if the explanation is sensitive to the model parameters.

where ρ denotes the Spearman rank correlation coefficient and f_l is the true model with additive perturbed weights of layer l (see Adebayo et al. (2018) for further details).

We also consider the Random Logit score $q_{\text{RL},m}$ (Sixt et al. 2020), which can be defined as e.g. structural similarity index ($SSIM$) or Pearson correlation between an explanation map of a random class $\hat{c}$ (with $f(\mathbf{x}) = c$, $\hat{c} \neq c$) and an explanation map of the predicted class c (see also Sixt et al. (2020) for further details and visualization):

$$q_{\text{RL},m} = \text{SSIM}(\Phi^m(f, c, \mathbf{x}), \Phi^m(f, \hat{c}, \mathbf{x})). \tag{12}$$

Here the metrics return scores $q_{m,n} := q_{\text{MPT/RL},m,n}$ with $n \in \{1, \ldots, N\}$ for either all layers (Randomization metric) $N = L$ or all other classes ($c \neq c_{true}$) with $N = \Gamma$. Thus, we average across L or Γ to obtain q_m, as follows:

$$q_{\text{MPT/RL},m} = \frac{1}{N} \sum_{n=1}^{N} q_{m,n}. \tag{13}$$

The metric scores of randomization and robustness metrics are interpreted similarly, i.e., low metric scores indicate strong performance.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

f. Metric score calculation

The differing scales of the evaluation metric output (e.g. Sparseness ranges between $0-1$, Faithfulness Correlation between -1 and 1, and Local Lipschitz Estimate between $0-\infty$) and their respective interpretation (e.g. for the first two metrics the best score would be 1, whereas for the latter the best score would be 0) complicate their comparison. Therefore, following Murphy and Daan (1985), we introduce a skill score, S, measuring the improvement in forecasts performance A_f over the performance of reference forecast, A_r, relative to the perfect performance A_p, where $A_p = 0$ if performance is measured by the mean-squared error (Murphy and Daan 1985; Murphy 1988). S is given by:

$$S(A_f) = \frac{A_f - A_r}{A_p - A_r}. \tag{14}$$

Here, we calculate the skill score $S(q_m)$ for an explanation method based on the metric scores in each property. The skill score allows us to compare the performance of explanation methods relative to a reference score $A_r = q_r$. To establish this reference score q_r, we create a uniform random baseline explanation similar to Rieger and Hansen (2020), maximizing the violation of each property's underlying assumptions and creating a bad-skill scenario (for details see Appendix A-b). The skill score then measures whether an explanation method improves upon this baseline score.

As the respective perfect score q^* varies across metrics and takes up values of both 0 (e.g. for Local Lipschitz Estimate) and 1 (e.g. for Sparseness), the skill score is:

$$S(q_m) = \begin{cases} 1 - \frac{q_m}{q_r} & \text{if } q^* = 0, \\ \frac{q_m - q_r}{1 - q_r} & \text{if } q^* = 1 \end{cases} \tag{15}$$

where $q_m \in \mathbb{R}$ represents the raw or aggregated metric score (for details see Appendix A-b).

4. Experiments

a. Network predictions, explanations and motivating example

In the following, we evaluate the network performance and discuss the application of the explanation methods for both network architectures. To ensure comparability between networks and

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

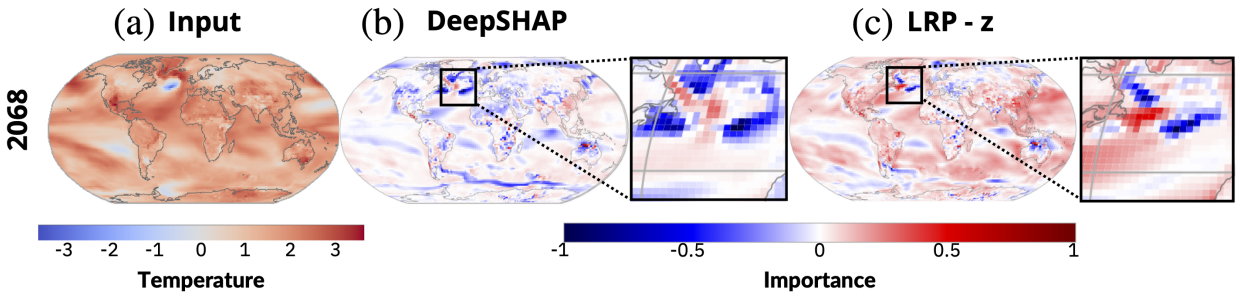

FIG. 7: Motivating example visualizing the difference between different XAI methods. Shown are the T2m-temperature map (a) for the year 2068 with the corresponding DeepSHAP (b) and LRP-z (c) explanation maps of the MLP. For both XAI methods, red indicates a pixel contributed positively, and blue indicates a negative contribution to the predicted class. Next to the explanation maps, a zoomed-in map of the North Atlantic region (NA, 10 – 80°W, 20 – 60°N) is shown, demonstrating different evidence for DeepSHAP and LRP-z.

comparability to our case study Labe and Barnes (2021), we use a similar set of hyperparameters for the MLP and the CNN during training. A detailed performance discussion is provided in Appendix B-a. The achieved similar performance ensures that XAI evaluation score differences between the MLP and the CNN are not caused by differences in network accuracy.

After training and performance evaluation, we explain all correctly predicted temperature maps in the training, validation, and test samples (see Appendix B-a for details). These explanations are most often subject to further research on physical phenomena learned by the network (Barnes et al. 2020; Labe and Barnes 2021; Barnes et al. 2021; Labe and Barnes 2022). We apply all XAI methods presented in Section 2c to both networks with the exception of the composite rule of LRP, converging to the LRP-z rule for the MLP model due to its dense layer architecture (Montavon et al. 2019). The corresponding explanation maps across all XAI methods and for both networks are displayed in Figures B4 and B5. Despite explaining the same network predictions, different methods assign different relevance values to the same areas, revealing the disagreement problem in XAI (Krishna et al. 2022).

To illustrate this explanation disagreement, we show the explanation maps for the year 2068 given by DeepSHAP and LRP-z, alongside the input temperature map in Figure 7. According to the primary publication Labe and Barnes (2021), the cooling patch in the North Atlantic (NA), depicted in the zoomed-in map sections of 10 – 80°W, 20 – 60°N of Figure 7, significantly contributes to the network prediction for all decades. Thus, its reasonable to assume high relevance values in this region. However, the two XAI methods display contrary signs of relevance in some areas,

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

impeding visual comparison and interpretation. The varying sign can be attributed to DeepSHAP being based on feature-removal and modified gradient backpropagation, while LRP-z, in contrast, being theoretically equivalent to input times gradient. Thus, the two explanations potentially display different aspects of the network decision (Clare et al. 2022) and explanations can vary in sign depending on the input image (see also discussions on input shift invariance in Mamalakis et al. (2022a)). Nonetheless, we also find common features, as for example in Australia or throughout the antarctic region. Thus, a deeper understanding of explanation methods and their properties is necessary to enable an informed method choice.

b. Assessment of explanation methods

To introduce the application of XAI evaluation, we compare the different XAI methods applied to the MLP and calculate their skill scores across all five XAI method properties (see Section 3). For each property, two representative metrics (hyperparameters are listed in Appendix B-b) are computed and compared. Each skill score is averaged across 50 random samples drawn from the explanations of all correctly predicted inputs and we provide the standard error of the mean (SEM) (see Appendix A-b for details). To account for potential biases resulting from the choice of the testing period, we also compute the scores for random samples not limited to correct predictions. We report qualitatively robust findings (not shown) compared to the scores shown here. Our results are depicted in Figure 8.

For the *robustness* property, we find that all tested explanation methods result in similar, high, and closely distributed skill scores (≥ 0.85 and ≤ 0.93) for both the Average Sensitivity metric (hatches in Figure 8a) and Local Lipschitz Estimate metric (no hatches), where the latter shows slightly higher values overall. For both metrics, we find that salience (earthy tones) and sensitivity methods (violet tones) show a similar robustness skill and perturbation-based methods (SmoothGrad, NoiseGrad, and Integrated Gradients) do not significantly improve skill compared to the respective baseline explanations (gradient and input times gradient). We relate the latter finding to the low signal-to-noise ratio of the climate data and variability between different ensemble members, complicating the choice of an efficient perturbation threshold for the explanation methods. Nonetheless, these findings disagree with previous studies regarding suggested robustness improvements when

