LESSONS LEARNED

Carefully Choose the Baseline: Lessons Learned from Applying XAI Attribution Methods for Regression Tasks in Geoscience

ANTONIOS MAMALAKIS,[a] ELIZABETH A. BARNES,[a] AND IMME EBERT-UPHOFF[b,c]

[a] *Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado*
[b] *Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado*
[c] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*

(Manuscript received 25 July 2022, in final form 9 November 2022)

ABSTRACT: Methods of explainable artificial intelligence (XAI) are used in geoscientific applications to gain insights into the decision-making strategy of neural networks (NNs), highlighting which features in the input contribute the most to a NN prediction. Here, we discuss our "lesson learned" that the task of attributing a prediction to the input does not have a single solution. Instead, the attribution results depend greatly on the considered baseline that the XAI method utilizes—a fact that has been overlooked in the geoscientific literature. The baseline is a reference point to which the prediction is compared so that the prediction can be understood. This baseline can be chosen by the user or is set by construction in the method's algorithm—often without the user being aware of that choice. We highlight that different baselines can lead to different insights for different science questions and, thus, should be chosen accordingly. To illustrate the impact of the baseline, we use a large ensemble of historical and future climate simulations forced with the shared socioeconomic pathway 3-7.0 (SSP3-7.0) scenario and train a fully connected NN to predict the ensemble- and global-mean temperature (i.e., the forced global warming signal) given an annual temperature map from an individual ensemble member. We then use various XAI methods and different baselines to attribute the network predictions to the input. We show that attributions differ substantially when considering different baselines, because they correspond to answering different science questions. We conclude by discussing important implications and considerations about the use of baselines in XAI research.

SIGNIFICANCE STATEMENT: In recent years, methods of explainable artificial intelligence (XAI) have found great application in geoscientific applications, because they can be used to attribute the predictions of neural networks (NNs) to the input and interpret them physically. Here, we highlight that the attributions—and the physical interpretation—depend greatly on the choice of the baseline—a fact that has been overlooked in the geoscientific literature. We illustrate this dependence for a specific climate task, in which a NN is trained to predict the ensemble- and global-mean temperature (i.e., the forced global warming signal) given an annual temperature map from an individual ensemble member. We show that attributions differ substantially when considering different baselines, because they correspond to answering different science questions.

KEYWORDS: Artificial intelligence; Machine learning; Model interpretation and visualization; Neural networks

1. Introduction

Explainable artificial intelligence (XAI) aims to provide insights about the decision-making process of AI models and has been increasingly applied to the geosciences (e.g., Toms et al. 2021; Ebert-Uphoff and Hilburn 2020; Hilburn et al. 2021; Barnes et al. 2019, 2020; Mayer and Barnes 2021; Keys et al. 2021; Sonnewald and Lguensat 2021). XAI methods show promising results in calibrating model trust and assisting in learning new science (see for example, McGovern et al. 2019; Toms et al. 2020; Sonnewald and Lguensat 2021; Clare et al. 2022; Mamalakis et al. 2022a). A popular subcategory of XAI is the so-called local attribution methods, which compute the attribution of a model's prediction to the input variables (also referred to as "input features"). The final product typically comes in the form of a heat map in the shape of the original input. Because of the complex architecture of state-of-the-art AI models [e.g., neural networks (NNs)], the "attribution task" can be challenging, and many XAI methods have been shown to not honor desired properties (e.g., the so-called completeness property, input invariance property, etc.; see details in Kindermans et al. 2017; Ancona et al. 2018, 2019; Rudin 2019; Dombrowski et al. 2020). Intercomparison studies have shown that the faithfulness of the attributions (with respect to the network's decision-making process) and their comprehensibility (the degree to which they can be understood by the user) depend on the prediction setting and the model architecture, and that no XAI method likely exists that maximizes both for any application (Mamalakis et al. 2022b,c).

Apart from the issue of the fidelity of the methods, another very important aspect of the attribution task is that one needs

Corresponding author: A. Mamalakis, amamalak@colostate.edu

DOI: 10.1175/AIES-D-22-0058.1 e220058

to define the *baseline* that the attribution is calculated for, that is, the baseline is the reference point to which the prediction is compared so it can be understood. Without having defined a baseline, the attribution task cannot be solved—it is an ill-defined task. Defining a baseline is necessary not only when explaining a complex AI model, but also for explaining very simple models: thus, defining a baseline is tied to the attribution task itself and it is necessary independently from what model or what type of XAI method (model agnostic versus model specific) is being used. To illustrate the above, let us consider the simple case of a linear model with no constant term: $y = f(\mathbf{x}) = \sum_i w_i x_i$, and let us suppose that we want to attribute a prediction y_n to the input. A trivial solution to the attribution task for this case seems to be that any input feature x_i contributes $w_i x_{i,n}$ to the prediction y_n. However, this is a correct attribution rule only if we assume a baseline $\hat{\mathbf{x}} = \mathbf{0}$. Thus, even in cases in which the attribution task seems to have exactly one trivial solution, this solution corresponds to a baseline that has been assumed implicitly. The complete way to attribute the prediction y_n to the input is to consider the *set of solutions*: $w_i(x_{i,n} - \hat{x}_i)$. It can be observed that the attribution depends on the considered baseline and that the previous, trivial solution is simply a special case of the complete set of solutions. Another special case that is typically very informative to scientists is to choose a baseline of the form $\hat{\mathbf{x}} = E(\mathbf{x})$. In this case, the attribution would highlight those features in the input for which the deviation from the average state is important for the prediction.

