

# Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience

ANTONIOS MAMALAKIS,<sup>a</sup> ELIZABETH A. BARNES,<sup>a</sup> AND IMME EBERT-UPHOFF<sup>b,c</sup>

<sup>a</sup> *Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado*

<sup>b</sup> *Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado*

<sup>c</sup> *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*

(Manuscript received 9 February 2022, in final form 9 June 2022)

**ABSTRACT:** Convolutional neural networks (CNNs) have recently attracted great attention in geoscience because of their ability to capture nonlinear system behavior and extract predictive spatiotemporal patterns. Given their black-box nature, however, and the importance of prediction explainability, methods of explainable artificial intelligence (XAI) are gaining popularity as a means to explain the CNN decision-making strategy. Here, we establish an intercomparison of some of the most popular XAI methods and investigate their fidelity in explaining CNN decisions for geoscientific applications. Our goal is to raise awareness of the theoretical limitations of these methods and to gain insight into the relative strengths and weaknesses to help guide best practices. The considered XAI methods are first applied to an idealized attribution benchmark, in which the ground truth of explanation of the network is known a priori, to help objectively assess their performance. Second, we apply XAI to a climate-related prediction setting, namely, to explain a CNN that is trained to predict the number of atmospheric rivers in daily snapshots of climate simulations. Our results highlight several important issues of XAI methods (e.g., gradient shattering, inability to distinguish the sign of attribution, and ignorance to zero input) that have previously been overlooked in our field and, if not considered cautiously, may lead to a distorted picture of the CNN decision-making strategy. We envision that our analysis will motivate further investigation into XAI fidelity and will help toward a cautious implementation of XAI in geoscience, which can lead to further exploitation of CNNs and deep learning for prediction problems.

**KEYWORDS:** Artificial intelligence; Classification; Data science; Deep learning; Model interpretation and visualization; Neural networks

## 1. Introduction

In recent years, convolutional neural networks (CNNs) and deep learning in general have seen great application in a plethora of problems in geoscience (Lary et al. 2016; Karpatne et al. 2019; Reichstein et al. 2019), ranging from solid Earth science (Bergen et al. 2019), marine science and hydrology (Shen 2018; Sit et al. 2020) to climate science and meteorology (Barnes et al. 2019; Rolnick et al. 2019; Ham et al. 2019). The popularity of CNNs has risen mainly because of their ability to capture nonlinear system behavior and to extract predictive spatiotemporal patterns (LeCun et al. 2015), which makes them of particular interest to geoscientists. Another important reason is the increasing availability of observational and simulated data in this decade (Overpeck et al. 2011; Guo 2017; Agapiou 2017; Reinsel et al. 2018) that helps to meet the requirement to train CNNs with large datasets.

Despite their potential, an important issue about the application of CNNs in the geosciences is their black-box nature, which makes it hard for scientists to interpret predictions and to assess the model from a physical perspective, that is,

---

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/AIES-D-22-0012.s1>.

---

*Corresponding author:* Antonios Mamalakis, [amamalak@rams.colostate.edu](mailto:amamalak@rams.colostate.edu)

beyond using prediction performance as the only criterion. The interpretability issue is considered a key issue for deep learning in general, and it has prompted the emergence of a new subfield in computer science, namely, explainable artificial intelligence (XAI; Buhrmester et al. 2019; Tjoa and Guan 2019; Das and Rad 2020). The goal of XAI methods is to explain, in a postprediction setting (typically referred to as *post hoc explanation*), the decision strategy of a model that otherwise is inherently not interpretable. One common way to do this is to highlight the most important variables in the input space (typically referred to as *features*) that helped the model to make a specific prediction. These methods are referred to as “local” XAI methods because they focus on a specific prediction, in contrast to “global” XAI methods that identify important features across all samples (Buhrmester et al. 2019).

XAI methods have already proven to be of great utility for explaining black-box models in computer science and beyond, and they have also seen recent application in geoscience (McGovern et al. 2019; Ebert-Uphoff and Hilburn 2020; Toms et al. 2020; Mamalakis et al. 2022b). Specifically, recent work shows how XAI can help to calibrate model trust (Sonnewald and Lguensat 2021; Mayer and Barnes 2021; Hilburn et al. 2021; Keys et al. 2021), to identify ways to fine-tune models that are performing poorly (Ebert-Uphoff and Hilburn 2020), and also to accelerate learning new science (Barnes et al. 2020; Toms et al. 2021). The results of these recent studies indicate that XAI can be a real game changer for prediction and classification problems (Mamalakis et al. 2022b)

DOI: 10.1175/AIES-D-22-0012.1 e220012

© 2022 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) ([www.ametsoc.org/PUBSReuseLicenses](http://www.ametsoc.org/PUBSReuseLicenses)).

and can help to further exploit the potential of deep learning in geoscience in our era of big data.

Despite the above, many XAI methods have been shown to exhibit theoretical and practical limitations in explaining black-box models (Sundararajan et al. 2017; Kindermans et al. 2019; Ancona et al. 2018; Rudin 2019; Dombrowski et al. 2022; Zhou et al. 2022). Moreover, XAI results are not typically assessed on the basis of a ground truth of explanation but rather are based on the subjective evaluation by the analyst/scientist about whether the explanation is physically reasonable. However, even if an explanation makes physical sense to a human, it does not necessarily mean that this is the strategy the model in question is actually using (and vice versa). In other words, the human perception of an explanation alone is not a solid criterion for assessing its trustworthiness. Also, what physically makes sense depends on the a priori understanding of the problem that the scientist has and, thus, might be different across individuals, especially in problem settings of high levels of complexity. The theoretical and practical limitations of XAI methods, together with the issue of subjectivity in their assessment that may propagate individual biases, have been recognized in the literature (Leavitt and Morcos 2020) and call for a more objective and systematic investigation of XAI methods' fidelity for a range of different applications and model architectures.

In an effort to introduce more objectivity in the assessment of XAI methods for geosciences, our group proposed a generic approach to develop simple *attribution benchmark datasets* for benchmarking XAI methods (Mamalakis et al. 2022a). Attribution benchmark datasets consist of synthetic inputs and outputs, where the functional relationship between the two is known. This allows for deriving the ground truth of what the explanation of the network should look like for each prediction. In this way, the assessment of XAI methods is no longer based on subjective criteria, but rather it is based on the direct comparison of the XAI results with the ground truth of the explanation. As a first example, Mamalakis et al. (2022a) generated a large attribution benchmark inspired from a climate prediction setting and applied XAI methods to explain the predictions of a fully connected neural network. Other studies have also developed similar benchmarks in the field of computer science (Arras et al. 2021; Zhou et al. 2022).

Here, we build on previous studies that deal with the assessment and the benchmarking of XAI methods and we shift our focus to *convolutional* neural networks, with the aim to investigate XAI fidelity in CNN applications in the geosciences. Our goal is to raise awareness of the theoretical limitations of XAI methods and gain insight into the relative strengths and weaknesses to help guide best practices. We focus on some of the most popular XAI methods (e.g., Gradient, Smooth Gradient, Integrated Gradients, Layerwise Relevance Propagation, among many others)<sup>1</sup> and apply them to explain the

predictions of CNNs for two specific classification problems. First, we consider an idealized attribution benchmark dataset, where the CNN is trained to classify pictures of circular and square frames depending on which class of frames covers more area. The simplicity of the prediction task allows us to derive the ground truth of the explanation and assess XAI methods in an objective manner. Thus, this first problem helps us to gain insight into limitations that might be overlooked in cases for which no ground truth of the explanation is available. In the second problem, we consider a prediction setting with a climate-related task, namely, predicting the number of the atmospheric rivers in daily snapshots of climate simulations. In this setting there is no ground truth of the explanation, as is the case in most geophysical studies. The second problem aims to validate our insights about XAI in a more climate-related setting and to illustrate how explanations should be regarded and interpreted so as to avoid reaching false conclusions about the strategy of the network.

In section 2, we introduce the two datasets, discuss CNN architectures and prediction performance, and describe the XAI methods considered in the study. In section 3, we present and discuss the results of the XAI methods when applied to explain the CNN decision strategy, and in section 4, we state our conclusions.

## 2. Data and methods

### a. Synthetic attribution benchmark

For our first classification problem we develop and use a synthetic attribution benchmark dataset to objectively assess XAI methods. An attribution benchmark consists of a synthetic input  $\mathbf{X}$  and a synthetic output  $Y$ , with the latter being a known function  $F$  of the former (Mamalakis et al. 2022a). On the functional form of  $F$ , Mamalakis et al. (2022a) noted that the function  $F$  can be of any arbitrary choice (depending on what type of network the analyst wants to benchmark: fully connected, a CNN, etc.), as long as it has such a form so that the attribution of any output to the corresponding input is objectively derivable.

We herein consider an idealized classification task that is specifically designed for CNN applications and is inspired by remote sensing tasks in geosciences where spatial patterns (such as cloud objects, weather fronts, etc.) need to be tracked and extracted (see, e.g., Hilburn et al. 2021). We generate a series of inputs that consist of 2D (single channel) images where circular and square frames are present, and we task the CNN to classify each image depending on which class of frames covers more area (i.e., has more pixels). More specifically, each input image consists of  $65 \times 65$  pixels (i.e., a total of  $d = 4225$  pixels), with the input features being binary variables,  $\mathbf{X} \in \{0, 1\}^d$ . If a pixel  $i$  for a sample  $n$  belongs to a circular or a square frame then  $x_{i,n} = 1$  (see, e.g., the dark-red pixels of the square frame in the top-left plot of Fig. 1), whereas  $x_{i,n} = 0$  otherwise; see Fig. 1 for some examples of the synthetic input images. The number of frames per class, the size of the frames, and the positioning of the frames are random in each image, and no frame overlap is allowed to occur. In terms of the output of the dataset, two separate classes into

<sup>1</sup> Ed. Note: Standard American Meteorological Society style is to lowercase method names, but for consistency with previous XAI literature and to allow easy visual identification in this paper, the authors' use of capitalization in XAI method names will be retained.