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

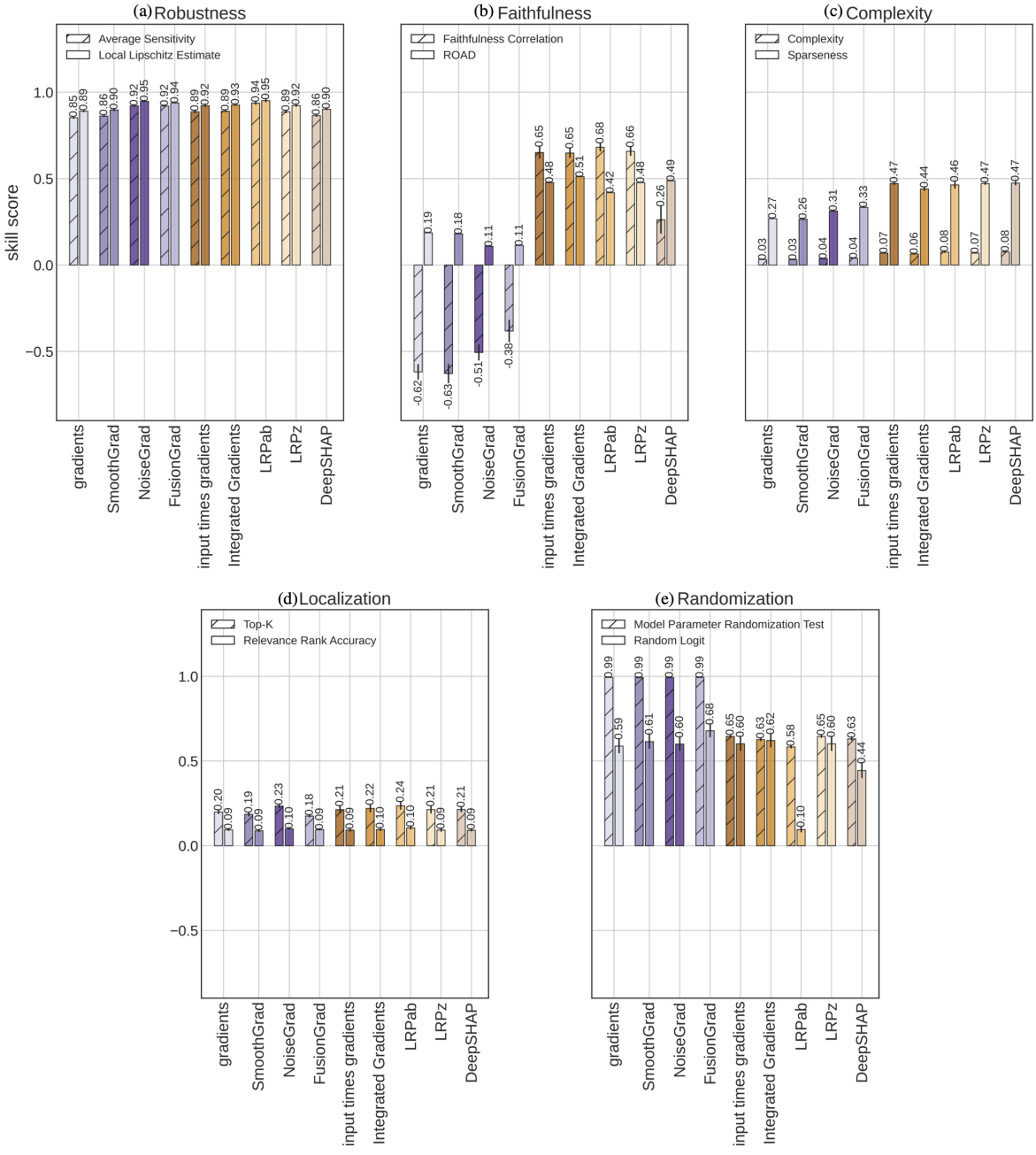

FIG. 8: Barplot of skill scores based on the random baseline reference for two different metrics in each, (a) the robustness, (b) faithfulness, (c) complexity, (d) localization, and (e) randomization property. The different metrics are indicated by hatches or no hatches on the bar. We report the mean skill score (as bar labels) and the standard error of the mean (SEM), indicated by the error bars in black on each bar. The bar color scheme indicates the grouping of the XAI methods into sensitivity (violet tones) and salience/attribution methods (earthy tones).

applying salience and perturbation-based methods (Smilkov et al. 2017; Sundararajan et al. 2017; Bykov et al. 2022b; Mamalakis et al. 2022a).

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

For *faithfulness*, we find pronounced skill score differences between both metrics, with ROAD scores indicating positive skill for all methods, whereas Faithfulness Correlation scores include negative values for the sensitivity methods (hatched violet bars in Fig. 8b). This disparity arises from the calculation of Faithfulness Correlation metric scores using the correlation coefficient, and the distinct interpretations of relevance values in salience maps versus sensitivity maps. Since sensitivity maps display the network's sensitivity towards the change in the value of each pixel (the sign conveys the direction), the impact of the masking value depends on the discrepancy between the original pixel value and the masking value, leading to a negative correlation. Nonetheless, across metrics, the best skill scores ≤ 0.6 are achieved by input times gradient, Integrated Gradients, and LRP-z, followed by $S(q_{\text{LPR-}\alpha-\beta}) \leq 0.42$. Furthermore, sensitivity methods (violet tones) achieve overall lower skill scores. Although DeepSHAP exhibits a lower faithfulness correlation skill (which we attribute to the challenge of applying Shapley values to continuous data (Han et al. 2022) and vulnerability towards feature correlation Flora et al. (2022)), the method still outperforms the sensitivity methods, indicating salience (or attribution) methods provide more faithful relevance values. However, this is due to salience methods indicating the contribution of each pixel to the prediction as required by faithfulness. Thus, sensitivity methods inherently result in less faithful explanations. We note that the input multiplication of salience methods can lead to a loss of information when using standardized input pixels, as zero values in the input (i.e., values close to climatology) will result in zero values in the explanation regardless of the networks sensitivity to it (see Section 2c and Mamalakis et al. (2022a) discussing "ignorant to zero input").

For *complexity* (Figure 8c), all explanation methods exhibit low Complexity scores compared to Sparseness, indicating the explanations on climate data exhibit similar entropy to uniformly sampled values. This similarity in entropy can be attributed to the increased variability and subsequently low signal-to-noise ratio of climate data (Sonnewald and Lguensat 2021; Clare et al. 2022). For the Sparseness metric, skill scores show skill improvement for salience (attribution) methods. We also find slight skill score improvements for NoiseGrad and FusionGrad, suggesting that incorporating network perturbations may decrease explanation complexity.

To compute the results of the *localization* metrics, Top-K (hatches in Fig. 8d) and Relevance Rank Accuracy (no hatches), we select the region in the North Atlantic ($10-80°$W, $20-60°$N) as our ROI, with the cooling in this region being a recognized feature of climate change Labe and

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Barnes (2021). In both metrics, all explanation methods yield low skill scores. This is consistent with lower Sparseness skill scores in complexity (≤ 0.47) indicating that high-relevance values are spread out, with the ROI also including fewer distinct features. In addition, high relevance in the ROI depends on whether the network learned this specific area. Thus, our results potentially indicate an inadequate choice of the ROI (either size or location) and show that localization metrics can identify a learned region. Nonetheless, LRP-$\alpha-\beta$ yields the highest skill across metrics, indicating that attributing only positive relevance values improves the distinctiveness of features in the NA region. Similar to complexity, salience methods (earthy tones) yield a slightly higher localization skill than sensitivity methods (violet tones) with the exception of NoiseGrad.

Lastly, we present the randomization results (Figure 8e). For the Random Logit metric, all XAI methods yield lower skill scores (≥ 0.1 and ≤ 0.58). This can be attributed to the network task classes being defined based on decades with an underlying continuous temperature trend. Thus, the differences in temperature maps can be small for subsequent years, and the network decision and explanation for different classes may include similar features. Nonetheless, we find salience (earthy tones) and sensitivity methods (violet tones) to yield no clear separation. Instead, XAI methods using perturbation result in higher skill scores, with mean improvements for FusionGrad exceeding the SEM, as well as a slight improvement for NoiseGrad and SmoothGrad over gradient and Integrated Gradients over input times gradient. Thus, while input perturbations already slightly improve the class separation in the explanation, also including network perturbation yields favorable improvement. For the Model Parameter Randomization Test scores, skill scores are overall higher (≥ 0.58 and ≤ 0.99) across all explanation methods, and sensitivity methods outperform salience methods, the latter aligning with Mamalakis et al. (2022b). Similar to the complexity results, the DeepSHAP skill score aligns with other salience method results. In addition, LRP-α-β yields the worst skill across metrics, potentially due to neglecting negatively contributing neurons during backpropagation (see Eq. (A4) in Appendix A-a) and corresponding variations across classes and under parameter randomization.

c. Network-based comparison

To compare the performance of explanation methods for the MLP and CNN networks, we selected one metric per property: Local Lipschitz Estimate for robustness, ROAD for faithfulness,

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Sparseness for complexity, Top-K for localization, and Model Parameter Randomization Test for randomization.

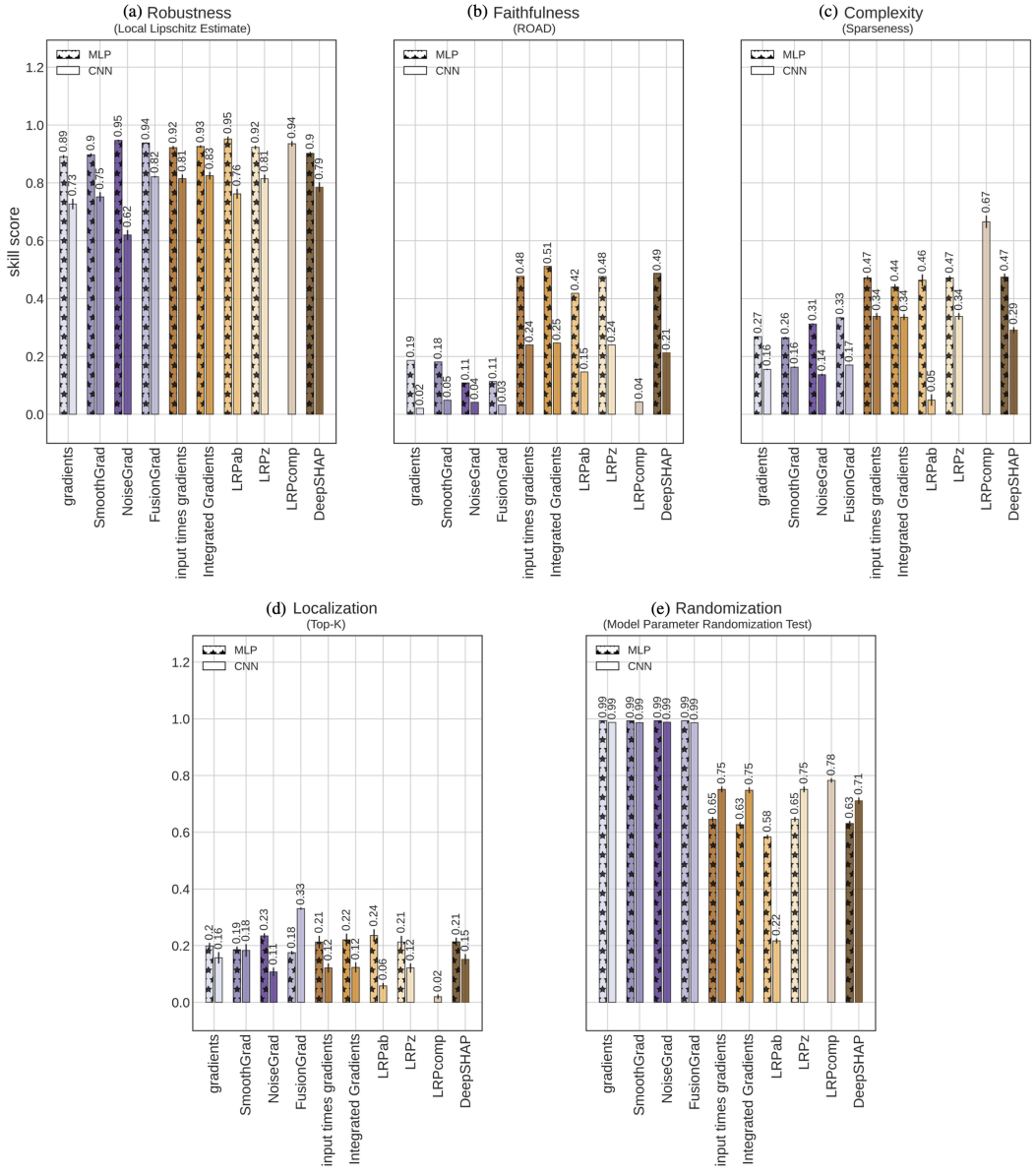

FIG. 9: Barplot of skill scores based on the random baseline reference for MLP (star hatches) and CNN (no hatches) in each, the robustness (a), faithfulness (b), the complexity (c), localization (d), and randomization (e) property. We report the skill score (as bar labels) and the standard error of the mean (SEM) of all scores, indicated by the error bars in black on each bar. The bar color scheme indicates the grouping of the XAI methods into sensitivity (violet tones) and salience/attribution methods (earthy tones). Note that, for LRP-composite (LRP-$comp$) we only report the CNN results (for details, see Section 4a).