In this paper, our aim is to highlight our lesson learned that the attribution task is highly dependent on the choice of baseline—a fact that seems to have been overlooked so far in applications of XAI to the geosciences. Also, we show by example how the use of different baselines offers an opportunity to answer different science questions. To do so, we use a large ensemble of historical and future climate simulations forced with the shared socioeconomic pathway 3-7.0 (SSP3-7.0) climate change scenario (ensemble of 80 members from the climate model CESM2; Rodgers et al. 2021) and train a fully connected NN to predict the ensemble- and global-mean temperature (i.e., the forced component of the global mean temperature) given an annual temperature map from an individual ensemble member. We then use XAI methods and different baselines to attribute the network predictions to the input. We show that attributions differ substantially when considering different baselines, as they correspond to answering different science questions.

In section 2, we provide details about the data, prediction task, and methods, and in section 3 we present our results. Section 4 discusses some considerations about the use of baselines in XAI research, and in section 5 we provide a summary of the key points of our study.

2. Data and methods

a. Data

We use yearly mean surface air temperature from the CESM2 Large Ensemble Community project (Rodgers et al. 2021), spanning the years from 1850 to 2100 (publicly available at https://www.cesm.ucar.edu/projects/community-projects/LENS2/data-sets.html). Historical forcing is applied to the climate system over the 1850–2014 period, and the SSP3-7.0 climate change scenario is applied over the 2015–2100 period. We use the output of 80 members bilinearly regridded to a 2.5° × 2.5° resolution from an approximate 1° × 1° resolution to reduce dimensionality of the prediction task.

b. Prediction task and network architecture

Given a yearly mean temperature map as an input (10 512 pixels), we train a fully connected NN to predict the ensemble- and global-mean temperature of the same year as the input map (see graphic in Fig. 1). This task requires the NN to recognize the forced signal in the temperature field (i.e., the signal originating from natural forcings, such as solar radiation and volcanic eruptions, or anthropogenic forcings, such as changes in greenhouse gases concentrations, tropospheric aerosols, land use, etc.) so as to estimate the forced global mean temperature while ignoring any internal variability signals that may be present in the input [e.g., an active El Niño–Southern Oscillation (ENSO)]. Thus, the NN needs to be able to separate the forced temperature response from that of the internal climate "noise."

Our network consists of two hidden layers with eight and five neurons each (with ReLU activation functions) and one output neuron with no activation. Dropout (with probability of 0.3) and ridge regularization (with a regularization factor of 0.005) are applied in the input layer to avoid overfitting, and training is performed using the Adam optimizer (Kingma and Ba 2017), a batch size of 32, and a learning rate of 0.0001. The mean square error is used as the loss function during training. We train on 70 members and test on the remaining 10 members. The performance of the network is satisfying, with a mean absolute error on the order of 0.068°C for the testing data [and a coefficient of determination (R^2) on the order of 99%].

c. XAI methods

The main goal of the current work is the explanation rather than the prediction itself. To explain the predictions of the network we use two local attribution methods, namely, the integrated gradients (Sundararajan et al. 2017) and deep Shapely additive explanations (SHAP) (Lundberg and Lee 2017). We have chosen these methods for two main reasons: (i) Both methods allow the user to define the baseline for which the attribution is derived, which is the focus of this work and allows us to gain insights into different science questions (see Fig. 1). (ii) Both methods satisfy the *completeness* property [also referred to as *local accuracy* in Lundberg and Lee (2017) or *sensitivity-N* in Ancona et al. (2018)], which holds that the feature attributions must add up to the difference between the NN output at the current sample $\mathbf{x}_n$ and the baseline $\hat{\mathbf{x}}$. We briefly describe the two methods below.

1) Integrated gradients

This method (Sundararajan et al. 2017) is a local attribution method that builds on the input*gradient method (Shrikumar