## Examples of input of the synthetic dataset

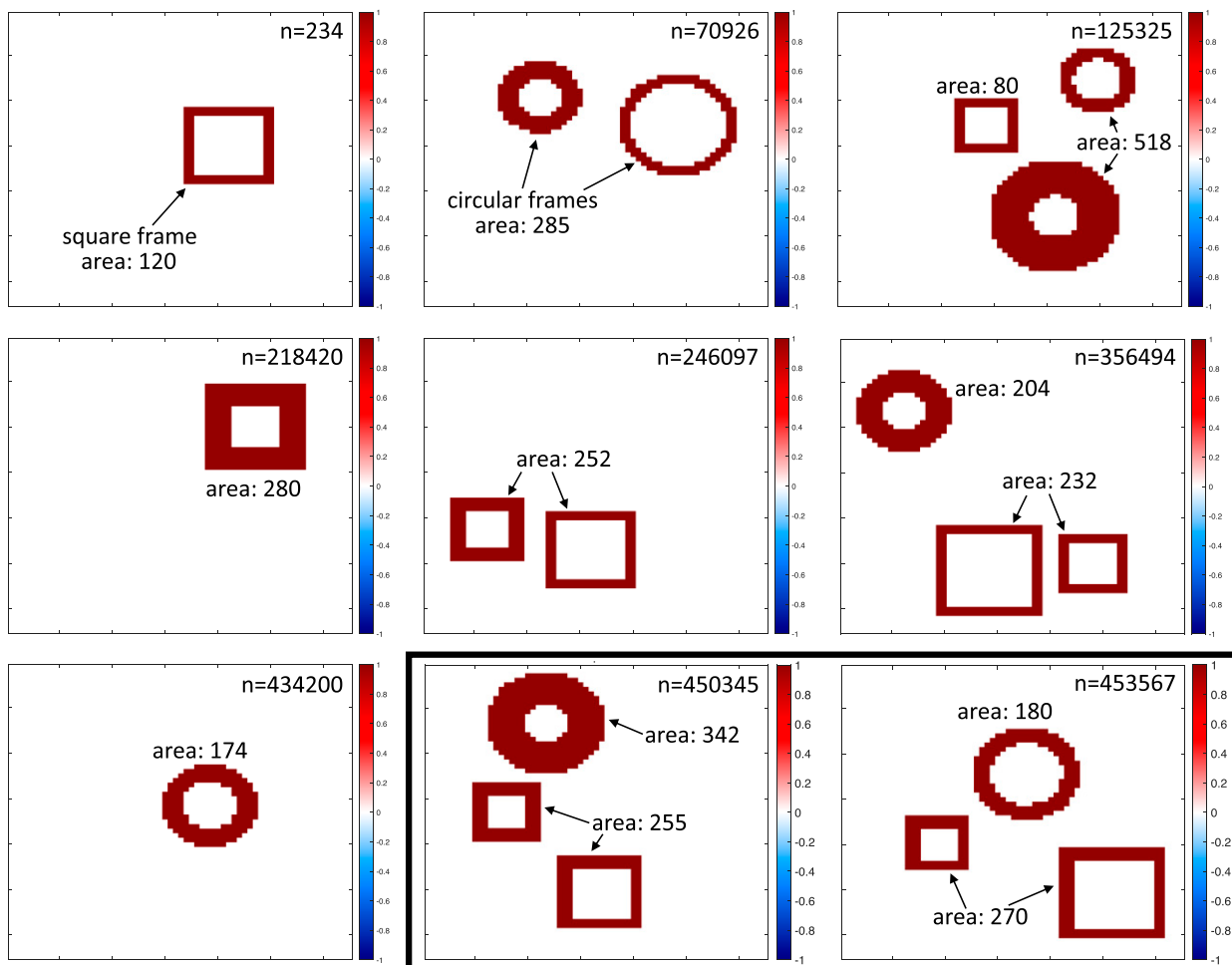


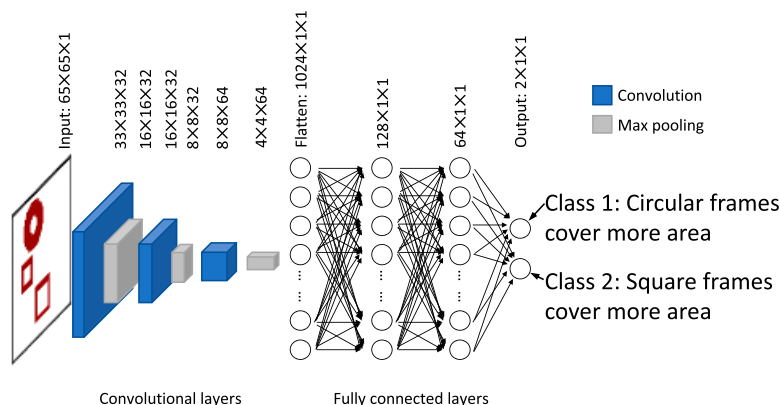
FIG. 1. Examples of input of the synthetic attribution benchmark dataset. Details about this synthetic dataset are provided in [section 2a](#). In many cases, the answer as to which class of frames covers more area is easy to get simply with visual inspection. However, there are also cases in which the answer is more difficult to disentangle (such as sample  $n = 356494$  or  $450345$ ). The testing performance of the trained CNN (see architecture in [Fig. 2a](#), below) was slightly above 99% accuracy, which is higher performance than what a human eye would do. The examples highlighted in the black-outlined box are from the testing dataset and are analyzed further in [section 3](#) (see [Figs. 3](#) and [4](#), below).

which all input images are classified are defined: in class 1, the circular frames in the image cover more area than the square frames; in class 2, the square frames in the image cover more area than the circular frames. Thus, the synthetic output of the dataset is a series of logical values indicating to which class each input image corresponds, and the network is trained to classify the images between the two classes.

By choosing this simple, idealized classification task we achieve three things. First, for any model to be able to correctly classify these synthetic images, it needs to be able to extract spatial patterns of different shapes. This makes a CNN (our study's focus) be the most suitable type of network to address this classification task ([LeCun et al. 2015](#)). Second, the simplicity of the current task makes it possible for us to objectively derive the ground truth of the attribution: pixels in an

image that belong to any circular or square frame contribute positively or negatively, respectively, to the probability that class 1 is true and negatively or positively, respectively, to the likelihood of class 2; note that this is valid when considering a blank image as the baseline. Third, an immediate consequence of the latter rule of attribution is that (as we will see in [Figs. 3–6](#), described in more detail below) there are many cases in which both positive and negative contributions appear in the same explanation. This means that with this benchmark, we can assess which XAI methods can disentangle the sign of the contribution of specific input patterns to the output, an aspect that is often overlooked ([Kohlbrenner et al. 2020](#)). In summary, this synthetic dataset fits our current scope to objectively assess XAI for CNN applications; for the connection of this dataset with the mathematical framework

## a) Convolutional Neural Network used for the synthetic benchmark



## b) Convolutional Neural Network used for the ClimateNet dataset

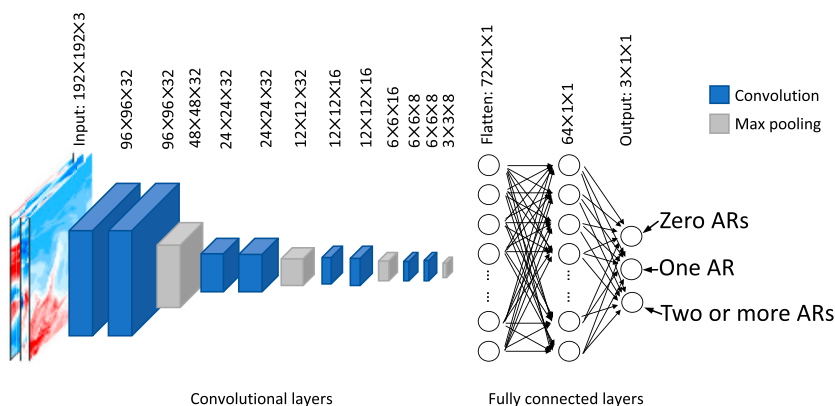


FIG. 2. Specific architectures of the CNNs that were used in the two classification problems of our study.

of *additively separable functions* introduced by Mamalakis et al. (2022a), see appendix A.

The CNN that we use for this classification task consists of three pairs of convolutional and max pooling layers followed by three fully connected layers (Fig. 2a). The output layer consists of two neurons, with the first or the second respectively calculating the likelihood that the circular or square frames cover more area. We use rectified linear unit (ReLU) activations in all layers apart from the output layer, where we use the softmax function. The CNN is trained using 450 000 samples, and 50 000 samples are used for testing. The testing classification accuracy is slightly above 99%; that is, less than 1% of the testing images are misclassified by the network. The reason that we chose to generate such an unrealistically high sample size is so that the CNN can learn almost perfectly the underlying function  $F$ . Only under this condition is it fair to use the ground truth of attribution as a benchmark for the XAI methods since any deviation between the two should mostly arise from XAI limitations and to a lesser degree from poor training of the

network. However, we note that discrepancies between XAI output and the ground truth shall always exist because the CNN is a close approximation (and not identical) to the function  $F$ .

### b. ClimateNet dataset

As a second application, we employ a more climate-related task where there is no ground truth of the explanation available (as is the case in most geophysical studies). This second task aims to validate the insights about XAI gained from the first task in a more climate-related setting. For our second classification problem we use the ClimateNet dataset (Prabhat et al. 2021). The ClimateNet dataset is a publicly available dataset (<https://portal.nersc.gov/project/ClimateNet/>) that consists of daily outputs of climate simulations from the Community Atmospheric Model (CAM5.1). Each daily output includes snapshots of many different variables like precipitation, vertically integrated precipitable water and temperature and wind velocities at different pressure levels. Also, for each simulated day in the ClimateNet dataset a labeled world map is available,

where expert meteorologists and scientists have detected the locations over which atmospheric rivers (AR: narrow elongated bands of enhanced water vapor in the atmosphere) and tropical cyclones occur around the world on that specific day. This labeling has currently been done for 456 days of the simulated historical years 1996–2013, which is the total sample size of the dataset (see an example of a simulated day in Supplementary Fig. 1 in the online supplemental material).

We build a CNN to classify these daily snapshots from the ClimateNet dataset in terms of how many ARs occur on the corresponding day. More specifically, we use a three-channel image as our input with zonal and meridional wind velocities at 850-hPa pressure level and vertically integrated precipitable water constituting the three channels. Based on the expert labeling that is available in the dataset, the CNN is then trained to classify the input into three different classes: zero, one, and two or more ARs occurring on the corresponding simulated day. The architecture of the CNN consists of four sets of two convolutional layers and one max pooling layer followed by two fully connected layers (see Fig. 2b). We use ReLU activations in all layers of the network apart from the output layer, where we use the softmax function. The output layer consists of three neurons, with the first, second, and third neuron computing the likelihood that zero, one, and two or more ARs occur on the simulated day, respectively. Because a size of 456 samples is small to train and test a deep CNN, we cut each snapshot in the dataset into six equally sized segments (three segments in each hemisphere; see an example in Supplementary Fig. 1 in the online supplemental material). We use simulations in years 1996–2010 for training and in years 2011–13 for testing, where each input channel is standardized by using the all-sample and all-pixel mean and standard deviation of the corresponding variable. The end result of this preprocessing is that 2370 samples were used for training, 366 samples were used for testing, and the input image consists of  $192 \times 192 \times 3$  pixels. The trained CNN exhibits a classification accuracy of 62% for the testing data.

This climate-related classification task is fairly similar to the idealized classification task of the previous section in that both tasks require the adopted model to learn to extract (and compare or neglect) specific spatial patterns. In the idealized dataset, the model is required to learn to compare the area of two different classes of spatial patterns (square and circular frames), while in the ClimateNet dataset, the model is required to learn to extract spatial patterns that resemble ARs but to neglect all other spatial patterns that might be present (e.g., tropical cyclones). This similarity between the two problems allows us to validate the XAI insights that are gained from the idealized task in a similar but more climate-related second task.

### c. XAI methods

For our assessment, we consider some of the most popular XAI methods for CNNs that have been proposed in the computer science literature. To keep this section as concise as possible, we only briefly describe how each method explains the

network in the following list (the category that each method belongs to is provided in parenthesis; see also Table 1). For more details on the methods' analytical formulas, the reader is referred to appendix B and the corresponding studies cited below.

#### 1) GRADIENT (SENSITIVITY)

This method (Simonyan et al. 2014) assesses the importance of the input features based on *sensitivity*. Sensitivity refers to how much the value of the output will change for a unit change in a specific feature and is estimated here by the first partial derivative of the network's output with respect to the feature.

#### 2) SMOOTH GRADIENT (SENSITIVITY)

This method (Smilkov et al. 2017) also computes the gradient, but it does so by averaging the gradients over a perturbed number of inputs with added noise. This aims to increase the robustness of the results (i.e., reduce the noise).

#### 3) INPUT\*GRADIENT (ATTRIBUTION)

This method (Shrikumar et al. 2016, 2017) assesses the *attribution* of the output to the input (see detailed differences between *sensitivity* and *attribution* in appendix C). Attribution refers to the marginal contribution of an input feature to the output and is estimated here by multiplying (pixelwise) the input with the gradient.

#### 4) INTEGRATED GRADIENTS (ATTRIBUTION)

This method (Sundararajan et al. 2017) uses a reference vector (e.g., for which the network's output is zero). It then estimates the contribution of each feature as the product of the average of the gradients at points along the straight-line path from the reference point to the input with the distance of that path. Integrated Gradients is similar to Input\*Gradient but is designed to account for nonlinearities in the model that is being explained.

#### 5) LRP (ATTRIBUTION)

This method (Bach et al. 2015) propagates the network's output back to neurons of lower layers, until the input layer is reached. In the backpropagation phase, the relevance/importance of each neuron to the output is estimated, based on different propagation rules. In this study, we consider the most popular Layerwise Relevance Propagation (LRP) rules: (i) the  $LRP_z$  rule (Bach et al. 2015), which distributes the relevance of each neuron based on the values of the localized preactivations that are directed to it, (ii) the  $LRP_{\alpha\beta 0}$  (Bach et al. 2015), which is similar to  $LRP_z$  but considers only positive preactivations, (iii) the  $LRP_{comp}$  (Kohlbrenner et al. 2020), which combines the two previous rules (it applies the  $LRP_z$  rule to distribute the relevance in the fully connected layers of the CNN and the  $LRP_{\alpha\beta 0}$  rule in the convolutional layers), and (iv) the  $LRP_{comp/flat}$  (Bach et al. 2016; Kohlbrenner et al. 2020), which is similar to  $LRP_{comp}$  but additionally applies a flat rule



TABLE 1. Summary of XAI methods considered in this study. Practical strengths (check marks) and weaknesses (crosses) of each method are also reported.