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

For robustness (see Figure 9a), XAI methods applied to the CNN yield strong skill score variations, with the MLP results showing overall higher skill scores. For the CNN, LRP-composite provides the best robustness skill. We find salience methods to exhibit slightly higher skill scores, the exception being FusionGrad outperforming LRP-α-β and DeepShap. This suggests that due to the differences in learned patterns between CNN and MLP, including both network and input perturbations yields more robust explanations, while the combination of a removal-based technique (Covert et al. 2021) with a modified gradient backpropagation (Ancona et al. 2019)as in DeepSHAP and neglecting negatively contributing neurons as in LRP-α-β worsens robustness compared to other salience methods. Moreover, explanation methods using input perturbations improve sensitivity explanation robustness for the CNN (SmoothGrad and FusionGrad), while methods using only network perturbations decrease robustness skill (NoiseGrad).

In the faithfulness property (see Figure 9b), salience explanation methods (Integrated Gradients, input times gradient, and LRP) achieve higher skill for both networks, aligning with previous research (Mamalakis et al. 2022b,a) and the theoretical differences (see Section 4b). However, LRP-composite is the exception, adding additional insight to the findings of other studies Mamalakis et al. (2022a), as LRP-composite sacrifices faithful evidence for a less complex (human-aligned Montavon et al. (2019)) and more robust explanation. Moreover, perturbation-based explanation methods (SmoothGrad, NoiseGrad, FusionGrad, and Integrated Gradients) do not significantly increase the faithfulness skill compared to their respective baseline explanations (gradient and input times gradient), except for Integrated Gradients for the MLP. Similar to the MLP results, LRP-α-β acts as an outlier compared to other salience methods. For the CNN also the DeepSHAP's faithfulness skill is decreased, contradicting theoretical claims and other findings (Lundberg and Lee 2017; Mamalakis et al. 2022a). Since the CNN learns more clustered patterns (groups of pixel according to the filter-based architecture), we attribute this outcome to both DeepSHAP's theoretical definitions (Han et al. 2022) and vulnerability towards feature correlation (Flora et al. 2022), with the latter making partitionSHAP a more suitable option (Flora et al. 2022).

In complexity, salience methods exhibit slight skill improvement over sensitivity methods across networks, except for LRP-α-β for the CNN (Figure 9c). This indicates that neglecting feature relevance is more influential for the CNN's explanation, leading to fewer distinct features in the ex-

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

planation, while the lower DeepSHAP skill further confirms the previously discussed disadvantages of DeepSHAP for the CNN.

In localization, both MLP and CNN show similar low overall skill scores (≤ 0.33), indicating that the size or location of the ROI was not optimally chosen for the case study. Nonetheless, the skill scores across XAI methods are in line with the complexity results, except for the worst and best skill scores. LRP-composite yields the lowest localization skill, further confirming its trade-off between faithfulness and interpretability, also in the ROI. FusionGrad provides the highest localization skill for the CNN. In contrast, LRP-α-β yields the highest skill for the MLP but the second lowest skill score for the CNN. The difference in results across networks for complexity and localization can be attributed to differences in learned patterns (as discussed above), affecting properties that assess the spatial distribution of evidence in the image.

Lastly, for randomization (see Figure 9e), regardless of the network sensitivity methods outperform salience methods, indicating a decreased susceptibility to changes in the network parameters. While slightly lower, the randomization skill score of DeepSHAP does agrees with other salience methods aligning with Mamalakis et al. (2022b,a).

Overall, our results show that while explanation methods applied to different network architectures retain similar faithfulness and randomization properties, their robustness, complexity, and localization properties depend on the specific architecture.

d. Choosing a XAI method

Evaluation metrics enable the comparison of different explanation methods based on various properties for different network architectures, allowing us to assess their suitability for specific tasks. Here, we propose a framework to select an appropriate XAI method.

Practitioners first determine which explanation properties are essential for their network task. For instance, for physically informed networks, randomization (the Model Parameter Randomization Test) is crucial, as parameters are meaningful and explanations should respond to their successive randomization. Similarly, localization might be less important if an ROI cannot be determined beforehand. Second, practitioners calculate evaluation scores for each selected property across various XAI methods. We suggest calculating the skill score (see Section 3f) to improve score interpretability. Third and last, the optimal XAI method for the task can be chosen based on the

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

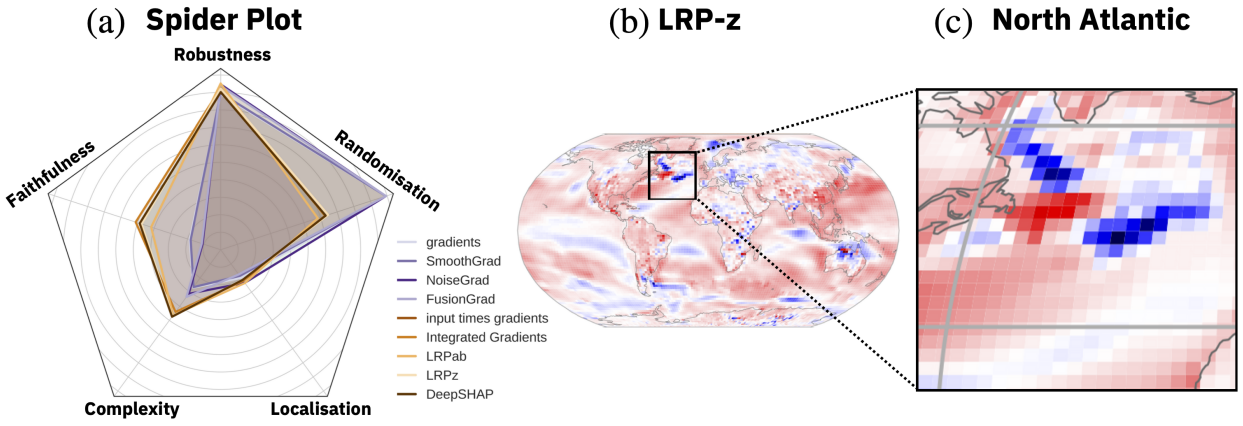

FIG. 10: Visualization of the proposed procedure to choose an appropriate XAI method. In the spider plot (a) the mean skill scores for all properties across nine explanation methods (MLP explanations) are visualized, according to Figure 9. The spider plot can be used as a visual aid alongside the skill scores or ranks in each essential property to identify the best-performing XAI method. In the plot, the best results correspond to the furthest distance from the center of the graph. The LRP-z explanation map of the decade prediction on the temperature map of 2068 is shown in (b) and the North Atlantic (NA) region in (c).

skill scores independently or rank of explanation method, as in previous studies (Hedström et al. 2023b; Tomsett et al.; Rong et al. 2022b; Brocki and Chung 2022; Gevaert et al. 2022).

In our case study, for example, the explanation method should exhibit robustness towards variation across climate model ensemble members, display concise features (complexity) without sacrificing faithfulness, and capture randomization of the network parameter (randomization). Using the Quantus XAI evaluation library (Hedström et al. 2023b), we visualize the evaluation results for the MLP using a spider plot (Figure 10a), with the outermost line indicating the best-performing XAI method in each property. All methods yield similar robustness skill but differ in randomization, faithfulness and complexity skills. LRP-z (light beige), input times gradient (ocher), Integrated Gradients (orange), and DeepShap (brown) provide the most faithful explanations (similar to findings in Mamalakis et al. (2022a)), with DeepShap providing a slightly worsened randomisation and robustness skill.

Based on the different strengths and weaknesses, we would select LRP-z to explain the MLP predictions (Figure 10b) and analyze the impact of the NA region (Figure 10c) on the network predictions. According to the explanation, the network heavily depends on the North Atlantic region and the cooling patch pattern, suggesting its relevance in correctly predicting the decade in this global warming simulation scenario. However, we also stress that additionally applying a

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

sensitivity method such as gradient-based SmoothGrad potentially illuminates more aspects of this network decision, as sensitivity methods provide strong randomization, in contrast to LRP-z.

5. Discussion and Conclusion

AI models, particularly DNNs, can learn complex relationships from data to predict unseen points afterward. However, their black box character restricts the human understanding of the learned input-output relation, making DNN predictions challenging to interpret. To illuminate the model's behavior, local XAI methods were developed, that identify the input features responsible for individual predictions and offer novel insights in climate AI research (Camps-Valls et al. 2020; Gibson et al. 2021; Dikshit and Pradhan 2021; Mayer and Barnes 2021; Labe and Barnes 2021; Van Straaten et al. 2022; Labe and Barnes 2022). Nevertheless, the increasing number of available XAI methods and their visual disagreement (Krishna et al. 2022), illustrated in our motivating example (Figure 7), raise two important questions: Which explanation method is trustworthy, and which is the appropriate choice for a given task?

To address these questions, we introduced XAI evaluation to climate science, building upon existing climate XAI research as our case study (Labe and Barnes 2021). We evaluate and compare various local explanation methods for an MLP and a CNN network regarding five properties, i.e., *robustness*, *faithfulness*, *randomization*, *complexity* and *localization*, that are provided by the Quantus library (Hedström et al. 2023b). Furthermore, we improve the interpretation of the evaluation scores by calculating a skill score in reference to a random uniform explanation.

In the first experiment, we showcase the application of XAI evaluation on the MLP explanations using two metrics for each property (Alvarez-Melis and Jaakkola 2018; Montavon et al. 2019; Yeh et al. 2019; Bhatt et al. 2020; Arras et al. 2020; Rong et al. 2022a; Hedström et al. 2023b). Our results indicate that salience methods (i.e., input times gradient, Integrated Gradients, LRP) yield an improvement in faithfulness and complexity skill but a reduced randomization skill. Contrary to salience methods, sensitivity methods (gradient, SmoothGrad, NoiseGrad, and FusionGrad) show higher randomization skill scores while sacrificing faithfulness and complexity skills. These results indicate that a combination of explanation methods can be favourable depending on the explainability context. We also establish that evaluating explanation methods in a climate context mandates careful consideration. For example, due to the natural variability in the data, the

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Sparseness metric is best suited for determining explanation complexity. Further, the Random Logit metric is favored for classification with pronounced class separations rather than datasets with continuous features spanning multiple classes. Lastly, we highlight the importance of the correct identification of an ROI to ensure an informative localization evaluation and that localization metrics enable probing the network regarding learned physical phenomena.

In the second experiment, we compare the properties of MLP and CNN explanations across all XAI methods. Both localization and complexity evaluation show larger variations between networks, due to differences in how the networks learn features in the input. The robustness results exhibit similar variation, with the CNN showing higher skill scores for all input perturbation-based methods like SmoothGrad, FusionGrad, and Integrated Gradients, contrary to the MLP, with the exception of NoiseGrad. Independent of network architecture, explanations using averages across input perturbations, like SmoothGrad and Integrated Gradients, do not consistently increase and, in some cases, even decrease the faithfulness skill. Furthermore, sensitivity methods result in less faithful and more complex explanations but capture network parameter changes more reliably. In contrast, salience methods are less complex, except for LRP-α-β explaining the CNN. Moreover, salience methods exhibit a higher faithfulness skill and lower randomization skill compared to sensitivity methods, consistent with findings in Mamalakis et al. (2022b,a) and in line with salience methods presenting the contribution of each input pixel rather than sensitivity (see Section 4b), due to input multiplication. Contrary to previous research (Mamalakis et al. 2022a), LRP-$composite$ was an outlier among salience methods, sacrificing a faithful explanation for an improved complexity skill and higher robustness. Similarly, LRP-α-β and DeepSHAP stands out as an exception among salience methods applied to the CNN due to almost consistently lower skill scores. We attribute both findings to the mathematical definition of each method. While LRP-$composite$ is optimized towards improved interpretation resulting in less feature content, DeepSHAP is based on feature-removal and modified gradient backpropagation, and is vulnerable towards feature correlation, for CNN features and LRP-α-β neglecting negatively contributing neurons during backpropagation.

Lastly, we propose a framework using XAI evaluation to support the selection of an appropriate XAI method for a specific research task. The first step is to identify important XAI properties for the network and data, followed by calculating evaluation skill scores across the properties

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

for different XAI methods. Then, the resulting skill scores across XAI methods can be ranked or compared directly to determine the best-performing method or combination of methods. In our case study, LRP-z (alongside input times gradient and Integrated Gradients) yields suitable results in the MLP task, allowing the reassessment of our motivating example (Figure 7) and the trustworthy interpretation of the NA region as a contributing input feature.

Overall, our results demonstrate the value of XAI evaluation for climate AI research. Due to their technical and theoretical differences (Letzgus et al. 2022; Han et al. 2022; Flora et al. 2022), the various explanation methods can reveal different aspects of the network decision and exhibit different strengths and weaknesses. Evaluation metrics allow to compare explanation methods by assessing their suitability and properties, in different explainability contexts. Next to benchmark datasets, evaluation metrics also contribute to the benchmarking of explanation methods. XAI evaluation can support researchers in the choice of an explanation method, independent of the network structure and targeted to their specific research problem.

Acknowledgments. This work was funded by the German Ministry for Education and Research through project Explaining 4.0 (ref. 01IS200551). M.K. acknowledges funding from XAIDA (European Union's Horizon 2020 research and innovation program under grant agreement No 101003469). The authors also thank the CESM Large Ensemble Community Project (Kay et al. 2015) for making the data publicly available. Support for the Twentieth Century Reanalysis Project version 3 dataset is provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER), the National Oceanic and Atmospheric Administration Climate Program Office, and by the NOAA Earth System Research Laboratory Physical Sciences Laboratory.

Data availability statement. Our study is based on the RPC8.5 configuration of the CESM1 Large Ensemble simulations (`https://www.cesm.ucar.edu/projects/community-projects/LENS/instructions.html`). The data is freely available (`https://www.cesm.ucar.edu/projects/community-projects/LENS/data-sets.html`). The source code for all experiments will be accessible at (`https://github.com/philine-bommer/Climate_X_Quantus`). All experiments and code are based on Python v3.7.6, Numpy v1.19 (Harris et al. 2020), SciPy v1.4.1 (Virtanen et al. 2020), and colormaps provided by Matplotlib v3.2.2 (Hunter 2007). Additional Python packages used for the development of the ANN, explanation methods, and evaluation

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

include Keras/TensorFlow (Abadi et al. 2016), iNNvestigate (Alber et al. 2019) and Quantus (Hedström et al. 2023b). We implemented all explanation methods except for NoiseGrad and FusionGrad using iNNvestigate (Alber et al. 2019). For XAI methods by (Bykov et al. 2022b) and Quantus (Hedström et al. 2023b) we present a Keras/TensorFlow (Abadi et al. 2016) adaptation in our repository. All dataset references are provided throughout the study.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

APPENDIX A

Additional Methodology

a. Explanations

To provide a theoretical background we provide formulas for the different XAI methods we compare, in the following Section.

Gradient

The gradient method is the weak derivative $\nabla_x := \nabla f(\mathbf{x})$ of the network output $f(\mathbf{x})$ with respect to each entry of the temperature map $\mathbf{x} \in \mathbf{X}$ (Baehrens et al. 2010).

$$\Phi(f(\mathbf{x})) = \nabla_x \tag{A1}$$

Accordingly, the raw gradient has the same dimensions as the input sample $\nabla_x, \mathbf{x} \in \mathbb{R}^D$.

input times gradient

input times gradient explanations are based on a point-wise multiplication of the impact of each temperature map entry on the network output, i.e., the weak derivative ∇_x, with the value of the entry in the explained temperature map $\mathbf{x}$. All explanations are calculated as follows:

$$\Phi(f(\mathbf{x})) = \nabla_x \mathbf{x} \tag{A2}$$

with $\Phi(f(\mathbf{x})), \nabla_x, \mathbf{x} \in \mathbb{R}^D$

Integrated Gradients

The Integrated Gradients method aggregates gradients along the straight line path from the baseline $\overline{\mathbf{x}}$ to the input temperature map $\mathbf{x}$. The relevance attribution function is defined as follows:

$$\Phi(f(\mathbf{x})) = (\mathbf{x} - \overline{\mathbf{x}}) \odot \int_0^1 \nabla f(\overline{\mathbf{x}} + \alpha(\mathbf{x} - \overline{\mathbf{x}}))\, \mathrm{d}\alpha, \tag{A3}$$

where $\odot$ denotes the element-wise product and α is the step-width from $\overline{\mathbf{x}}$ to $\mathbf{x}$.

Layerwise Relevance Propagation (LRP)

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

For LRP, the relevances of each neuron i in each layer l are calculated based on the relevances of all connected neurons j in the higher layer $l+1$ (Samek et al. 2017; Montavon et al. 2017).

For the α-β-**rule** the weighted contribution of a neuron j to a neuron i, i.e., $z_{ij} = a_i^{(l)} w_{ij}^{(l,l+1)}$ with $a_i^{(l)} = x_i$, are separated in a positive z_{ij}^+ and negative z_{ij}^- part. Accordingly, the propagation rule is defined by:

$$R_i^{(l)} = \sum_j \left(\alpha \frac{z_{ij}^+}{\sum_i z_{ij}^+} + \beta \frac{z_{ij}^-}{\sum_i z_{ij}^-} \right) \tag{A4}$$

with α as the positive weight, β as negative weight and $\alpha + \beta = 1$ to maintain relevance conservation. We set $\alpha = 1$ and $\beta = 0$

The z-**rule** accounts for the bounding that input images in image classification are exhibiting, by multiplying positive network weights w_{ij}^+ with the lowest pixel value l_i in the input and the negative weights w_{ij}^- by the highest input pixel value h_i (Montavon et al. 2017). The relevance is calculated as follows:

$$R_i^{(l)} = \sum_j \frac{z_{ij} - l_i w_{ij}^+ - h_i w_{ij}^+}{\sum_i z_{ij} - l_i w_{ij}^+ - h_i w_{ij}^+} \tag{A5}$$

For the *composite*-**rule** the relevances of the last layers with high neuron numbers are calculated based on LRP-0 (see Bach et al. (2015)), which we drop due to our small network. In the middle layers propagation is based on LRP-ϵ, defined as:

$$R_i^{(l)} = \sum_j \alpha \frac{a_j (w_{ij} + \gamma w_{ij}^+)}{\sum_i a_j (w_{ij} + \gamma w_{ij}^+)} \tag{A6}$$

The relevance of neurons in the layer before the input follows from LRP-γ

$$R_i^{(l)} = \sum_j \alpha \frac{z_{ij}}{\sum_i z_{ij}} \tag{A7}$$

and the relevance of the input layer is calculated based on Eq. A5.

SmoothGrad

The SmoothGrad explanations are defined as the average over the explanations of M perturbed

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

input images $\mathbf{x}+\mathbf{g}_i$ with $i = [1,\ldots,M]$.

$$\Phi(f(\mathbf{x})) = \frac{1}{M+1}\sum_{i=0}^{M}\Phi_0(f(\mathbf{x}+\mathbf{g}_i)) \tag{A8}$$

The additive noise $\mathbf{g}_i \sim \mathcal{N}(0,\sigma)$ is generated using a Gaussian distribution.

NoiseGrad

NoiseGrad samples N sets of perturbed network parameters $\hat{\boldsymbol{\theta}}_i = \boldsymbol{\eta}_i\boldsymbol{\theta}$ using multiplicative noise $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{1},\boldsymbol{\sigma})$. Each set of perturbed parameters $\hat{\boldsymbol{\theta}}_i$ results in a perturbed network $f_i(\mathbf{x}) := f(\mathbf{x};\hat{\boldsymbol{\theta}}_i)$, which are all explained by a baseline explanation method $\Phi_0(f(\mathbf{x}))$. The NoiseGrad explanation is calculated as follows:

$$\Phi(f(\mathbf{x})) = \frac{1}{N+1}\sum_{i=0}^{N}\Phi_0(f_i(\mathbf{x})) \tag{A9}$$

with $f_0(\mathbf{x}) = f(\mathbf{x})$ being the unperturbed network.

FusionGrad

For FusionGrad the NG procedure is extended by combining the SG procedure using M perturbed input samples with NG calculations. Accordingly, FG can be calculated as follows:

$$\Phi(f(\mathbf{x})) = \frac{1}{M+1}\frac{1}{N+1}\sum_{j=0}^{M}\sum_{i=0}^{N}\Phi_0(f_i(\mathbf{x}_j)) \tag{A10}$$

Deep SHapley Additive exPlanations (DeepSHAP) (Lundberg and Lee 2017)

The Deep SHAP Explainer, uses the concept of DeepLift (Shrikumar et al. 2016) to approximate Shapley values. Formally, we can express the Shapley values as follows:

$$\phi_{d_i}(f_W,x) = \sum_{S\subset d\setminus d_i}\frac{|S|!(|d|-|S|-1)!}{|d|!}[f(x)-f(x_S)], \tag{A11}$$

where x is the input with features d and individual features $d_i \in d$, f is our model and $x_S := x \setminus d_i$ is the masked input, only containing the features in $S \subset d \setminus \{d_i\}$, all subsets that do not contain the feature d_i. For DeepSHAP, the network f is separated into individual components f_i according to the layer structure as proposed in DeepLift. Similar to Integrated Gradients, DeepSHAP uses a reference value (here chosen as an all-zero reference image), relative to which the contributions

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

of each feature are calculated. This is achieved by determining the multiplicators for each layer according to the DeepLift multiplicators and the multiplicators are back-propagated to the input layer (Shrikumar et al. 2016; Lundberg and Lee 2017).