XAI method	Brief summary of the method	Desired property for CNN applications as explored in this study			Extra comments/insights
		Disentangles the sign of relevance	Insensitive to gradient shattering	Not ignorant to zero input	
Gradient (Simonyan et al. 2014)	Calculates the first partial derivative of the model output with respect to the input (sensitivity)	✓	×	✓	Estimates the sensitivity of the output to the input, which is not the same as the attribution; see <a href="#">appendix C</a>
Smooth Gradient (Smilkov et al. 2017)	Calculates the average gradient across many perturbed inputs (sensitivity)	✓	×	✓	
Input*Gradient (Shrikumar et al. 2017)	Multiplies the input with the gradient (attribution)	✓	×	×	
Integrated Gradients (Sundararajan et al. 2017)	Multiplies the average gradient along the straight line between the input point and a reference point with the corresponding distance between the two points (attribution)	✓	×	✓	
LRP <sub><math>\alpha/\beta_0</math></sub> (Bach et al. 2015)	Layerwise backpropagation of each neuron's relevance based on the $\alpha/\beta_0$ rule (attribution)	×	✓	×	Considers only positive preactivations
LRP <sub>z</sub> (Bach et al. 2015)	Layerwise backpropagation of each neuron's relevance based on the z rule (attribution)	✓	×	×	Equivalent to Input*Gradient for networks using ReLU activations
LRP <sub>comp</sub> (Kohlbrenner et al. 2020)	Layerwise backpropagation of each neuron's relevance by combining the $\alpha/\beta_0$ rule and the z rule (attribution)	✓	✓	×	Combines the strengths of LRP <sub>z</sub> and LRP <sub><math>\alpha/\beta_0</math></sub>
LRP <sub>comp/flat</sub> (Kohlbrenner et al. 2020)	Layerwise backpropagation of each neuron's relevance by combining the $\alpha/\beta_0$ rule, the z rule, and the flat rule (attribution)	✓	✓	✓	Provides a coarser picture of attribution; not suitable if local accuracy is necessary
Deep Taylor (Montavon et al. 2017)	Applies Taylor decomposition of the relevance function for each neuron recursively (attribution)	×	✓	×	Equivalent to LRP <sub><math>\alpha/\beta_0</math></sub> for networks using ReLU activations; not defined for negative predictions
PatternNet (Kindermans et al. 2018)	Calculates the signal in the input for each neuron recursively (signal)	×	✓	✓	Estimates the signal (not the same as the attribution)
PatternAttribution (Kindermans et al. 2018)	Calculates the attribution in the direction of the signal for each neuron recursively (attribution)	×	✓	✓	
Deep SHAP (Lundberg and Lee 2017)	Approximates Shapley values for each neuron recursively (attribution)	✓	×	✓	Based on well-founded theory; computationally expensive

in the very lowest layer(s). The flat rule distributes relevance uniformly to all connected neurons without considering the preactivations values.

#### 6) DEEP TAYLOR DECOMPOSITION (ATTRIBUTION)

This method (Montavon et al. 2017) applies a local Taylor decomposition to decompose each neuron’s relevance to the neurons of the lower layer. It is applied recursively until the importance of the input features is obtained. Deep Taylor is equivalent to  $LRP_{\alpha1\beta0}$  for networks that use ReLU activations.

#### 7) PATTERNNET (SIGNAL) AND PATTERNATTRIBUTION (ATTRIBUTION)

These methods (Kindermans et al. 2018) are based on the idea that every input image consists of a signal component (all of the information in the input that is relevant to the prediction task) and a distractor (all of the distracting information that is irrelevant to the prediction task). The method PatternNet performs a layerwise backprojection of the signal to the input space. In each layer, the signal is approximated as a superposition of neuronwise, local signal estimators. This is done recursively, until the signal of the network’s output in the input image is estimated. PatternAttribution aims to estimate the attribution of the network’s output to the input (i.e., not simply the signal), by applying the same layerwise backprojection approach, but also considering the weight vector that connects subsequent layers.

#### 8) DEEP SHAP (ATTRIBUTION)

This method (Lundberg and Lee 2017) approximates the Shapley values (originally discovered in the field of the cooperative game theory; Shapley 1953) for the entire network by computing the Shapley values for smaller components of the network and propagating them backward until the input layer is reached (similar in philosophy to LRP, PatternNet, and PatternAttribution). Shapley values have been shown to satisfy desired properties with regard to the explanation (e.g., local accuracy, missingness, and consistency; Lundberg and Lee 2017), which is not necessarily the case with other XAI methods (LRP, Input\*Gradient, etc.).

### 3. Results

In this section, we present the results of applying the XAI methods first to the synthetic dataset and then to ClimateNet. We highlight that, for the synthetic dataset, methods Gradient, Smooth Gradient, and PatternNet are not directly comparable to the derived ground truth of attribution, since they estimate the *sensitivity* or the *signal* of the output to the input; rather than the *attribution* of the output to the input (see appendix C for the difference between sensitivity and attribution). However, they are included in the intercomparison because of their popularity and for the sake of completeness.

#### a. Synthetic dataset

In Figs. 3 and 4, we explore the CNN strategy for two samples from the synthetic dataset. In both samples, two square frames and one circular frame are present. In Fig. 3, the square frames cover more area (specifically 270 pixels vs 180), and in Fig. 4, the circular frame covers more area (specifically 342 pixels vs 255), while in both cases, the CNN has correctly classified the input images. In terms of the ground truth of attribution in Fig. 3, we expect that the pixels of the square frames increased the certainty of the network (i.e., increased the likelihood of class 2), while the pixels of the circular frame decreased it (this is valid when considering a blank image as our baseline). That is, if it were not for the circular frame, the certainty of the model would have been higher. The opposite is true in Fig. 4.

As is evident in both figures, despite all methods being applied to explain the same exact prediction, different XAI methods lead to different explanations. Specifically, despite most methods identifying the frames as important features, some methods exhibit relative noisier results, and there is no consensus about the sign of the attribution. If this were a classification problem about which we knew nothing (as could be the case for a typical geoscience setting), it would be difficult to reach certain conclusions about the decision strategy of the network. However, by knowing the ground truth of attribution in these examples, we can assess the fidelity of each of the methods and also understand the lack of consensus in the results.

First, Gradient is shown to produce somewhat noisy patterns. For shallow networks, some studies suggest that the gradient resembles a Brownian motion and exhibits spatial coherence, while for deeper networks the gradient converges to white noise and the spatial autocorrelation vanishes (Balduzzi et al. 2017). This phenomenon is known in the computer science literature as “gradient shattering” (Balduzzi et al. 2017). Although our network is not very deep (less than 10 layers), the noise in the results of the gradient can be partially attributed to gradient shattering. Despite this, one can see that the square or circular frames are respectively highlighted with mostly positive or negative values in Fig. 3, whereas the opposite results are shown in Fig. 4, which is consistent with what we expect in both cases. Moreover, in both figures, the gradient vanishes away from the frames. This means that the CNN has correctly learned that if one were to increase the value of any pixel away from the frames this would not affect the chances of either class, because isolated pixels constitute neither a circular nor a square frame. Smooth Gradient produces very different results, namely, mostly negative gradients in Fig. 3 and mostly positive gradients in Fig. 4.

Results from the Input\*Gradient and Integrated Gradients methods are very similar and close to the ground truth of attribution (pattern correlation with the ground truth is on the order of 0.5–0.6 in all examples). In Fig. 3 the square or circular frames are highlighted with mostly positive or negative attributions, respectively, whereas in Fig. 4 we obtain the opposite results. Pixels outside the frames receive zero attribution. However, both methods may suffer from the effects of

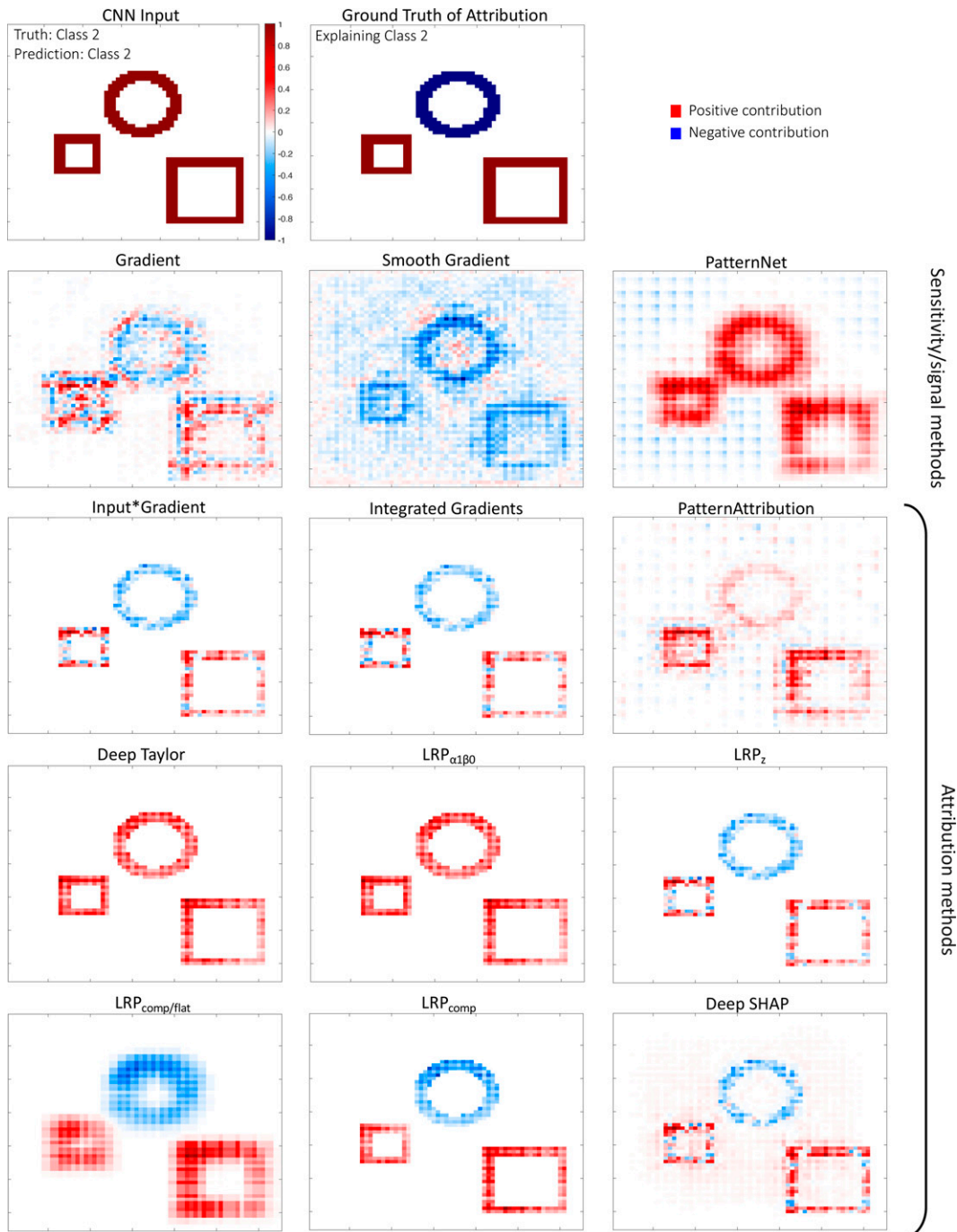


FIG. 3. Explanations from different XAI methods of the strategy of the CNN for the synthetic dataset and sample 453567. The CNN has successfully classified this image to class 2; i.e., the square frames cover more area (area: 270) than the circular frames do (area: 180). XAI methods are applied to explain the successful prediction. For each heat map, we divided all values by the maximum (in absolute terms) value. The ground truth of attribution is derived using a blank image (image with zeros) as a baseline.

gradient shattering in the same way as Gradient, since they are directly connected to the latter [see Eqs. (B3) and (B4) in appendix B]. Indeed, as we can see in Figs. 3 and 4, attributions exhibit some level of noise, which for a deeper network

might be so high that it can severely limit comprehensibility (e.g., see Fig. 7, below).

The PatternNet method correctly highlights all three frames (as well as some pixels away from the frames) as important in



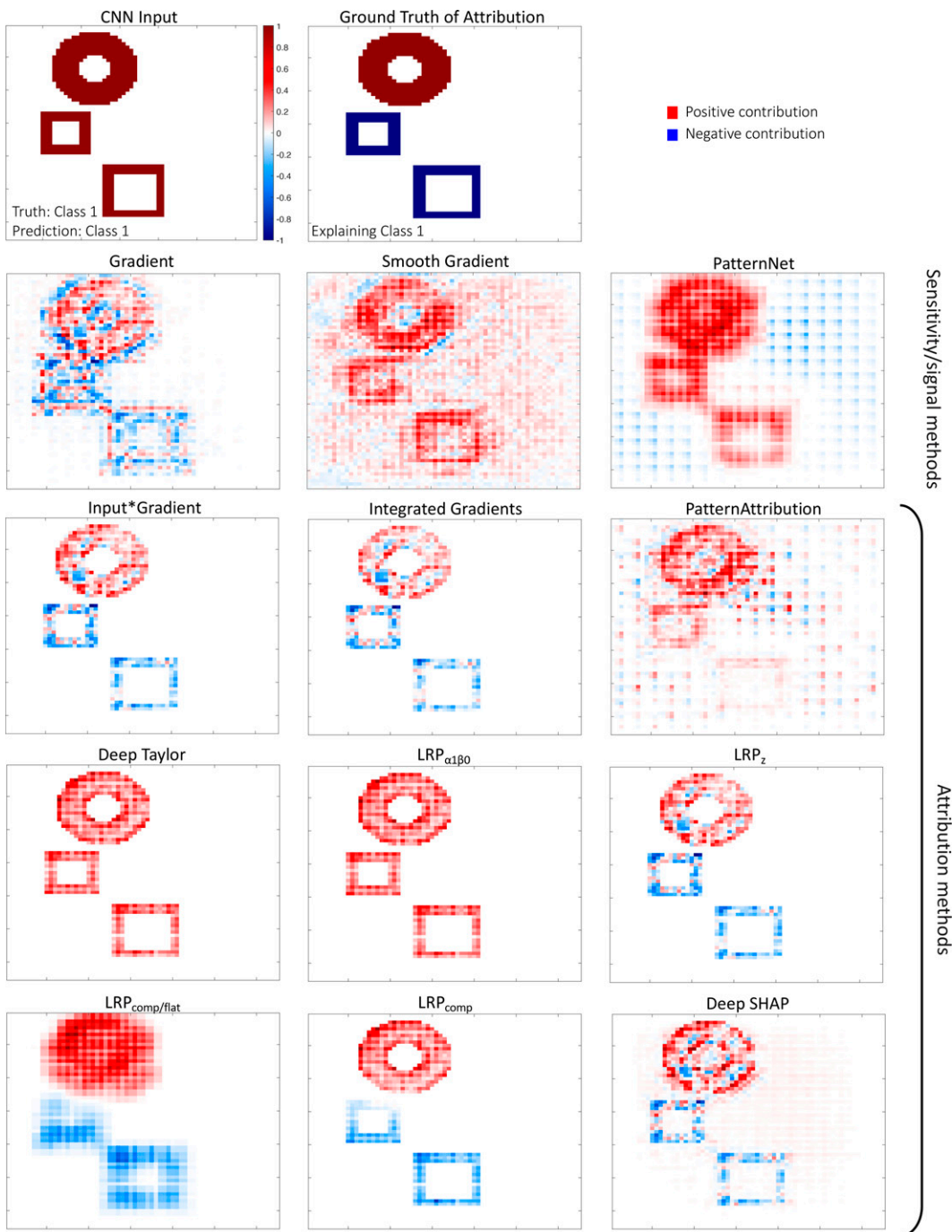


FIG. 4. As in Fig. 3, but for the sample 450345. The CNN has successfully classified this image to class 1; i.e., the circular frames cover more area (area: 342) than the square frames do (area: 255). XAI methods are applied to explain the successful prediction.

containing information (a signal) for the decision of the CNN. PatternAttribution correctly highlights the two square frames in Fig. 3 and the circular frame in Fig. 4 as contributing positively to the CNN’s decision. However, PatternAttribution does not very effectively distinguish between positive and negative contributions in either example, because in both

cases, it assigns positive attribution to the frames that are actually contributing negatively to the CNN’s decision. Its pattern correlation with the ground truth is on the order of 0.4 in both examples.