For visualizations, as depicted in Figure B4 and B5 we maintain comparability of the relevance maps $\Phi(f(\mathbf{X}_{i,t})) = \bar{R}^{(i,t)} \in \mathbb{R}^{vxh}$ across different methods, by applying a *min-max normalization* to all explanations:

$$\bar{R}^i = \frac{\mathbb{I}_{\max} R^i}{\max\left(r_{jk} | r_{jk} \in R^i \forall j \in [1,v] \forall k \in [1,h]\right)} - \frac{\mathbb{I}_{\min} R^i}{\min\left(r_{jk} | r_{jk} \in R^i \forall j \in [1,v] \forall k \in [1,h]\right)} \quad \text{(A12)}$$

with $\mathbb{I}_{\min}, \mathbb{I}_{\max} \in \mathbb{R}^{vxh}$ defining corresponding minimum/maximum indicator masks, i.e., for the minimum indicator each entry $\mathbf{i}_{\min}^{(jk)} = 1, \forall r_{jk} < 0$ and $\mathbf{i}_{\min}^{(jk)} = 0 \forall r_{jk} \geq 0$, for the maximum indicator entries are defined reversely $\mathbf{i}_{\max}^{(jk)} = 1, \forall r_{jk} \geq 0$ and $\mathbf{i}_{\max}^{(jk)} = 0$ otherwise. The normalization maps pixel-wise relevance $r_{jk} \mapsto \bar{r}_{jk}$ with $\bar{r}_{jk} \in [-1, 1]$ for methods identifying positive and negative relevance and $\bar{r}_{jk} \in [0, 1]$ for methods contributing only positive relevance values.

b. Evaluation Metrics

(i) Random Baseline Similar to Rieger and Hansen (2020), we establish a random baseline as an uninformative baseline explanation. The artificial explanation $\Phi_{\text{rand}} \in \mathbb{R}^{h \times v}$ is drawn from a Uniform distribution $\Phi_{\text{rand}} \sim U(0, 1)$. Each time a metric reapplies the explanation function, for example in the robustness metrics when the perturbed input is subject to the explanation method, we redraw each random explanation. The only exception for the re-explanation step is the randomization metric as it aims for a maximally different explanation. Thus, to maximally violate the metric assumptions, we fix the explanation, emulating a constant explanation for a changing network $\Phi(\mathbf{x}, f_\theta) \approx \Phi(\mathbf{x}, f_{\hat{\theta}})$.

(ii) Score Calculation As discussed in Section 3e, we calculate the skill score according to the optimal metric outcome. Thus, skill scores reported for the Average Sensitivity, the Local Lipschitz Estimate, the ROAD, the Complexity, the Model Parameter Randomization Test, and the Random

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Logit metrics are calculated based on the first case of Eq. (15), while the skill scores calculation based on Faithfulness Correlation, Top-K, Relevance Rank Accuracy, and Sparseness scores are calculated follows the bottom case of Eq. (15).

We calculate the mean skill scores Q^m and corresponding SEM reported in Figures 8−9 based on the skill scores of $I = 50$ explanation samples. We choose this number of samples to provide valid statistics, while maintaining computational efficiency, for both networks. All samples are drawn randomly from the calculated explanations (both training and test data). For each explanation method M, both mean skill scores Q^m and corresponding SEM are calculated as follows:

$$\begin{aligned} Q_m &= \frac{1}{I}\sum_{j=1}^{I} \bar{q}_{m,j} \\ \bar{\mathrm{s}}^m &= \frac{\mathrm{s}}{\sqrt{I}} \end{aligned} \tag{A13}$$

with s being the standard deviation of the normalized scores $\bar{q}_i^m$ (see Section 3) across explanation samples.

An exception is the ROAD metric, as discussed in Section 3, the curve used in the AUC calculation results from the average of $N = 50$ samples. Thus, we repeat the AUC calculation for $V = 10$ draws of $N = 50$ samples and calculate the mean skill score and the SEM.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

APPENDIX B

Additional Experiments

a. Network and Explanation

Aside from the learning rate l ($l_{\mathrm{CNN}} = 0.001$), we maintain a similar set of the hyperparameters to Labe and Barnes (2021) and use the fuzzy classification setup for the performance validation. To assess the predictions of the network for each individual input we include the network predictions for 20CRv3 Reanalysis data, i.e., observations (Slivinski et al. 2019). We measure performance using both the $RMSE = R$ between true $\hat{y}_{\mathrm{true}}$ and predicted year $\hat{y}$ as well as the accuracy on the test set. Both the MLP and the CNN have a similar performance compared to the primary publication. We show in Figure B3 the regression curves for the model data (grey) and reanalysis data (blue) of A) the MLP and B) CNN (see also Figure 3c in Labe and Barnes (2021)). We train both networks to exhibit no significant performance differences and prevent overfitting. The learning curves for the MLP, achieving a test accuracy of $\mathrm{Acc}_{\mathrm{MLP}} = 67 \pm 4\%$ and CNN with $\mathrm{Acc}_{\mathrm{CNN}} = 71 \pm 2\%$ (estimated across 50 trained networks), are shown in Figures B1 and B2 respectively. Additionally, we consider the RMSE of the predicted years and see comparable RSME for the Test Data with $R_{\mathrm{MLP}} = 5.1$ and $R_{\mathrm{CNN}} = 4.5$.

In Figure B3 we also show the number of correct predictions for both architectures (all points on

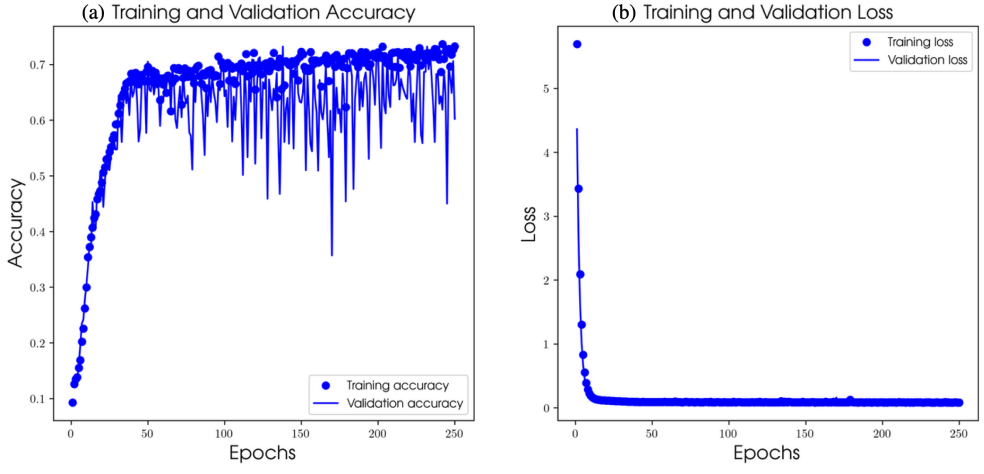

FIG. B1: Learning curve of the MLP including accuracy (a) and loss (b). In both plots, the scatter graph represents the training performance, and the line graph the performance on the validation data.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

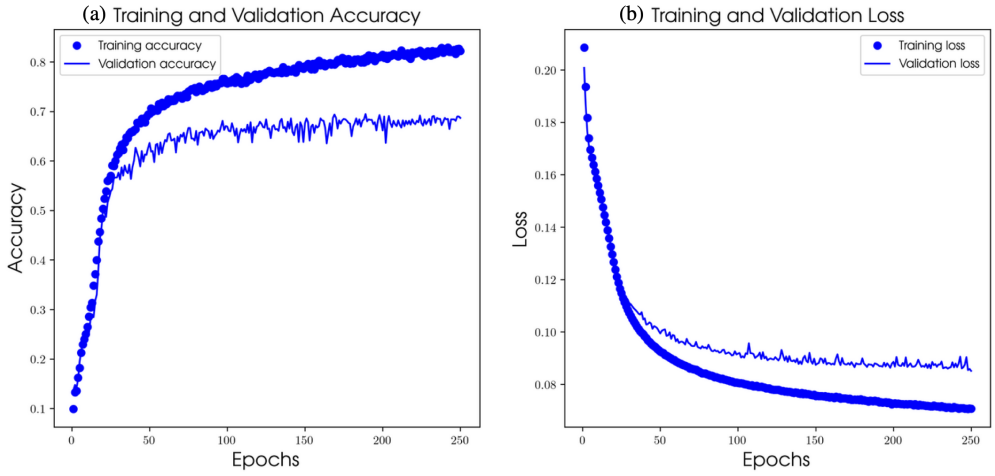

FIG. B2: Learning curve of the CNN including accuracy (a) and loss (b). In both plots, the scatter graph represents the training performance, and the line graph the performance on the validation data.

the regression line). In these graphs, we observe changing numbers of correct predictions across

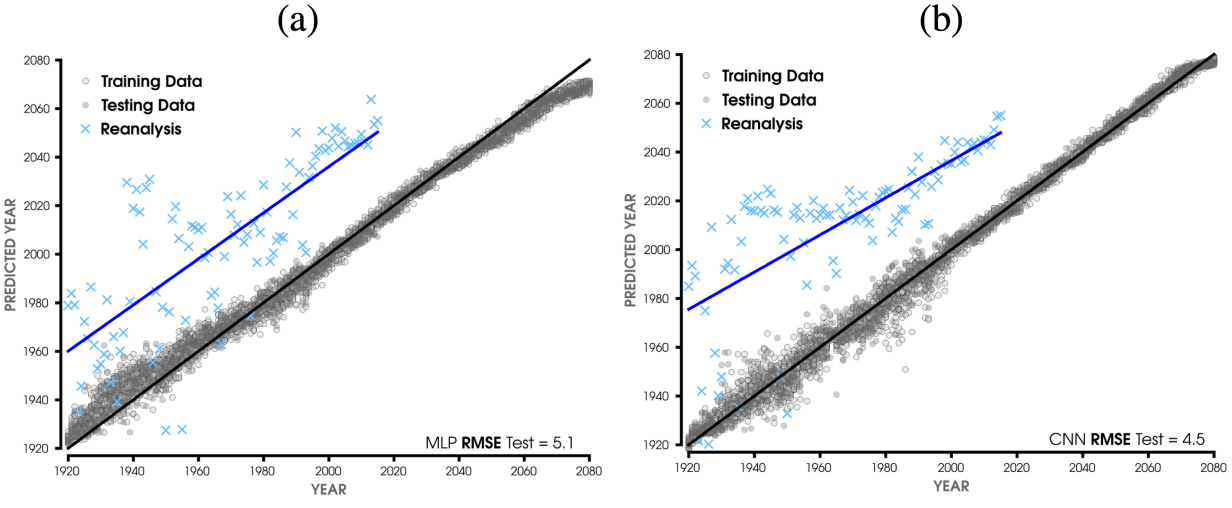

FIG. B3: Network performance based on the RMSE of the predicted years to the true years of both A) MLP and B) CNN (compare to Figure 3c in Labe and Barnes (2021)). The red dots correspond to the agreement of the predictions based on the training and validation data to the actual years and the grey dots show agreement between the predictions on the test set and the actual years, with the black line showing the linear regression across the full model data (training, validation and test data). In blue, we also included the predictions on the reanalysis data with the linear regression line in dark blue.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

different years. Thus, we apply all explanation methods to the full model data Ω, to ensure access to correct samples across all years.

We show examples of MLP and CNN across all explanation methods in figure B4 and B5. Following Labe and Barnes (2021), we adopt a criterion requiring a correct year regression within an error of ±2 years, to identify a correct prediction. We average correct predictions across ensemble members and display time periods of 40 years based on the temporal average of explanations (see Figure 6 in Labe and Barnes (2021)).

In comparison, both figures highlight the difference in spatial learning patterns, with the CNN relevance focusing on pixel groups whereas the MLP relevance can change pixel-wise. In table B1, we list the hyperparameters of the explanation methods, compared in our experiments. We use the notation introduced in Appendix A-a. We use Integrated Gradients with the baseline $\bar{\mathbf{x}}$ generated per default by iNNestigate.

b. Evaluation metrics

(i) Hyperparmeters In table B2 we list the hyperparameters of the different metrics. We list only the adapted parameters for all others (see Hedström et al. (2023b)) we used the Quantus default

TABLE B1: The hyperparameters of the XAI methods. Note that parameters vary across explanation methods. We report only adjusted parameters, for all others we write −. We denote maximum and minimum values across all temperature maps $\mathbf{X}$ in the dataset Ω as $x_{\max}$ and $x_{\min}$ respectively.

	α	β	N	M	$\sigma_{\mathbf{SG}}$	$\sigma_{\mathbf{NG}}$	$\Phi_0(f(\mathbf{x}))$	$\bar{\mathbf{x}}$
gradient	–	–	–	–	–	–	–	–
SmoothGrad	–	–	150	–	$0.25(x_{\max}-x_{\min})$	–	gradient	–
NoiseGrad	–	–	–	20	–	0.25	gradient	–
FusionGrad	–	–	20	20	$0.25(x_{\max}-x_{\min})$	0.125	gradient	–
input times gradients	–	–	–	–	–	–	–	–
Integrated Gradients	–	–	–	–	–	–	–	**0**
LRP-α-β	1	0	–	–	–	–	–	–
LRP-z	–	–	–	–	–	–	–	–
LRP-composite	–	–	–	–	–	–	–	–
DeepSHAP	–	–	–	–	–	–	–	**0**

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

TABLE B2: We show the hyperparameters of the XAI evaluation metrics based on the QUANTUS package calculations (Hedström et al. 2023b). We consider the metrics, Average Sensitivity (AS), Local Lipschitz Estimate (LLE), Faithfulness Correlation (FC), ROAD, Model Parameter Randomization Test (MPT), Random Logit (RL), Complexity (COM), Sparseness (SPA), Top-K and Relevance Rank Accuracy (RRA). Note that parameters vary across metrics and we report settings only for existing parameters in each metric (for all others we write –).

	Robustness		*Faithfulness*		*Randomization*		*Complexity*		*Localisation*	
Hyperparameters	**AS**	**LLE**	**FC**	**ROAD**	**MPT**	**RL**	**COM**	**SPA**	**TopK**	**RRA**
Normalization	True	True	True	True	True	True	True	True	True	True
Perturbation function	$\mathcal{N}(0,0.1)$	$\mathcal{N}(0,0.1)$	Indices	Linear	–	–	–	–	–	–
Similarity function	Difference	Lipschitz Constant	Pearson Corr.	–	Pearson Corr.	Pearson Corr.	–	–	–	–
Num. of samples/runs	10	10	50	–	–	–	–	–	–	–
Norm nominator	Frobenius	Euclidean	–	–	–	–	–	–	–	–
Norm denominator	Frobenius	Euclidean	–	–	–	–	–	–	–	
Subset size	–	–	40	–	–	–	–	–	–	–
Percentage range	–	–	–	1 – 50%	–	–	–	–	–	
k	–	–	–	–	–	–	–	–	$0.1d$	–
Perturbation baseline	–	–	$U(0,1)$	$U(0,1)$	–	–	–	–	–	–
Number of Classes	–	–	–	–	–	20	–	–	–	–
Layer Order	–	–	–	–	bottom_up	–	–	–	–	–

values. The normalization parameter refers to an explanation of normalization according to Eq. A12.

Faithfulness. In table B2 the perturbation function 'Indices' refers to the baseline replacement by indices of the highest value pixels in the explanation and 'Linear' refers to noisy linear imputation (see Rong et al. (2022a) for details). Please, note that the evaluation of the faithfulness property strongly depends on the choice of perturbation baseline. Thus, we advise the reader to choose the uniform baseline, as determined here for standardized weather data, as it most strongly resembles noise.

Randomization. For the Model Parameter Randomization Test score calculations, we perturb the layer weights starting from the output layer to the input layer, referred to as 'bottom_up' in table B2. To ensure comparability we use the Pearson correlation as the similarity function for both metrics.

Localisation. For top-k we consider $k = 0.1d$, which are the 10% most relevant pixels of all pixels d in the temperature map.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

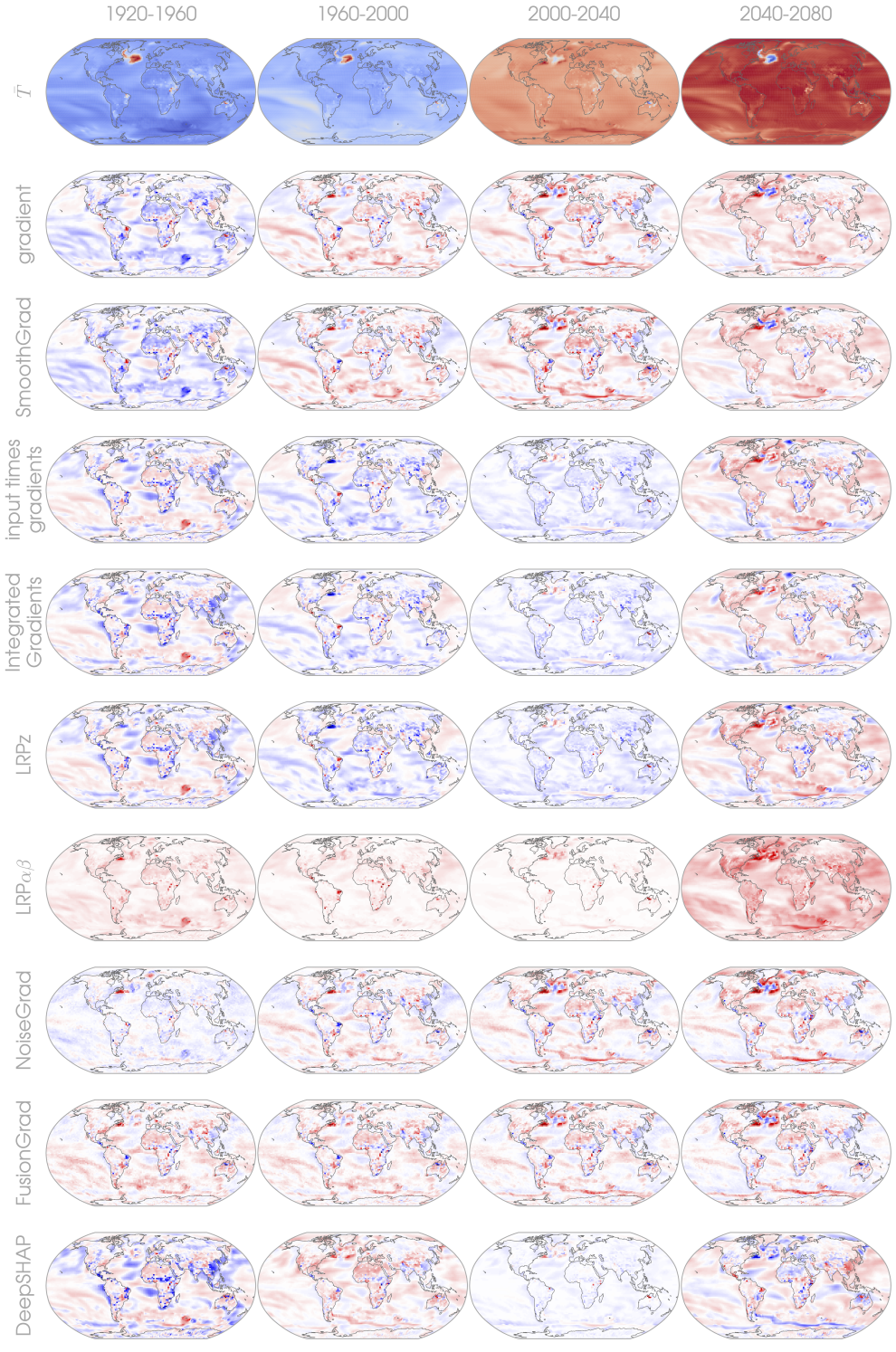

FIG. B4: MLP explanation map average over 1920 – 1960,1960 – 2000, 2000 – 2040 and 2040 – 2080 for all XAI methods. The first row shows the average input temperature map $\bar{T}$ with the color bar ranging from maximum (red) to minimum(blue) temperature anomaly. All consecutive lines show the explanation maps of the different XAI methods with the color bar ranging from 1 (red) to −1 (blue).