The results of Deep Taylor and the  $LRP_{\alpha1\beta0}$  rule are identical, since these two methods are equivalent for networks with

ReLU activations (Samek et al. 2016; Montavon et al. 2017). Both methods pick up the corresponding three frames, and pixels outside the frames receive zero attribution. However, all frames receive positive attributions in both figures, which is not consistent with the ground truth of attribution (the results of these two methods exhibit a correlation with the ground truth of only about 0.1–0.2). It has recently been noted that  $\text{LRP}_{\alpha 1 \beta 0}$  propagates the sign of the before-softmax value back to the input (e.g., Kohlbrenner et al. 2020), and, thus, it is not able to distinguish between positive and negative contributions of different features.<sup>2</sup> Because of this property,  $\text{LRP}_{\alpha 1 \beta 0}$  is known to provide smoother (not very noisy) results than do other LRP rules—however, with limited local accuracy, since the negative preactivations are not being considered in this rule [see Eq. (B6) in appendix B]. Results from  $\text{LRP}_z$  are the same as those from the Input\*Gradient since the two methods are equivalent when explaining networks that use ReLU activations (Ancona et al. 2018; 2019).

The results from  $\text{LRP}_{\text{comp}}$  seem to be the most consistent and very similar to the ground truth of attribution (this method exhibits the highest correlation with the ground truth, on the order of 0.8–0.9). As mentioned in the previous section, this method combines  $\text{LRP}_{\alpha 1 \beta 0}$  and  $\text{LRP}_z$  in an attempt to get the best from both rules: as shown in Figs. 3 and 4, it is able to maintain local accuracy, and thus, distinguish between positive and negative contributions (owing to the use of  $\text{LRP}_z$ ), while at the same time returning smooth results, thus, eliminating the effect of gradient shattering (owing to the use of  $\text{LRP}_{\alpha 1 \beta 0}$ ). The rule  $\text{LRP}_{\text{comp}/\text{flat}}$  is shown to provide a coarser but similar picture of attribution to the  $\text{LRP}_{\text{comp}}$  (correlation with the ground truth on the order of 0.7). This verifies arguments in previous studies (Bach et al. 2016) that if the analyst/scientist is not interested in local accuracy, but they only need to obtain a coarse picture of the attribution, this is a suitable rule to use. Last, the method Deep SHAP is shown to provide attributions that are close to the ground truth, but relatively noisier (correlation with the ground truth on the order of 0.5–0.6). Results are similar to the results of Input\*Gradient, Integrated Gradients, and  $\text{LRP}_z$ .

In Fig. 5, we repeat the results of Fig. 4, but now we aim to detect which features in the input made the CNN assign a very small probability to class 2. We note that in geoscientific applications, it is always good practice to use XAI to explain not only the predicted class but also the rejected class(es), since this may provide further insight. The ground truth of

<sup>2</sup>This can be explained easily by looking at the formula in Eq. (B6) and setting  $\alpha = 1$  and  $\beta = 0$  to obtain the  $\text{LRP}_{\alpha 1 \beta 0}$  rule: because the ratio  $z_{ij}^+/z_{ij}^-$  is by definition a positive number, then the relevance of any neuron in the lower layer  $R_i^{(l)}$  has the same sign as the relevance of the neuron in the upper layer  $R_j^{(l+1)}$ , and this sign is maintained and recursively propagated back to the input layer. Thus, when the before-softmax value of the class that is being explained is a positive number then the corresponding heat map will show only nonnegative values and when it is a negative number then the corresponding heat map will show only nonpositive values.

attribution in Fig. 5 shows the opposite of what shown in Fig. 4: the pixels of the square frames increase the likelihood of class 2, while the pixels of the circular frame decrease it. The XAI results verify most of the arguments made in the discussion of the previous figures. First, methods like Gradient, Input\*Gradient, Integrated Gradients,  $\text{LRP}_z$ , and Deep SHAP are able to disentangle the sign of the attribution but might be partially affected by gradient shattering. The  $\text{LRP}_{\alpha 1 \beta 0}$  rule provides smooth results but cannot disentangle the sign of the attribution. Specifically, it assigns negative attributions to all frames (similarly to PatternNet and PatternAttribution), because the before-softmax value that corresponds to class 2 is a negative number in this example. Deep Taylor does not return any results, since this method is only defined for positive network outputs (Montavon et al. 2017). Last, the method  $\text{LRP}_{\text{comp}}$  is again shown here to provide the most consistent attribution relative to the ground truth, because it is able to provide smooth results and also disentangle the sign of the attribution.

To explore the above insights more quantitatively and across many samples we have calculated the distribution of the correlation with the ground truth for each of the XAI attribution methods (not shown). This analysis showed that  $\text{LRP}_{\text{comp}}$  exhibits systematically the strongest correlation with the ground truth,  $\text{LRP}_{\alpha 1 \beta 0}$  exhibits the weakest correlation, while the rest of the methods fall in between, similar to what Figs. 3–5 suggest.

Next, we explore the sensitivity of the XAI results to input transformations. In geoscientific applications, input transformations may represent modifications of the units of an input variable (e.g., from kelvins to degrees Celsius) or the scaling (anomalies about zero vs raw measurements), thus, it is of high importance to investigate their effect. To do so, we perform the following experiment that is inspired by Kindermans et al. (2019): we apply a uniform shift of  $s = -1$  to all pixels in all input images of the synthetic dataset. The features of the shifted input are binary variables with  $\mathbf{X}^* \in \{-1, 0\}^d$ : the pixels of the frames are equal to 0, and the nonframe pixels are equal to  $-1$ . We then consider the already trained CNN and simply change the biases of its first layer to account for the shift in the input: for any  $j$  neuron in the first hidden layer the new bias term is modified as  $b_j^* = b_j - \sum_i w_{ij} s = b_j + \sum_i w_{ij}$ . With this modification, the predictions of the modified CNN (denoted CNN\*) when using the shifted input (denoted  $\mathbf{X}^*$ ) are the same as those of the CNN in the original setting.<sup>3</sup>

In Fig. 6, we apply XAI methods to explain the decision strategy of this modified CNN\* for the same prediction as in Fig. 4. The methods Gradient, Smooth Gradient, PatternAttribution,  $\text{LRP}_{\text{comp}/\text{flat}}$ , and Deep SHAP provide similar results, which makes them “input shift invariant” (Kindermans et al. 2019).

<sup>3</sup> Any activation value of the neurons in the first hidden layer  $x_j^*$  is equal to the corresponding activation in the original setting:  $x_j^* = \text{ReLU}(\sum_i w_{ij} x_i^* + b_j^*) = \text{ReLU}[\sum_i w_{ij}(x_i - 1) + b_j + \sum_i w_{ij}] = \text{ReLU}[\sum_i (w_{ij} x_i - w_{ij}) + b_j + \sum_i w_{ij}] = \text{ReLU}(\sum_i w_{ij} x_i + b_j) = x_j$ ; see also Kindermans et al. (2019).

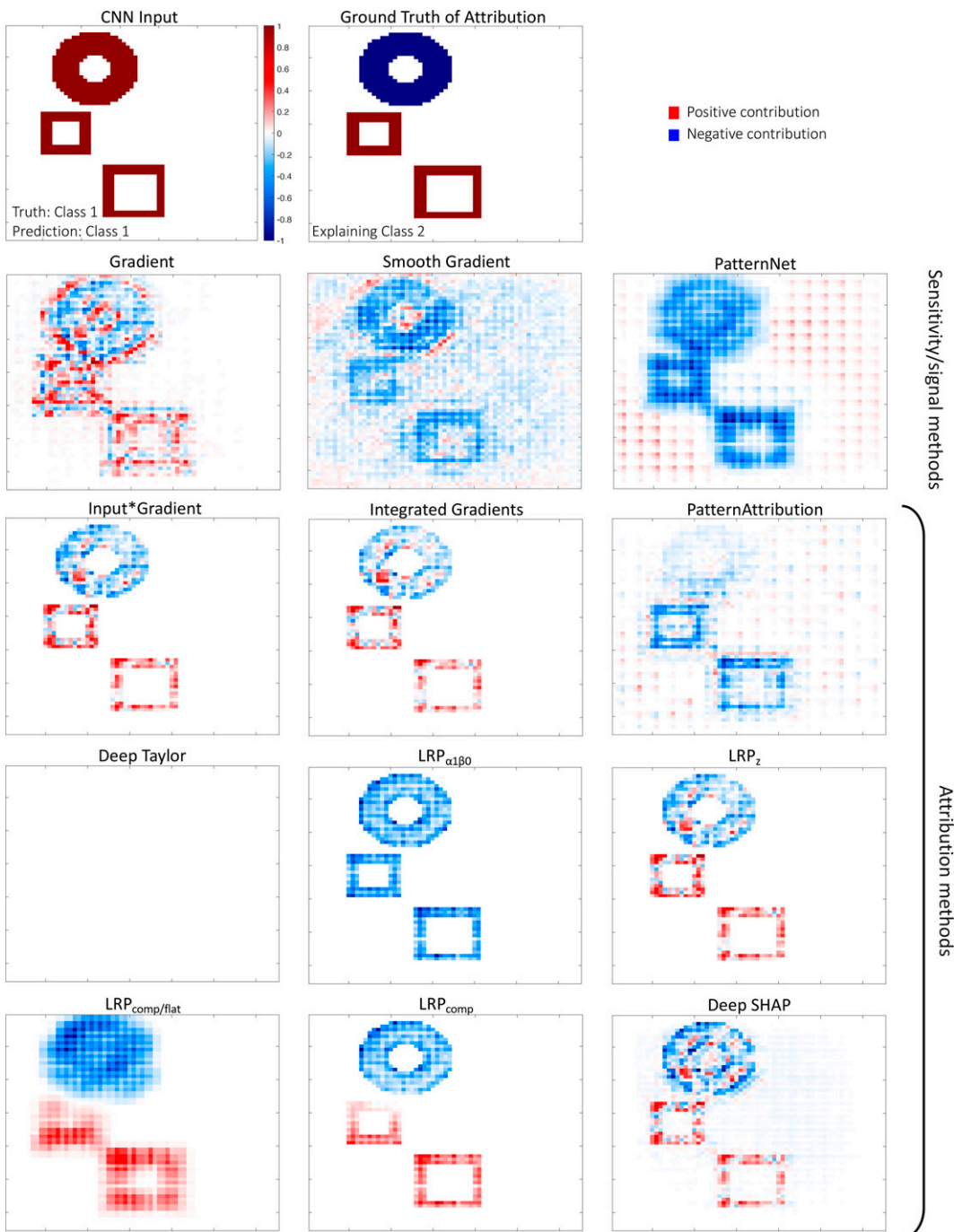


FIG. 5. As in Fig. 4, but here XAI methods are applied to explain why the CNN correctly predicted that class 2 is not true (i.e., explaining the low probability that the CNN assigned to class 2).

The reason for the invariance in their results is as follows: (i) The gradient of a constant is zero, so methods Gradient and Smooth Gradient are expected to be “input shift invariant.” (ii) The method  $LRP_{comp/flat}$  applies a flat rule in the lowest layers, which distributes relevance uniformly to any input feature that is connected to a neuron in the upper layer, without considering the value of preactivations [see

Eq. (B7)]. Thus, since the architecture of the modified CNN\* and all preactivations in all layers except the lowest one are the same as in the original setting, the feature attributions are the same. (iii) Both PatternAttribution and Deep SHAP use the range of variability of the input features in the training dataset to assess feature importance; thus, the input shift is taken into account.

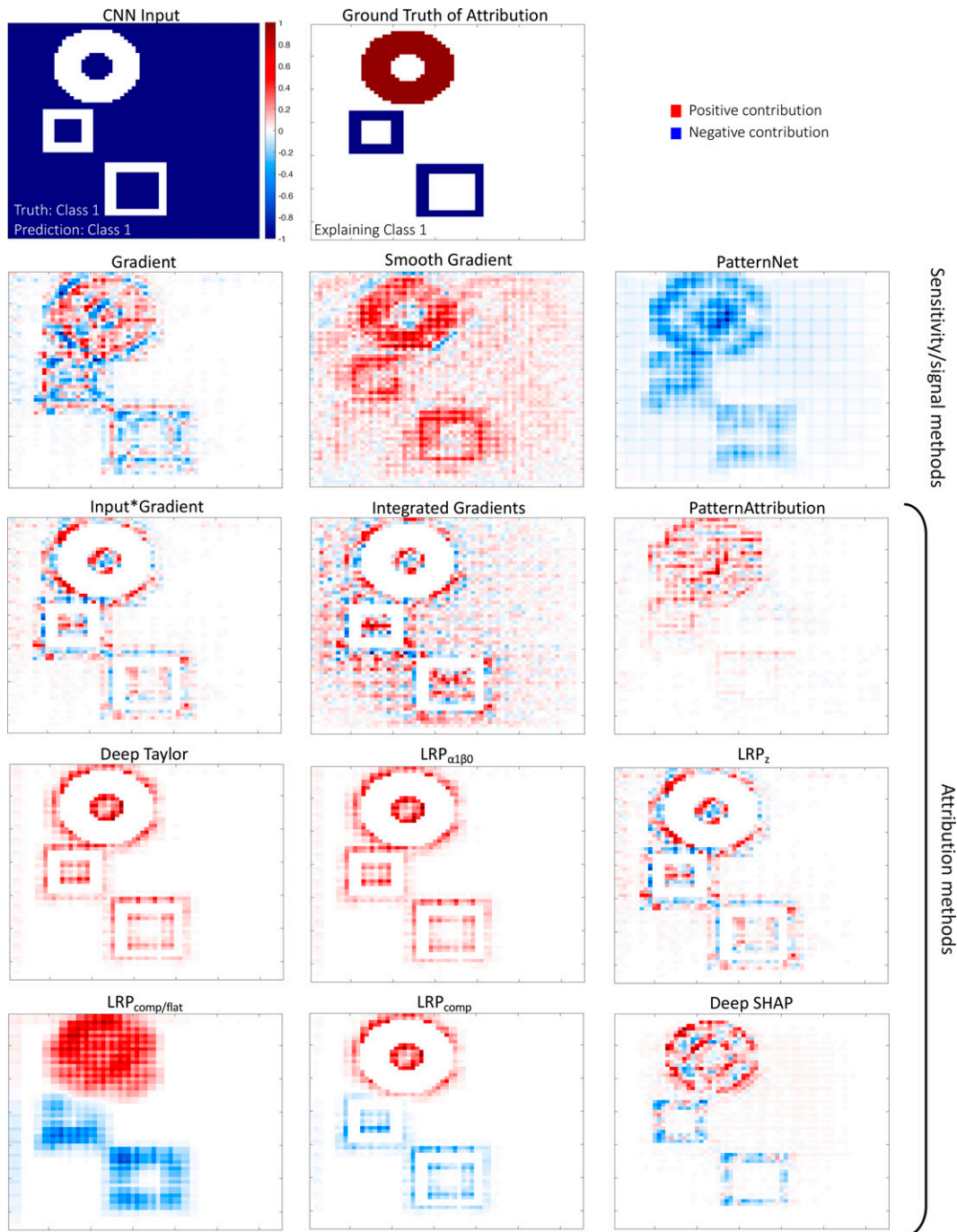


FIG. 6. As in Fig. 4, but after a shift of  $-1$  has been applied to the input. Also note that the ground truth of attribution is derived using a baseline image with all feature values equal to  $-1$ .

In contrast, Input\*Gradient, Integrated Gradients, and all the rest of the LRP rules show very different results with the shifted input when compared with Fig. 4, with most of these methods highlighting the perimeter of the three frames while the body of the frames receives zero attribution. This indicates a sensitivity of these methods to input transformations. In our example, this sensitivity originates from the fact that these methods

are theoretically unable to assign attribution to a zero value in the input (i.e., the body of the frames in Fig. 6). Indeed, the formulas in Eqs. (B3) and (B4) in appendix B show that Input\*Gradient and Integrated Gradients (when using a blank image as reference) always assign a zero attribution to a zero input by construction. Similarly, all LRP rules except  $LRP_{comp/flat}$  perform the relevance redistribution based on the preactivation



value  $w_{ij}x_i$ , thus zero inputs automatically receive a zero attribution. For the rest of the paper, we will refer to this systematic behavior of assigning zero attribution to a zero input and ignoring the impact it could have to the network's output as the "ignorant to zero input" issue. Input\*Gradient and all LRP rules except  $\text{LRP}_{\text{comp/flat}}$  are ignorant to zero input. The ignorant-to-zero-input issue did not show up in Figs. 3 and 4 (if anything, it worked to the advantage of these methods), since pixels with a zero value were expected to receive zero attribution in those examples. In general, however, this issue can provide a distorted picture of the decision strategy of the CNN. A clear example is Fig. 6, where, according to Input\*Gradient and most LRP rules, the frames are not important to the CNN decision.<sup>4</sup> We also note that, if we train a completely new CNN to classify the shifted images and use XAI to explain its predictions (see Supplementary Fig. 2 in the online supplemental material), we observe very similar results with Fig. 6, which further verifies the validity of the above remarks.

The results of this section highlight three important issues of XAI methods, namely, the effect of gradient shattering, the issue of disentangling the sign of the attribution, and the ignorant-to-zero-input issue (see Table 1). All of these issues may limit the user's understanding of the decision-making strategy of a CNN, and no method was shown to be optimal.

#### b. Climate net

In Figs. 7 and 8, we apply the same XAI methods to explain CNN predictions for the ClimateNet dataset. For this dataset, there is no clearly defined ground truth for the attribution of the output to the input. Even though the dataset contains labeled maps by experts (i.e., Supplementary Fig. 1d in the online supplemental material), these cannot act as a ground truth for the attribution, as the CNN may employ patterns or climate information outside the regions of the ARs for making its predictions. Thus, we cannot assess the XAI fidelity for this application as we did for the synthetic dataset. Instead, we use this dataset to examine whether, and how, the properties and artifacts of different XAI methods that were identified in the previous section manifest in a more climate-related prediction setting. Furthermore, by providing this example we seek to illustrate how the knowledge of relative strengths and weaknesses of each XAI method affects our interpretation of the corresponding XAI results.

In the specific sample that we consider, two ARs have been detected by the expert scientists, and the CNN correctly assigned this input to the class of two or more ARs. In Fig. 7, we present the XAI results that explain which features in

the first channel of the input image (the zonal wind at 850 hPa pressure level U850) the CNN used to make this prediction. Similar to the previous dataset, the obtained results are very different when using different XAI methods, which makes the interpretation of the decision-making strategy of the CNN challenging. First, in accordance with the remarks of the previous section, one can see that the results of the methods Gradient, Smooth Gradient, Input\*Gradient, Integrated Gradients, and  $\text{LRP}_z$  are very noisy (Fig. 7), and based on these methods, one cannot make any robust inferences about the CNN's strategy. For ClimateNet, the CNN that we use is almost 2 times as deep as in the previous dataset (see Fig. 2), and, thus, the gradient shattering has a detrimental effect on the explanations.

Focusing on the rest of the methods, PatternNet highlights all features in the input where zonal wind is positive, indicating that these features contain important information for the network. PatternAttribution seems to primarily highlight one of the two wind patterns that are associated with the two ARs. The methods Deep Taylor and  $\text{LRP}_{\alpha, \beta 0}$  provide only positive attributions to all highlighted features (recall here from the previous section that these methods do not disentangle the sign of the attribution), and they assign the highest attribution to the two positive wind patterns that are associated with the ARs. The same features are highlighted more clearly when using the methods  $\text{LRP}_{\text{comp}}$  and Deep SHAP. These two methods are relatively more insightful, since (i)  $\text{LRP}_{\text{comp}}$  is a "best practice" implementation of LRP (Kohlbrenner et al. 2020) and it combines the strengths of the rules  $\text{LRP}_{\alpha, \beta 0}$  and  $\text{LRP}_z$ , and (ii) Deep SHAP has been proven to satisfy desirable properties of consistency, local accuracy, and missingness (Lundberg and Lee 2017), and it successfully disentangles the sign of the attribution and does not exhibit the ignorant-to-zero-input issue. Last, the results from  $\text{LRP}_{\text{comp/flat}}$  show how the attribution is distributed when one considers all three channels together; recall here that this rule applies a flat (uniform) rule of relevance distribution in the lowest layers; thus, the obtained heat map is determined by how relevance is distributed spatially across the neurons in the upper layers. The results show that the important features for this prediction form two spatial patterns that are closely aligned with the locations over which the two ARs were detected by the experts (we do not wish to further assess this alignment quantitatively since there is no exact grid-by-grid correspondence between the U850 and the labeled fields). Thus, we can conclude that the network classified this input to the right class based on the wind features that are associated with the labeled ARs locations, which may add to the model's trustworthiness. The XAI results for the other two channels of V850 (meridional wind) and integrated precipitable water are presented in Supplementary Figs. 3 and 4 in the online supplemental material.

In Fig. 8, we consider the same input and use XAI to explain why the CNN assigned a small probability to the class of zero ARs. We again observe that the effect of gradient shattering is drastic and makes the results of Gradient, Smooth Gradient, Input\*Gradient, Integrated Gradients, and  $\text{LRP}_z$

<sup>4</sup> As a second example, let us consider that we wanted to explain the prediction of a (supposedly perfectly trained) network that simulates the function  $F(\mathbf{X}) = \sum_{i=1}^d \cos(X_i)$  at the point  $\mathbf{x} = \mathbf{0}$ . An ignorant-to-zero-input method would assign a zero attribution to all input features, just because  $x_i = 0, \forall i$ . This ignores the fact that each feature is actually contributing  $\cos(0) = 1$  to the total sum and leads to a distorted picture of the network's predictive strategy.



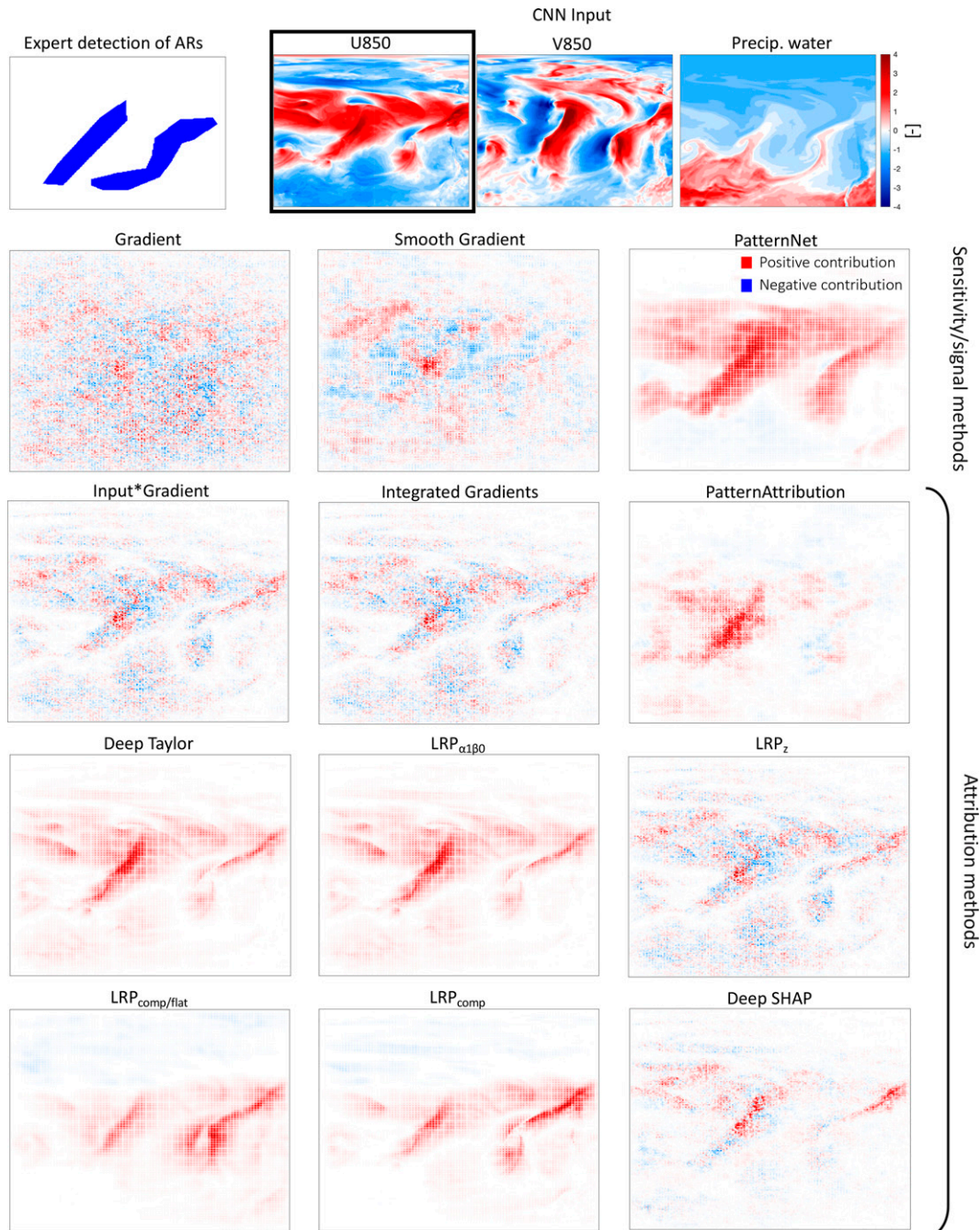


FIG. 7. Explanations from different XAI methods of the strategy of the CNN for the ClimateNet dataset. The CNN has successfully classified the input image to the class of two or more ARs. XAI methods are applied to explain the successful prediction, and results correspond to the U850 channel. For each heat map, we divided all values by the maximum (in absolute terms) value.