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

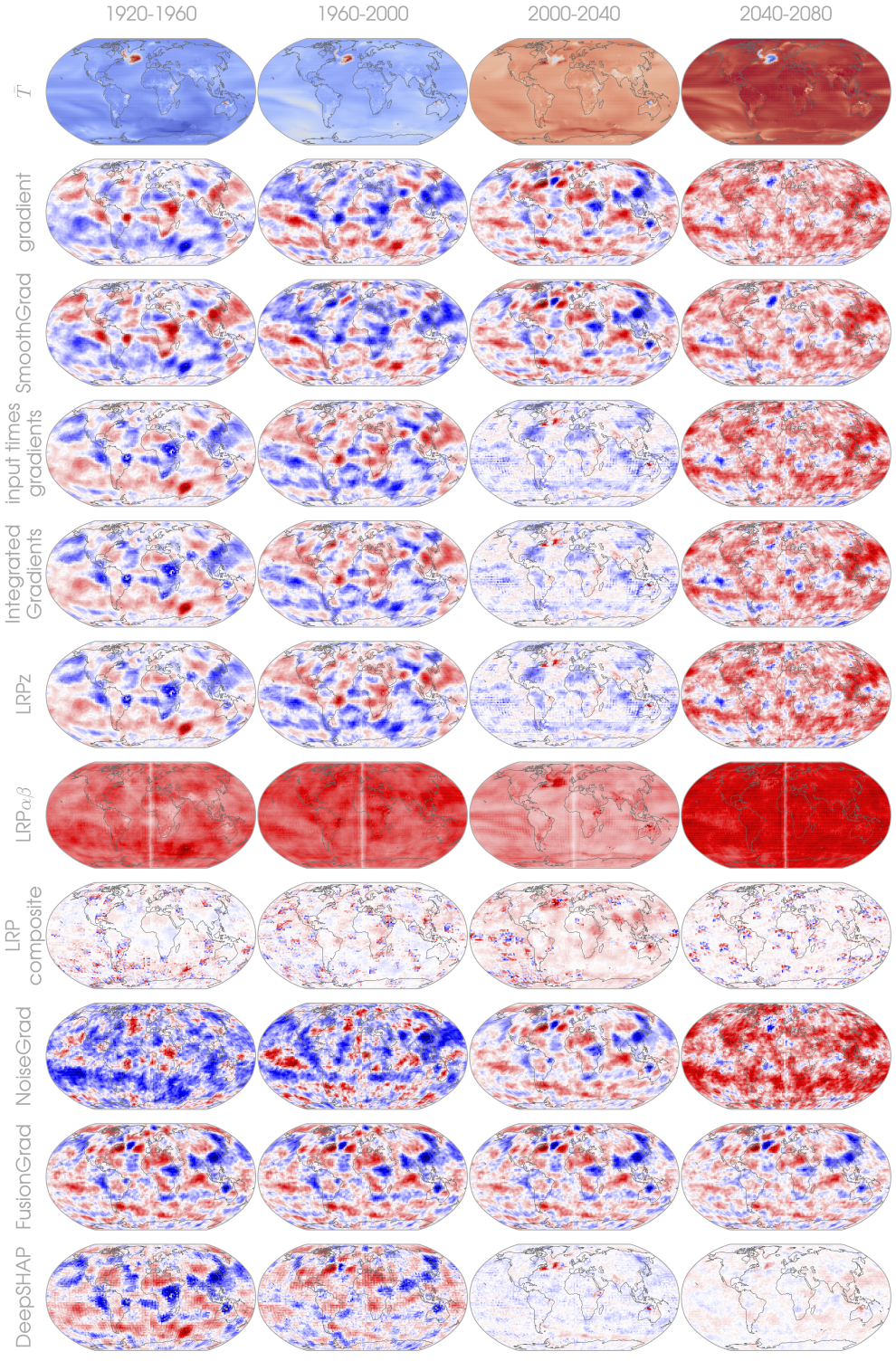

FIG. B5: CNN explanation map average over 1920 – 1960,1960 – 2000, 2000 – 2040 and 2040 – 2080 for all XAI methods. The first row shows the average input temperature map $\bar{T}$ with the color bar ranging from maximum (red) to minimum(blue) temperature anomalies. All consecutive lines show the explanation maps of the different XAI methods with the color bar ranging from 1 (red) to −1 (blue).

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

References

Abadi, M., and Coauthors, 2016: Tensorflow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, USENIX Association, USA, 265–283, OSDI'16.

Adebayo, J., J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, 2018: Sanity checks for saliency maps. *Advances in neural information processing systems*, **31**.

Agarwal, C., S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju, 2022: OpenXAI: Towards a transparent evaluation of model explanations. *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alber, M., and Coauthors, 2019: iNNvestigate neural networks! *Journal of Machine Learning Research*, **20 (93)**, 1–8, URL http://jmlr.org/papers/v20/18-540.html.

Alvarez Melis, D., and T. Jaakkola, 2018: Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, **31**.

Alvarez-Melis, D., and T. S. Jaakkola, 2018: On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Anantrasirichai, N., J. Biggs, F. Albino, and D. Bull, 2019: A deep learning approach to detecting volcano deformation from satellite imagery using synthetic datasets. *Remote Sensing of Environment*, **230**, 111 179, https://doi.org/10.1016/j.rse.2019.04.032.

Ancona, M., E. Ceolini, C. Öztireli, and M. Gross, 2019: Gradient-based attribution methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer International Publishing, 169–191, https://doi.org/10.1007/978-3-030-28954-6_9.

Arias-Duart, A., F. Parés, D. Garcia-Gasulla, and V. Giménez-Ábalos, 2022: Focus! rating xai methods and finding biases. *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8.

Arras, L., A. Osman, and W. Samek, 2020: Ground truth evaluation of neural network explanations with CLEVR-XAI. *arXiv preprint arXiv:2003.07258*.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Arrieta, A. B., and Coauthors, 2020: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, **58**, 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, 2015: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, **10 (7)**, e0130 140, https://doi.org/10.1371/journal.pone.0130140.

Baehrens, D., T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, 2010: How to explain individual classification decisions. *The Journal of Machine Learning Research*, **11**, 1803–1831.

Balduzzi, D., M. Frean, L. Leary, J. Lewis, K. W.-D. Ma, and B. McWilliams, 2017: The shattered gradients problem: If resnets are the answer, then what is the question? *CoRR*, **abs/1702.08591**, URL http://arxiv.org/abs/1702.08591, 1702.08591.

Barnes, E. A., R. J. Barnes, and N. Gordillo, 2021: Adding uncertainty to neural network regression tasks in the geosciences. *arXiv preprint arXiv:2109.07250*.

Barnes, E. A., B. Toms, J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2020: Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, **12 (9)**, e2020MS002 195.

Bhatt, U., A. Weller, and J. M. Moura, 2020: Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*.

Brocki, L., and N. C. Chung, 2022: Evaluation of interpretability methods and perturbation artifacts in deep neural networks. *CoRR*, **abs/2203.02928**, https://doi.org/10.48550/arXiv.2203.02928, 2203.02928.

Bromberg, C. L., C. Gazen, J. J. Hickey, J. Burge, L. Barrington, and S. Agrawal, 2019: Machine learning for precipitation nowcasting from radar images. *Workshop at the 33rd Conference on Neural Information Processing Systems*, 4.

Bykov, K., M. Deb, D. Grinwald, K.-R. Müller, and M. M.-C. Höhne, 2022a: Dora: Exploring outlier representations in deep neural networks. *arXiv preprint arXiv:2206.04530*.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Bykov, K., A. Hedström, S. Nakajima, and M. M.-C. Höhne, 2022b: Noisegrad: enhancing explanations by introducing stochasticity to model weights. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, 6132–6140.

Bykov, K., M. M.-C. Höhne, A. Creosteanu, K.-R. Müller, F. Klauschen, S. Nakajima, and M. Kloft, 2021: Explaining bayesian neural networks. *arXiv preprint*, 2108.10346.

Camps-Valls, G., M. Reichstein, X. Zhu, and D. Tuia, 2020: Advancing Deep Learning for Earth Sciences From Hybrid Modeling To Interpretability. *IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2020, Waikoloa, HI, USA, September 26 - October 2, 2020*, IEEE, 3979–3982, https://doi.org/10.1109/IGARSS39084.2020.9323558.

Chalasani, P., J. Chen, A. R. Chowdhury, S. Jha, and X. Wu, 2020: Concise explanations of neural networks using adversarial training. *Proceedings of the 37th International Conference on Machine Learning*, JMLR.org, ICML'20.

Chen, C., O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, 2019: This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, **32**.

Chen, K., P. Wang, X. Yang, N. Zhang, and D. Wang, 2020: A model output deep learning method for grid temperature forecasts in tianjin area. *Applied Sciences*, **10 (17)**, 5808, https://doi.org/10.3390/app10175808.

Clare, M. C., M. Sonnewald, R. Lguensat, J. Deshayes, and V. Balaji, 2022: Explainable artificial intelligence for bayesian neural networks: toward trustworthy predictions of ocean dynamics. *Journal of Advances in Modeling Earth Systems*, **14 (11)**, e2022MS003 162, https://doi.org/10.1002/essoar.10511239.1.

Covert, I. C., S. Lundberg, and S.-I. Lee, 2021: Explaining by removing: A unified framework for model explanation. *The Journal of Machine Learning Research*, **22 (1)**, 9477–9566.

Das, A., and P. Rad, 2020: Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Dikshit, A., and B. Pradhan, 2021: Interpretable and explainable ai (xai) model for spatial drought prediction. *Science of the Total Environment*, **801**, 149 797, https://doi.org/10.1016/j.scitotenv.2021.149797.

Ebert-Uphoff, I., and K. Hilburn, 2020: Evaluation, tuning and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society*, 1–47, https://doi.org/10.1175/bams-d-20-0097.1.

Felsche, E., and R. Ludwig, 2021: Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations. *Natural Hazards and Earth System Sciences*, **21 (12)**, 3679–3691.

Flora, M., C. Potvin, A. McGovern, and S. Handler, 2022: Comparing explanation methods for traditional machine learning models part 1: An overview of current methods and quantifying their disagreement. *arXiv preprint arXiv:2211.08943*.

Gautam, S., A. Boubekki, S. Hansen, S. Salahuddin, R. Jenssen, M. Höhne, and M. Kampffmeyer, 2022: Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, **35**, 17 940–17 952.

Gautam, S., M. M.-C. Höhne, S. Hansen, R. Jenssen, and M. Kampffmeyer, 2023: This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. *Pattern Recognition*, **136**, 109 172.

Gevaert, A., A.-J. Rousseau, T. Becker, D. Valkenborg, T. De Bie, and Y. Saeys, 2022: Evaluating feature attribution methods in the image domain. *CoRR*, **abs/2202.12270**, URL https://arxiv.org/abs/2202.12270, 2202.12270.

Gibson, P. B., W. E. Chapman, A. Altinok, L. Delle Monache, M. J. DeFlorio, and D. E. Waliser, 2021: Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment*, **2 (1)**, 159, https://doi.org/10.1038/s43247-021-00225-4.

Grinwald, D., K. Bykov, S. Nakajima, and M. M.-C. Höhne, 2022: Visualizing the diversity of representations learned by bayesian neural networks. *arXiv preprint arXiv:2201.10859*.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Ham, Y.-G., J.-H. Kim, and J.-J. Luo, 2019: Deep learning for multi-year enso forecasts. *Nature*, **573 (7775)**, 568–572.

Han, L., J. Sun, W. Zhang, Y. Xiu, H. Feng, and Y. Lin, 2017: A machine learning nowcasting method based on real-time reanalysis data. *Journal of Geophysical Research: Atmospheres*, **122 (7)**, 4038–4051, https://doi.org/10.1002/2016jd025783.

Han, T., S. Srinivas, and H. Lakkaraju, 2022: Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *arXiv preprint arXiv:2206.01254*.

Harder, P., D. Watson-Parris, D. Strassel, N. Gauger, P. Stier, and J. Keuper, 2021: Emulating aerosol microphysics with machine learning. *arXiv preprint arXiv:2109.10593*.

Harris, C. R., and Coauthors, 2020: Array programming with NumPy. *Nature*, **585 (7825)**, 357–362, https://doi.org/10.1038/s41586-020-2649-2.

He, S., X. Li, T. DelSole, P. Ravikumar, and A. Banerjee, 2021: Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 169–177.

Hedström, A., P. Bommer, K. K. Wickstrøm, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, 2023a: The meta-evaluation problem in explainable ai: Identifying reliable estimators with metaquantus. *arXiv preprint arXiv:2302.07265*.

Hedström, A., L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, 2023b: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, **24 (34)**, 1–11.

Hengl, T., and Coauthors, 2017: SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, **12 (2)**, e0169 748, https://doi.org/10.1371/journal.pone.0169748.

Hilburn, K. A., 2023: Understanding spatial context in convolutional neural networks using explainable methods: Application to interpretable GREMLIN. *Artificial Intelligence for the Earth Systems*, **2 (3)**, https://doi.org/10.1175/aies-d-22-0093.1.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Hilburn, K. A., I. Ebert-Uphoff, and S. D. Miller, 2021: Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using goes-r satellite observations. *Journal of Applied Meteorology and Climatology*, **60 (1)**, 3–21, https://doi.org/10.1175/jamc-d-20-0084.1.

Hoffman, R. R., S. T. Mueller, G. Klein, and J. Litman, 2018: Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hunter, J. D., 2007: Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, **9 (3)**, 90–95, https://doi.org/10.1109/mcse.2007.55.

Hurley, N., and S. Rickard, 2009: Comparing measures of sparsity. IEEE, 4723–4741 pp., https://doi.org/10.1109/mlsp.2008.4685455.

Hurrell, J. W., and Coauthors, 2013: The community earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, **94 (9)**, 1339–1360, https://doi.org/10.1175/bams-d-12-00121.1.

Janzing, D., L. Minorics, and P. Blöbaum, 2020: Feature relevance quantification in explainable ai: A causal problem. *International Conference on artificial intelligence and statistics*, 2907–2916.

Kay, J. E., and Coauthors, 2015: The community earth system model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, **96 (8)**, 1333–1349, https://doi.org/10.1175/bams-d-13-00255.1.

Krishna, S., T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, 2022: The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Labe, Z. M., and E. A. Barnes, 2021: Detecting climate signals using explainable AI with single-forcing large ensembles. *Journal of Advances in Modeling Earth Systems*, **13 (6)**, https://doi.org/10.1029/2021ms002464.

Labe, Z. M., and E. A. Barnes, 2022: Comparison of climate model large ensembles with observations in the arctic using simple neural networks. *Earth and Space Science*, **9 (7)**, e2022EA002 348, https://doi.org/10.1002/essoar.10510977.1.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, 2019: Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, **10**, https://doi.org/10.1038/s41467-019-08987-4.

Leinonen, J., D. Nerini, and A. Berne, 2021: Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, **59 (9)**, 7211–7223, https://doi.org/10.1109/tgrs.2020.3032790.

Letzgus, S., P. Wagner, J. Lederer, W. Samek, K.-R. Müller, and G. Montavon, 2022: Toward explainable artificial intelligence for regression models: A methodological perspective. *IEEE Signal Processing Magazine*, **39 (4)**, 40–58.

Lundberg, S. M., and S.-I. Lee, 2017: A Unified Approach to Interpreting Model Predictions. *Advances in neural information processing systems*, **30**.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2022a: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, 1–42, https://doi.org/10.1175/aies-d-22-0012.1.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2020: Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, Springer, 315–339.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2022b: Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science*, **1**, e8.

Mayer, K. J., and E. A. Barnes, 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophysical Research Letters*, **48 (10)**, e2020GL092 092, https://doi.org/10.1029/2020gl092092.

McGovern, A., R. Lagerquist, D. John Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, **100 (11)**, 2175–2199, https://doi.org/10.1175/bams-d-18-0195.1.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Mohseni, S., N. Zarei, and E. D. Ragan, 2021: A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, **11 (3-4)**, 1–45.

Montavon, G., A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, 2019: Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193–209.

Montavon, G., S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, 2017: Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, **65**, 211–222.

Montavon, G., W. Samek, and K.-R. Müller, 2018: Methods for interpreting and understanding deep neural networks. *Digital signal processing*, **73**, 1–15, https://doi.org/10.1016/j.dsp.2017.10.011.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly weather review*, **116 (12)**, 2417–2424.

Murphy, A. H., and H. Daan, 1985: Forecast evaluation. *Probability, statistics, and decision making in the atmospheric sciences*, 379–437.

Nguyen, A., A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, 2016: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, 3387–3395, URL https://proceedings.neurips.cc/paper/2016/hash/5d79099fcdf499f12b79770834c0164a-Abstract.html.

Nguyen, A.-p., and M. R. Martínez, 2020: On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*.

Pegion, K., E. J. Becker, and B. P. Kirtman, 2022: Understanding predictability of daily southeast u.s. precipitation using explainable machine learning. *Artificial Intelligence for the Earth Systems*, **1 (4)**, https://doi.org/10.1175/aies-d-22-0011.1.

Petsiuk, V., A. Das, and K. Saenko, 2018: Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Ribeiro, M. T., S. Singh, and C. Guestrin, 2016: "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rieger, L., and L. K. Hansen, 2020: Irof: a low resource evaluation metric for explanation methods. *arXiv preprint arXiv:2003.08747*.

Rong, Y., T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, 2022a: A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*.

Rong, Y., T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, 2022b: Evaluating feature attribution: An information-theoretic perspective. *arXiv preprint arXiv:2202.00449*.

Samek, W., A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Muller, 2017: Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, **28**, 2660–2673, https://doi.org/10.1109/TNNLS.2016.2599820.

Samek, W., G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, 2019: *Explainable AI: interpreting, explaining and visualizing deep learning*, Vol. 11700. Springer Nature.

Sawada, Y., and K. Nakamura, 2022: C-senn: Contrastive self-explaining neural network. *arXiv preprint arXiv:2206.09575*.

Scher, S., and G. Messori, 2021: Ensemble methods for neural network-based weather forecasts. *Journal of Advances in Modeling Earth Systems*, **13 (2)**, https://doi.org/10.1029/2020ms002331.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 2017: Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shapley, L. S., 1951: Notes on the n-person game—ii: The value of an n-person game.(1951). *Lloyd S Shapley*.

Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, 2015: Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, **28**.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Shrikumar, A., P. Greenside, A. Shcherbina, and A. Kundaje, 2016: Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Simonyan, K., A. Vedaldi, and A. Zisserman, 2014: Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd International Conference on Learning Representations, ICLR 2014, Workshop Track Proceedings*, URL http://arxiv.org/abs/1312.6034.

Sixt, L., M. Granz, and T. Landgraf, 2020: When explanations lie: Why many modified bp attributions fail. *International Conference on Machine Learning*, 9046–9057.

Slivinski, L. C., and Coauthors, 2019: Towards a more reliable historical reanalysis: Improvements for version 3 of the twentieth century reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, **145 (724)**, 2876–2908, https://doi.org/10.1002/qj.3598.

Smilkov, D., N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, 2017: Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Sonnewald, M., and R. Lguensat, 2021: Revealing the impact of global heating on north atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, **13 (8)**, e2021MS002 496, https://doi.org/10.1029/2021ms002496.

Strumbelj, E., and I. Kononenko, 2010: An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, **11**, 1–18.

Sturmfels, P., S. Lundberg, and S.-I. Lee, 2020: Visualizing the impact of feature attribution baselines. *Distill*, **5 (1)**, e22, https://doi.org/10.23915/distill.00022.

Sundararajan, M., A. Taly, and Q. Yan, 2017: Axiomatic attribution for deep networks. *International conference on machine learning*, PMLR, 3319–3328.

Theiner, J., E. Müller-Budack, and R. Ewerth, 2022: Interpretable semantic photo geolocation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 750–760.

Tomsett, R., D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, ????: Sanity Checks for Saliency Metrics. *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, 6021–6029, https://doi.org/10.1609/aaai.v34i04.6064.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Van Straaten, C., K. Whan, D. Coumou, B. Van den Hurk, and M. Schmeits, 2022: Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in western and central europe. *Monthly Weather Review*, **150 (5)**, 1115–1134, https://doi.org/10.1175/mwr-d-21-0201.1.

Vidovic, M. M.-C., N. Görnitz, K.-R. Müller, and M. Kloft, 2016: Feature importance measure for non-linear learning algorithms. *arXiv preprint arXiv:1611.07567*.

Vidovic, M. M.-C., N. Görnitz, K.-R. Müller, G. Rätsch, and M. Kloft, 2015: Opening the black box: Revealing interpretable sequence motifs in kernel-based learning algorithms. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II 15*, Springer, 137–153.

Virtanen, P., and Coauthors, 2020: SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, **17 (3)**, 261–272, https://doi.org/10.1038/s41592-019-0686-2.

Wang, J., K. Gao, Z. Zhang, C. Ni, Z. Hu, D. Chen, and Q. Wu, 2021: Multisensor remote sensing imagery super-resolution with conditional gan. *Journal of Remote Sensing*, **2021**, https://doi.org/10.34133/2021/9829706.

Yang, M., and B. Kim, 2019: Benchmarking attribution methods with relative feature importance. *arXiv e-prints*, arXiv:1907.09701, https://doi.org/10.48550/arXiv.1907.09701, URL https://ui.adsabs.harvard.edu/abs/2019arXiv190709701Y, 1907.09701.

Yeh, C.-K., C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, 2019: On the (in)fidelity and sensitivity for explanations. *Advances in Neural Information Processing Systems*, **32**.

Yuval, J., and P. A. O'Gorman, 2020: Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature communications*, **11 (1)**, 3295, https://doi.org/10.1038/s41467-020-17142-3.

Zhang, J., S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, 2018: Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, **126 (10)**, 1084–1102, https://doi.org/10.1007/s11263-017-1059-x.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.

Zhou, Y., S. Booth, M. T. Ribeiro, and J. Shah, 2022: Do feature attribution methods correctly attribute features? *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, 9623–9633.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0074.1.