incomprehensible. On the basis of the rest of the methods, and by comparing Figs. 7 and 8, results show roughly the same patterns but with the opposite sign. This suggests that the features that made the network be certain about the occurrence of two or more ARs are also the features that made the network decide that the considered input is not likely a

simulation with zero ARs. Thus, in Fig. 8, we verify that the CNN based its decision on features that are associated with the two ARs.

The above results validate the conclusions of our analysis in the previous section and show that the effects observed for the different XAI methods for the synthetic benchmark

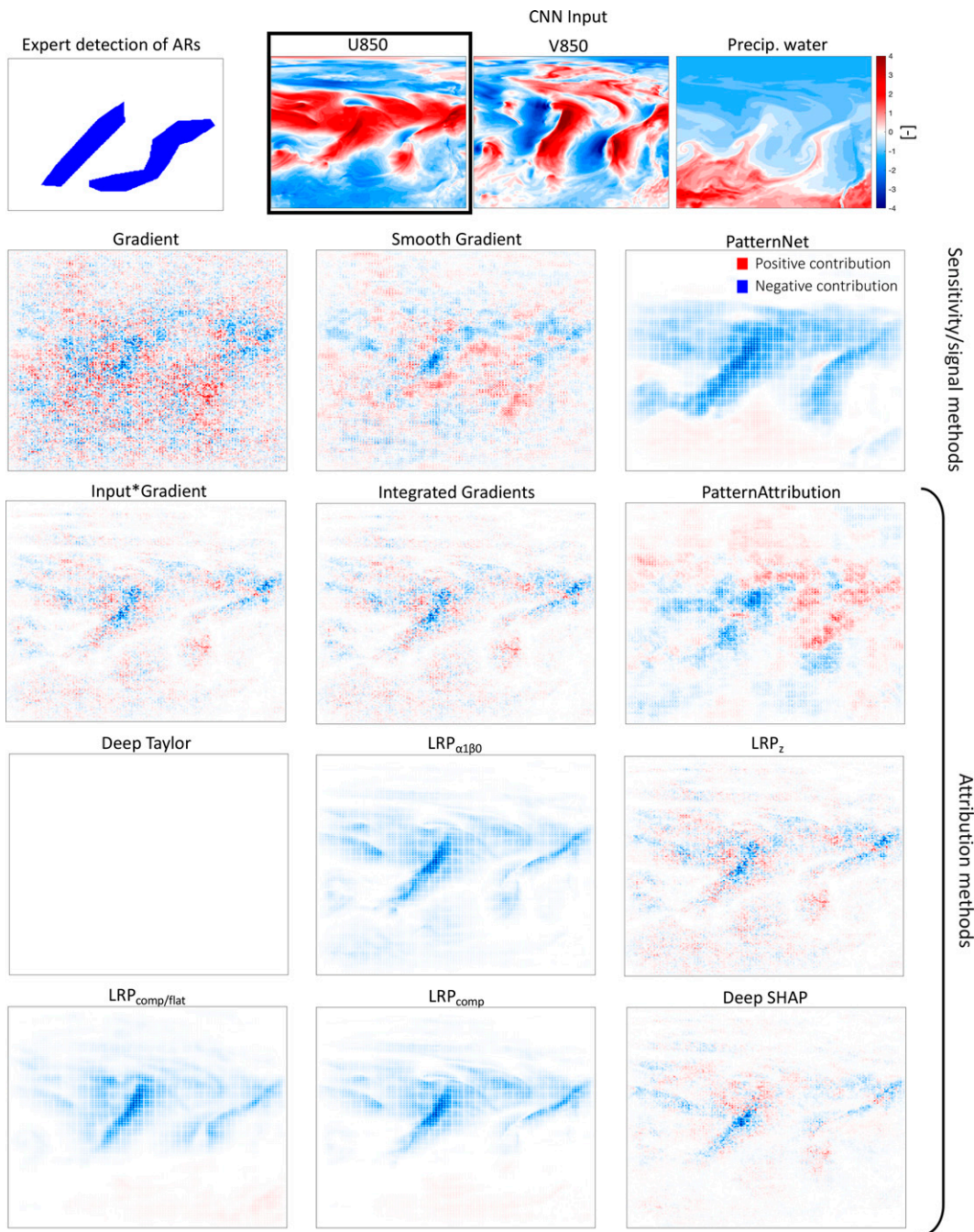


FIG. 8. As in Fig. 7, but here XAI methods are applied to explain why the CNN assigned a low probability to the class of zero ARs.

occur also for climate data and thus need to be taken into account when interpreting the results. In particular, no optimal method exists. Thus, in typical prediction applications, in which no ground truth of attribution exists, a more holistic approach should be taken. By considering the explanations from many XAI methods as a whole (as in Figs. 7 and 8) and knowing the relative strengths and weaknesses of each one, scientists may more effectively gain insights about the decision-

making strategy of the network, as opposed to the use of a single method.

#### 4. Conclusions

Explainable artificial intelligence has increasingly been receiving attention in the field of geoscience, as a means to explain black-box models of machine and deep learning that are



not inherently interpretable. Although the potential of XAI methods has already been documented in the computer science literature and in geosciences (McGovern et al. 2019; Ebert-Uphoff and Hilburn 2020; Barnes et al. 2020; Toms et al. 2020, 2021; Sonnewald and Lguensat 2021; Mayer and Barnes 2021; Hilburn et al. 2021; Keys et al. 2021; Mamalakis et al. 2022b), many studies have highlighted theoretical and practical limitations (Ancona et al. 2018; Kindermans et al. 2019; Rudin 2019; Dombrowski et al. 2022; Zhou et al. 2022). Moreover, the assessment of XAI has typically been based on subjective criteria in the recent literature (Mamalakis et al. 2022a; Leavitt and Morcos 2020). To shed more light on the XAI limitations and to gain insight into best practices, in this study we considered some of the most popular XAI methods and compared the fidelity of their explanations in applications for convolutional neural networks relevant to geoscience. To do so, we used a synthetic attribution benchmark, where the ground truth of attribution is a priori known, to objectively highlight relative strengths and weaknesses, and a dataset of climate simulations to validate our insights in a more typical prediction setting.

Our investigation revealed aspects that need to be considered when applying XAI methods. These include (i) gradient shattering (i.e., the phenomenon of noisy patterns in the gradient), the level of which is a function of the depth of the network. For very deep networks, gradient shattering might lead to overwhelmingly noisy patterns that make the explanation of any gradient-based method incomprehensible. (ii) Many of the considered methods are either theoretically unable or were shown in practice to be ineffective in disentangling positive and negative contributions. This may lead to a very distorted picture of what the network’s strategy is and possibly limit trust in the predictive model itself. (iii) Some methods automatically assign a zero attribution to zero values in the input, despite the fact that in specific settings a zero input value could be important for the prediction. We referred to this issue as the ignorant-to-zero-input issue. The results of these methods may be more informative if they are viewed as explanations that correspond to a blank image baseline (i.e., an image with only zeros). The effect and/or usefulness of assuming different baselines in XAI research will be the subject of a future study. A summary of the relative strengths and weaknesses that the considered methods exhibit for the types of applications in the current analysis is shown in Table 1.

Our investigation suggests that no optimal method exists for all prediction settings and network architectures. For example, previous studies in computer science and the geosciences have shown that for relatively shallow fully connected networks and for physical problems where a zero input contains no information, methods like Input\*Gradient, Integrated Gradients, and LRP<sub>z</sub> might perform well (Kohlbrenner et al. 2020; Mamalakis et al. 2022a). Yet in this investigation, we showed that for deep CNNs and/or for cases for which a zero input might be important for the prediction, these methods might provide a distorted picture of the decision strategy of the network. Having clarified that no universally optimal method exists, we note that for CNN applications, one might have relatively more good reasons to use methods like LRP<sub>comp</sub>,

LRP<sub>comp/flat</sub> and Deep SHAP than others. Yet, these methods are not perfect and require different computational resources, so we would argue that applying many methods and collectively comprehending the CNN strategy (a more holistic approach) is and will be the way to go for the foreseeable future. We conclude by saying that we envision our analysis and revealed insights to highlight even more the need for rigorous and objective assessment of XAI methods so as to successfully implement them in geoscience and leverage machine and deep learning for prediction.

*Acknowledgments.* This work was supported in part by the National Science Foundation under Grant OAC-1934668. Author Ebert-Uphoff also acknowledges support by the National Science Foundation under Grant ICER-2019758. The authors also thank the ClimateNet team (see Prabhat et al. 2021) for their efforts in making their data publicly available.

*Data availability statement.* The code that was used to generate the synthetic data and train and explain the CNN has been made publicly available ([https://github.com/amamalak/XAI\\_Fidelity\\_Assessment\\_CNN\\_GEO](https://github.com/amamalak/XAI_Fidelity_Assessment_CNN_GEO)). The ClimateNet dataset (Prabhat et al. 2021) is also publicly available (<https://portal.nersc.gov/project/ClimateNet/>).

## APPENDIX A

### The Use of Additively Separable Functions for Generating Synthetic Attribution Benchmarks and Their Connection to Our Study

As mentioned in section 2a of the main text, an attribution benchmark consists of a synthetic input  $\mathbf{X}$  and a synthetic output  $Y$ , with the latter being a known function  $F$  of the former (Mamalakis et al. 2022a). The functional form of  $F$  depends on what type of network one wants to benchmark (e.g., a fully connected network, a CNN), and Mamalakis et al. (2022a) noted that the function  $F$  can be of any arbitrary choice, as long as it has such a form so that the attribution of any output to the corresponding input is objectively derivable.

Mamalakis et al. (2022a) suggested that a simple form for  $F$  so that the above property is honored is when  $F$  is an *additively separable function*, that is, there exist local functions  $C_i$ , with  $i = 1, 2, \dots, d$ , so that

$$\begin{aligned} Y = F(\mathbf{X}) &= F(X_1, X_2, \dots, X_d) \\ &= C_1(X_1) + C_2(X_2) + \dots + C_d(X_d), \end{aligned} \quad (\text{A1})$$

where the form of  $C_i$  is chosen by the analyst depending on what type of network they want to benchmark. The important thing to notice is that because of the summation in Eq. (A1), and for any form of  $C_i$ , the contribution of any input feature  $X_i$  to the output  $y_n$  in the sample  $n$  is by definition equal to  $C_i(x_{i,n})$ ; that is when considering a zero baseline. This allows for deriving a ground truth of the

attribution for any sample  $n$ , and for any input feature  $X_i$ , and thus, a synthetic benchmark with  $F$  being an additively separable function allows for objectively benchmarking XAI methods.

As described in [section 2a](#), in this study, we generated a series of images where circular and square frames are present, and the task was to classify each image depending on which class of frames covers more area. This classification task can be shown to fall under the umbrella of additively separable functions as in [Eq. \(A1\)](#). Specifically, to generate the synthetic output of the current dataset, we may follow the framework of [Mamalakis et al. \(2022a\)](#), and define the output variable  $Y \in \mathbb{Z}^*$  as in [Eq. \(A1\)](#), but where

$$C_i(x_{i,n}) = \begin{cases} 1, & \text{if } i \text{ belongs to a square frame} \\ -1, & \text{if } i \text{ belongs to a circular frame.} \\ 0, & \text{otherwise} \end{cases} \quad (\text{A2})$$

By combining [Eqs. \(A1\)](#) and [\(A2\)](#), one can quickly notice that  $Y$  essentially represents the difference of the total area of square frames minus the total area of the circular frames in each image. If  $Y > 0$ , then the square frames cover more area in the corresponding image, and if  $Y < 0$ , the circular frames cover more area (note that during the simulation of the synthetic dataset, samples that happen to exhibit  $Y = 0$  may be disregarded). Thus, the classification task is simplified to predicting the sign of the output  $Y$ . A negative sign of  $Y$  corresponds to class 1, and a positive sign of  $Y$  corresponds to class 2, as these are defined in [section 2a](#).

The ground truth of the attribution is easily and objectively derivable, following [Eq. \(A2\)](#). In simple terms, and in accordance with the discussion in [section 2a](#), for a sample  $n$ , pixels that belong to any square frames contribute positively to the value of  $y_n$  (i.e., these pixels “push”  $y_n$  to have a positive sign), whereas pixels that belong to any circular frames contribute negatively (i.e., these pixels “push”  $y_n$  to have a negative sign). We highlight the latter rule of attribution is valid when considering a blank image as the baseline. Moreover, the contribution of each pixel to the output  $Y$  depends on whether the pixel belongs to a circular or square frame [see [Eq. \(A2\)](#)], thus, it depends on the values of the neighboring pixels. This inherent spatial dependency makes a CNN be the most suitable type of network to address this classification task.

## APPENDIX B

### Analytical Formulas of the Considered XAI Methods

In the Gradient method ([Simonyan et al. 2014](#)), one calculates the partial derivative of the network’s output with respect to each of the input features  $X_i$ , for the specific sample in question. The relevance (or importance) of the feature at grid point  $i$  for the network’s prediction of sample  $n$  is

$$R_{i,n} = \frac{\partial \hat{F}}{\partial X_i} \Big|_{X_i=x_{i,n}}, \quad (\text{B1})$$

where  $\hat{F}$  is the function learned by the CNN, as an approximation to the true function  $F$ . This method estimates the *sensitivity* of the network’s output to the input variable  $X_i$ . The motivation for using the Gradient method is that if changing the value  $x_{i,n}$  of a grid point is shown to cause a large change to the CNN output, then that grid point may be relevant for the prediction. Furthermore, calculation of the Gradient is very convenient, as it is readily available in any network training environment, contributing to the method’s popularity.

The Smooth Gradient sensitivity method was introduced in ([Smilkov et al. 2017](#)) and is very similar to the method Gradient except that it aims to obtain a more robust estimation of the local derivative by averaging the gradients over a perturbed number of inputs with added noise:

$$R_{i,n} = \frac{1}{m} \sum_{j=1}^m \frac{\partial \hat{F}}{\partial X_i} \Big|_{X_i=x_{i,n}+e_{i,n,j}}, \quad (\text{B2})$$

where  $m$  is the number of perturbations and  $e_{i,n,j}$  comes from a standard normal distribution.

As is evident from its name, the Input\*Gradient method ([Shrikumar et al. 2016; 2017](#)) multiplies the local gradient with the input itself to get the relevance:

$$R_{i,n} = x_{i,n} \frac{\partial \hat{F}}{\partial X_i} \Big|_{X_i=x_{i,n}}. \quad (\text{B3})$$

This method quantifies the *attribution* of the output to the input. Attribution methods aim to quantify the relative contribution of each input feature to the output value, something that is conceptually different from the sensitivity of the output to the input, as in the previous two methods; for a brief explanation of the difference between attribution and sensitivity see [appendix C](#).

The Integrated Gradients method ([Sundararajan et al. 2017](#)) is also an attribution method similar to the Input\*Gradient method but aims to account for the fact that in nonlinear problems the derivative is not constant. This method considers a reference (baseline) vector  $\hat{\mathbf{x}}$  [for which the network’s output is zero; that is,  $\hat{F}(\hat{\mathbf{x}}) = 0$ ]. Then the relevance is equal to the product of the distance of the input from the reference point with the average of the gradients at points along the straight-line path from the reference point to the input

$$R_{i,n} = (x_{i,n} - \hat{x}_i) \frac{1}{m} \sum_{j=1}^m \frac{\partial \hat{F}}{\partial X_i} \Big|_{X_i=\hat{x}_i + \frac{j}{m}(x_{i,n} - \hat{x}_i)}, \quad (\text{B4})$$

where  $m$  is the number of steps in the Riemann approximation.

LRP ([Bach et al. 2015; Samek et al. 2016](#)) is an attribution method that sequentially propagates the prediction  $\hat{F}(\mathbf{x}_n)$  (more specifically the before-softmax value) back to neurons of lower layers, obtaining the intermediate relevance for all neurons, until the input layer is reached and the relevance of all input features  $R_{i,n}$  is calculated. There are many different rules with which this relevance propagation

can be performed. Below we consider the most popular rules for CNNs.

#### a. $LRP_z$

In the  $LRP_z$  rule, the backpropagation is performed as follows:

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{z_j} R_j^{(l+1)}, \quad (\text{B5})$$

where  $R_j^{(l+1)}$  is the relevance of the neuron  $j$  at the upper layer ( $l + 1$ ) and  $R_i^{(l)}$  is the relevance of the neuron  $i$  at the lower layer ( $l$ ). The propagation is based on the ratio of the localized preactivations  $z_{ij} = w_{ij}x_i$  during prediction time and their respective aggregation  $z_j = \sum_i z_{ij} + b_j$  in the neuron  $j$ . Because this rule might lead to unbounded relevances when  $z_j$  approaches zero (Bach et al. 2015), additional advancements have been proposed.

#### b. $LRP_{\alpha\beta}$

In this rule, positive and negative preactivations  $z_{ij}$  are considered separately so that the denominators are always nonzero:

$$R_i^{(l)} = \sum_j \left( \alpha \frac{z_{ij}^+}{z_j^+} + \beta \frac{z_{ij}^-}{z_j^-} \right) R_j^{(l+1)}, \quad (\text{B6})$$

where

$$z_{ij}^+ = \begin{cases} z_{ij}; & z_{ij} > 0 \\ 0 & \end{cases} \quad \text{and} \quad z_{ij}^- = \begin{cases} 0 & \\ z_{ij}; & z_{ij} < 0. \end{cases}$$

In our study, we use the commonly used version of this rule where  $\alpha = 1$  and  $\beta = 0$ , which considers only positive preactivations (Bach et al. 2015).

#### c. $LRP_{comp}$

Because of the different strengths and weaknesses of the  $LRP_z$  and  $LRP_{\alpha\beta}$  rules that we discuss in the results section, a composite rule that combines these two rules has been recently suggested in the literature (Kohlbrenner et al. 2020). This composite rule essentially applies the  $LRP_z$  rule to propagate the relevance in the fully connected layers of the CNN and applies the  $LRP_{\alpha\beta}$  rule for the convolutional layers of the CNN. The aim is to combine the strengths and limit the effects of the weaknesses of the two rules. This rule has been suggested as a ‘‘best practice’’ implementation of LRP when explaining a deep CNN (Kohlbrenner et al. 2020).

#### d. $LRP_{comp/flat}$

This rule is an extension of  $LRP_{comp}$ . It implements the rules  $LRP_z$  and  $LRP_{\alpha\beta}$  exactly the same way as the  $LRP_{comp}$  rule but additionally implements a flat rule in the very lowest layer(s). The flat rule distributes the relevance of a neuron uniformly to all connected neurons in the lower layer. It is designed to be used for convolutional layers and it is not suitable for fully connected layers:

$$R_i^{(l)} = \sum_j \frac{1}{\sum_i 1} R_j^{(l+1)}. \quad (\text{B7})$$

The motivation behind this flat rule is that it allows the user to modify the resolution of the heat map by changing which layers the flat rule is applied to (e.g., only at the input layer or the lowest three layers). If the user is not interested in local accuracy, but they only need to obtain a coarse picture of the relevance, this is a suitable rule to use. Another important aspect of this rule is that it is invariant to any transformation of the input [see section 3 and Bach et al. (2016)].

For each neuron at an upper layer ( $l + 1$ ), the Deep Taylor decomposition attribution method (Montavon et al. 2017) computes a root point  $\hat{x}_i^l$  close to the input  $x_i$ , for which the neuron’s relevance is zero, and uses the difference ( $x_i - \hat{x}_i^l$ ) to estimate the relevance of the lower-layer neurons recursively. The relevance redistribution is performed as follows:

$$R_i^{(l)} = \sum_j \frac{\partial R_j^{(l+1)}}{\partial x_i} \Bigg|_{x_i = \hat{x}_i^l} (x_i - \hat{x}_i^l), \quad (\text{B8})$$

where  $R_j^{(l+1)}$  is the relevance of the neuron  $j$  at the upper layer ( $l + 1$ ) and  $R_i^{(l)}$  is the relevance of the neuron  $i$  at the lower layer ( $l$ ). It has been shown in (Samek et al. 2016; Montavon et al. 2017) that, for neural networks with ReLU activations, Deep Taylor leads to results that are identical to the  $LRP_{\alpha\beta 0}$  rule.

The PatternNet and PatternAttribution methods are based on the idea that every input consists of a signal component (all of the information in the input that is relevant to the prediction task) and a distractor (all of the distracting information that is irrelevant to the prediction task). Kindermans et al. (2018) argued that most existing XAI methods do not necessarily disentangle the signal and the distractor before attributing the output to the input. In fact, the authors showed that even for a simple linear regression model, the vector of weights (i.e., regression coefficients) that is typically used to interpret the model is not necessarily aligned with the direction of the signal in the input (Kindermans et al. 2018). Thus, Kindermans et al. (2018) argued that to explain a model one needs to develop an approach that distinguishes between the signal and the distractor in the input, and they proposed PatternNet to estimate the signal in the input and PatternAttribution to then attribute each prediction to the input. Both methods implement a layerwise propagation of the prediction back to lower layers until the input layer is reached and the signal or the attribution is obtained (i.e., similar to the LRP method). The propagation rules are

$$s_i^{(l)} = \sum_j \alpha_{ij} s_j^{(l+1)} \quad \text{and} \quad (\text{B9a})$$

$$R_i^{(l)} = \sum_j w_{ij} \alpha_{ij} R_j^{(l+1)} \quad (\text{B9b})$$



for PatternNet and PatternAttribution, respectively. Above,  $s_i^{(l)}$  and  $R_i^{(l)}$  are the signal and the attribution (relevance) of neuron  $i$  in the layer ( $l$ ). In both methods, the summation over  $j$  considers only the neurons in the upper layer ( $l + 1$ ) that were activated in the forward pass of the specific prediction. The symbol  $w_{ij}$  represents the weight from neuron  $i$  to neuron  $j$  and the vector  $\alpha_j = \{\alpha_{ij}, \forall i\}^T$  represents the direction of the signal in the neurons of the layer ( $l$ ) and the neuron  $j$  and is estimated using the training dataset as

$$\alpha_j = \frac{E_+(\mathbf{x}, z_j) - E_+(\mathbf{x})E(z_j)}{\mathbf{w}_j^T E_+(\mathbf{x}, z_j) - \mathbf{w}_j^T E_+(\mathbf{x})E(z_j)}, \quad (\text{B9c})$$

where  $\mathbf{w}_j = \{w_{ij}, \forall i\}^T$  is the weight vector,  $\mathbf{x} = \{x_j, \forall i\}^T$  is the vector with all of the activations of the neurons  $i$  in the layer ( $l$ ) and  $z_j$  is their linear projection in the neuron  $j$ . The symbol  $E_+$  indicates that the expectation is only taken over those training samples that correspond to positive  $z_j$ . Note that the expressions in the above ratio represent the covariance of  $\mathbf{x}$  and  $z_j$ .

Deep SHAP is an attribution method that is based on the use of Shapley values (Shapley 1953) and is specifically designed for deep neural networks (Lundberg and Lee 2017). The Shapley values originate from the field of cooperative game theory and represent the average expected marginal contribution of each player in a cooperative game, after all possible combinations of players have been considered (Shapley 1953). For the importance of Shapley values to XAI, it can be shown (Lundberg and Lee 2017) that across all *additive feature attribution methods* (a general class of attribution methods that unifies many popular XAI methods like LRP), the only method that satisfies all desired properties of local accuracy, missingness and consistency [see Lundberg and Lee (2017) for details on these properties] emerges when the feature attributions  $\varphi_i$  are equal to the Shapley values:

$$\varphi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{|M|} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)], \quad (\text{B10})$$

where  $M$  is the set of all input features,  $M \setminus \{i\}$  is the set  $M$  but with the feature  $x_i$  being withheld,  $|M|$  represents the number of features in  $M$ , and the expression  $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$  represents the net contribution (effect) of the feature  $x_i$  to the outcome of the model  $f$ , which is calculated as the difference between the model outcome when the feature  $x_i$  is present and when it is withheld. Thus, the Shapley value  $\varphi_i$  is the (weighted) average contribution of the feature  $x_i$  across all possible subsets  $S \subseteq M \setminus \{i\}$ . Because of computational constraints, Deep SHAP approximates the Shapley values for the entire network by computing the Shapley values for smaller components of the network and propagating them backward until the input layer is reached (similar in philosophy to LRP, PatternNet, and PatternAttribution).

## APPENDIX C

### Sensitivity versus Attribution

When explaining a black-box model to a human, one typically aims to disentangle which input features were important/relevant for a specific prediction made by the model. The way to define what “being an important feature” means is not unique, and different definitions or methods to estimate feature importance can lead to different insights. Two of the most important categories of methods that aim to estimate feature importance are the methods that estimate *sensitivity* and the methods that estimate *attribution*. Here we want to briefly clarify the conceptual difference between the two.

Sensitivity refers to how sensitive the value of the output is to a specific input feature. An obvious way to estimate sensitivity is to calculate the first partial derivative of the network function  $\hat{F}$  with respect to the input feature of interest. This is what methods like Gradient and Smooth Gradient aim to do. Attribution, on the other hand, refers to the relative contribution of a specific input feature to the output. When dealing with complex models like deep neural networks, estimating attribution becomes complicated and many methods like the Layerwise Relevance Propagation, Pattern Attribution, and Deep SHAP have been proposed for this task.

To give an illustrative example of the difference between sensitivity and attribution, let us consider a simple non-linear function  $Y = F(X_1, X_2) = \sin(X_1) + \cos(X_2)$ . We can easily calculate that at the point  $(X_1, X_2) = (0, 0)$ , we get  $Y_{0,0} = F(0, 0) = 0 + 1 = 1$ . If we were to explain this output  $Y_{0,0}$ , that is, if we were to argue about which feature from  $X_1, X_2$  was more important for it, we would get conceptually and numerically different answers using a sensitivity versus an attribution perspective. In terms of sensitivity, the output  $Y_{0,0}$  is more sensitive to the value of feature  $X_1$  than feature  $X_2$  because  $(\partial F / \partial X_1)|_{0,0} = \cos(0) = 1$ , whereas  $(\partial F / \partial X_2)|_{0,0} = -\sin(0) = 0$ . In terms of attribution, the opposite is true, that is, the feature  $X_2$  contributes more to prediction  $Y_{0,0}$  because  $\sin(X_1)|_{X_1=0} = \sin(0) = 0$ , whereas  $\cos(X_2)|_{X_2=0} = \cos(0) = 1$ .

Apart from the numerical difference, the conceptual difference between the sensitivity and attribution can be more clearly realized if we think about the units of the results in the two cases. When estimating sensitivity, the units of the importance or relevance are [units of output/units of input], whereas, when estimating attribution, the units of the results are [units of output]. Thus, these two ways of explaining a black-box model are conceptually (and numerically) different, but they can both be insightful in different ways to a human, and thus they are equally valuable.

## REFERENCES

- Agapiou, A., 2017: Remote sensing in a petabyte-scale: Satellite data and heritage Earth Engine© applications. *Int. J. Digital Earth*, **10**, 85–102, <https://doi.org/10.1080/17538947.2016.1250829>.

- Ancona, M., E. Ceolini, C. Öztireli and M. Gross, 2018: Towards better understanding of gradient-based attribution methods for deep neural networks. *Sixth Int. Conf. on Learning Representations*, Vancouver, Canada, ICLR, <https://openreview.net/pdf?id=Sy21R9JAW>.
- , —, —, and —, 2019: Gradient-based attribution methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek et al., Eds., Lecture Notes in Computer Science, Vol. 11700, Springer, 169–191.
- Arras, L., A. Osman, and W. Samek, 2021: Ground truth evaluation of neural network explanations with CLEVR-XAI. arXiv, 2003.07258v2, <https://doi.org/10.48550/arXiv.2003.07258>.
- 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**, e0130140, <https://doi.org/10.1371/journal.pone.0130140>.
- , —, K.-R. Müller, and W. Samek, 2016: Controlling explanatory heatmap resolution and semantics via decomposition depth. *Proc. IEEE Int. Conf. on Image Processing*, Phoenix, AZ, IEEE, 2271–2275, <https://doi.org/10.1109/ICIP.2016.7532763>.
- Balduzzi, D., M. Frean, L. Leary, J. P. Lewis, K. W.-D. Ma, and B. McWilliams, 2017: The shattered gradients problem: If re-snets are the answer, then what is the question? *Proc. 34th Int. Conf. on Machine Learning*, Sydney, Australia, ICML, 342–350, <http://proceedings.mlr.press/v70/balduzzi17b/balduzzi17b.pdf>.
- Barnes, E. A., J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2019: Viewing forced climate patterns through an AI Lens. *Geophys. Res. Lett.*, **46**, 13 389–13 398, <https://doi.org/10.1029/2019GL084944>.
- , B. Toms, J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2020: Indicator patterns of forced changed learned by an artificial neural network. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002195, <https://doi.org/10.1029/2020MS002195>.
- Bergen, K. J., P. A. Johnson, M. V. De Hoop, and G. C. Beroza, 2019: Machine learning for data-driven discovery in solid Earth geoscience. *Science*, **363**, eaau0323, <https://doi.org/10.1126/science.aau0323>.
- Buhrmester, V., D. Münch, and M. Arens, 2019: Analysis of explainers of black box deep neural networks for computer vision: A survey. *Mach. Learn. Knowl. Extr.*, **3**, 966–989, <https://doi.org/10.3390/make3040048>.
- Das, A., and P. Rad, 2020: Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv, 2006.11371v2, <https://doi.org/10.48550/arXiv.2006.11371>.
- Dombrowski, A.-K., C. J. Anders, K.-R. Müller, and P. Kessel, 2022: Towards robust explanations for deep neural networks. *Pattern Recognit.*, **121**, 108194, <https://doi.org/10.1016/j.patcog.2021.108194>.
- Ebert-Uphoff, I., and K. Hilburn, 2020: Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bull. Amer. Meteor. Soc.*, **101**, E2149–E2170, <https://doi.org/10.1175/BAMS-D-20-0097.1>.
- Guo, H., 2017: Big Earth data: A new frontier in Earth and information sciences. *Big Earth Data*, **1**, 4–20, <https://doi.org/10.1080/20964471.2017.1403062>.
- Ham, Y.-G., J.-H. Kim, and J.-J. Luo, 2019: Deep learning for multi-year ENSO forecasts. *Nature*, **573**, 568–572, <https://doi.org/10.1038/s41586-019-1559-7>.
- 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. *J. Appl. Meteor. Climatol.*, **60**, 3–21, <https://doi.org/10.1175/JAMC-D-20-0084.1>.
- Karpatne, A., I. Ebert-Uphoff, S. Ravela, H. A. Babaie and V. Kumarand, 2019: Machine learning for the geosciences: Challenges and opportunities. *IEEE Trans. Knowl. Data Eng.*, **31**, 1544–1554, <https://doi.org/10.1109/TKDE.2018.2861006>.
- Keys, P. W., E. A. Barnes, and N. H. Carter, 2021: A machine-learning approach to human footprint index estimation with applications to sustainable development. *Environ. Res. Lett.*, **16**, 044061, <https://doi.org/10.1088/1748-9326/abe00a>.
- Kindermans, P.-J., K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, 2018: Learning how to explain neural networks: PatternNet and PatternAttribution. *Sixth Int. Conf. on Learning Representations*, Vancouver, Canada, ICLR, <https://openreview.net/pdf?id=Hkn7CBaTW>.
- , S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, 2019: The (un)reliability of saliency methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek et al., Eds., Lecture Notes in Computer Science, Vol. 11700, Springer, 267–280.
- Kohlbrenner, M., A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lopuschkin, 2020: Towards best practice in explaining neural network decisions with LRP. *2020 Int. Joint Conf. on Neural Networks*, Glasgow, United Kingdom, IEEE, <https://doi.org/10.1109/IJCNN48605.2020.9206975>.
- Lary, D. J., A. H. Alavi, A. H. Gandomi, and A. L. Walker, 2016: Machine learning in geosciences and remote sensing. *Geosci. Front.*, **7**, 3–10, <https://doi.org/10.1016/j.gsf.2015.07.003>.
- Leavitt, M. L., and A. S. Morcos, 2020: Towards falsifiable interpretability research. arXiv, 2010.12016v1, <https://doi.org/10.48550/arXiv.2010.12016>.
- LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444, <https://doi.org/10.1038/nature14539>.
- Lundberg, S. M., and S. I. Lee, 2017: A unified approach to interpreting model predictions. *Proc. 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, NIPS, 4768–4777, <https://dl.acm.org/doi/10.5555/3295222.3295230>.
- Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2022a: Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environ. Data Sci.*, **1**, e8, <https://doi.org/10.1017/eds.2022.7>.
- , —, and —, 2022b: Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. *Beyond Explainable Artificial Intelligence*, A. Holzinger et al., Eds., Lecture Notes in Computer Science, Vol. 13200, Springer, 315–339.
- Mayer, K. J., and E. A. Barnes, 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophys. Res. Lett.*, **48**, e2020GL092092, <https://doi.org/10.1029/2020GL092092>.
- McGovern, A., R. Lagerquist, D. J. Gagne II, 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. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, <https://doi.org/10.1175/BAMS-D-18-0195.1>.
- Montavon, G., S. Lopuschkin, A. Binder, W. Samek, and K.-R. Müller, 2017: Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.*, **65**, 211–222, <https://doi.org/10.1016/j.patcog.2016.11.008>.
- Overpeck, J. T., G. A. Meehl, S. Bony, and D. R. Easterling, 2011: Climate data challenges in the 21st century. *Science*, **331**, 700–702, <https://doi.org/10.1126/science.1197869>.

- Prabhat, and Coauthors, 2021: ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geosci. Model Dev.*, **14**, 107–124, <https://doi.org/10.5194/gmd-14-107-2021>.
- Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, <https://doi.org/10.1038/s41586-019-0912-1>.
- Reinsel, D., J. Gantz, and J. Rydning, 2018: The digitization of the world: From edge to core. IDC Seagate White Paper, 28 pp., <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>.
- Rolnick, D., and Coauthors, 2019: Tackling climate change with machine learning. arXiv, 1906.05433v2, <https://doi.org/10.48550/arXiv.1906.05433>.
- Rudin, C., 2019: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Learn.*, **1**, 206–215, <https://doi.org/10.1038/s42256-019-0048-x>.
- Samek, W., G. Montavon, A. Binder, S. Lapuschkin, and K.-R. Müller, 2016: Interpreting the predictions of complex ML models by layer-wise relevance propagation. arXiv, 1611.08191v1, <https://doi.org/10.48550/arXiv.1611.08191>.
- Shapley, L. S., 1953: A value for n-person games. *Contributions to the Theory of Games*, H. W. Kuhn, and A. W. Tucker, Eds., Annals of Mathematics Studies, Vol. 2, Princeton University Press, 307–317.
- Shen, C., 2018: A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.*, **54**, 8558–8593, <https://doi.org/10.1029/2018WR022643>.
- Shrikumar, A., P. Greenside, A. Shcherbina, and A. Kundaje, 2016: Not just a black box: Learning important features through propagating activation differences. arXiv, 1605.01713v3, <https://doi.org/10.48550/arXiv.1605.01713>.
- , —, and A. Kundaje, 2017: Learning important features through propagating activation differences. *Proc. 34th Int. Conf. on Machine Learning*, Sydney, Australia, ICLM, 3145–3153.
- Simonyan, K., A. Vedaldi, and A. Zisserman, 2014: Deep inside convolutional networks: Visualizing image classification models and saliency maps. arXiv, 1312.6034v2, <https://doi.org/10.48550/arXiv.1312.6034>.
- Sit, M., B. Z. Demiray, Z. Xiang, G. J. Ewing, Y. Sermet, and I. Demir, 2020: A comprehensive review of deep learning applications in hydrology and water resources. *Water Sci. Technol.*, **82**, 2635–2670, <https://doi.org/10.2166/wst.2020.369>.
- Smilkov, D., N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, 2017: SmoothGrad: Removing noise by adding noise. arXiv, 1706.03825v1, <https://doi.org/10.48550/arXiv.1706.03825>.
- Sonnevald, M., and R. Lguensat, 2021: Revealing the impact of global heating on North Atlantic circulation using transparent machine learning. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002496, <https://doi.org/10.1029/2021MS002496>.
- Sundararajan, M., A. Taly, and Q. Yan, 2017: Axiomatic attribution for deep networks. arXiv, 1703.01365v2, <https://doi.org/10.48550/arXiv.1703.01365>.
- Tjoa, E., and C. Guan, 2019: A survey on explainable artificial intelligence (XAI): Towards medical XAI. arXiv, 1907.07374v5, <https://doi.org/10.48550/arXiv.1907.07374>.
- Toms, B. A., E. A. Barnes, and I. Ebert-Uphoff, 2020: Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *J. Adv. Model. Earth Syst.*, **12**, e2019MS002002, <https://doi.org/10.1029/2019MS002002>.
- , —, and J. W. Hurrell, 2021: Assessing decadal predictability in an earth-system model using explainable neural networks. *Geophys. Res. Lett.*, **48**, e2021GL093842, <https://doi.org/10.1029/2021GL093842>.
- Zhou, Y., S. Booth, M. T. Ribeiro, and J. Shah, 2022: Do feature attribution methods correctly attribute features? arXiv, 2104.14403v2, <https://doi.org/10.48550/arXiv.2104.14403>.