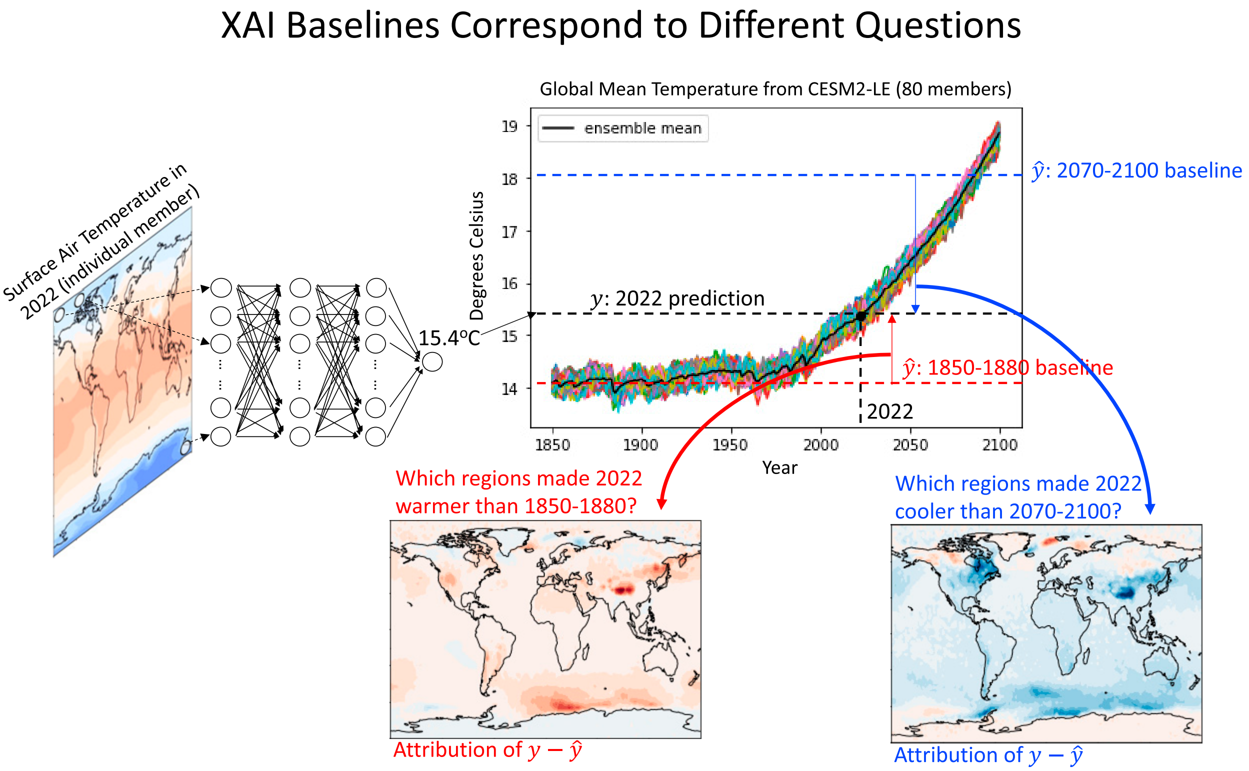
XAI Baselines Correspond to Different Questions
Surface Air Temperature in 2022 (individual member)
15.4°C
Global Mean Temperature from CESM2-LE (80 members)
ensemble mean
Degrees Celsius
19
18
17
16
15
14
1850
1900
1950
2000
2050
2100
Year
$\hat{y}$: 2070-2100 baseline
y: 2022 prediction
$\hat{y}$: 1850-1880 baseline
2022
Which regions made 2022 warmer than 1850-1880?
Which regions made 2022 cooler than 2070-2100?
Attribution of $y - \hat{y}$
Attribution of $y - \hat{y}$

FIG. 1. Schematic representation of the prediction task of the study and of the use of baselines to gain insights into different science questions.

et al. 2016, 2017). Namely, it aims to account for the fact that in nonlinear problems the gradient is not constant, and thus, the product of the local gradient with the input might not be a good approximation of the input's contribution. The integrated gradients method considers a baseline vector $\hat{\mathbf{x}}$, and the attribution for feature i and sample n is equal to the product of the distance of the input from the baseline with the average of the gradients at points along the straight-line path from the baseline to the input:

$$R_{i,n} = (x_{i,n} - \hat{x}_i) \times \frac{1}{m}\sum_{j=1}^{m} \frac{\partial \hat{F}}{\partial X_i}\bigg|_{X_i=\hat{x}_i+(j/m)(x_{i,n}-\hat{x}_i)}, \qquad (1)$$

where $\hat{F}$ represents the network and m is the number of steps in the Riemann approximation.

2) DEEP SHAP

Deep SHAP is a local attribution method that is based on the use of Shapley values (Shapley 1953) and is specifically designed for 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). Regarding the importance of Shapley values for XAI, it can be shown (Lundberg and Lee 2017) that across all *additive feature attribution methods* (a general class of local attribution methods that unifies many popular XAI methods), 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 φ_i are equal to the Shapley values:

$$\varphi_i = \sum_{S\subseteq M\setminus\{i\}} \frac{|S|!(|M|-|S|-1)!}{|M|}[f_{S\bigcup\{i\}}(x_{S\bigcup\{i\}}) - f_S(x_S)] \qquad (2)$$

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 φ_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 [i.e., implementing Eq. (2) recursively]. Deep SHAP may consider different baselines (defined by the user) with which to replace features x_i, when they need to be withheld.

3. Results

In this section, we focus on attributing a specific prediction of the NN using different baselines. We choose the 2022

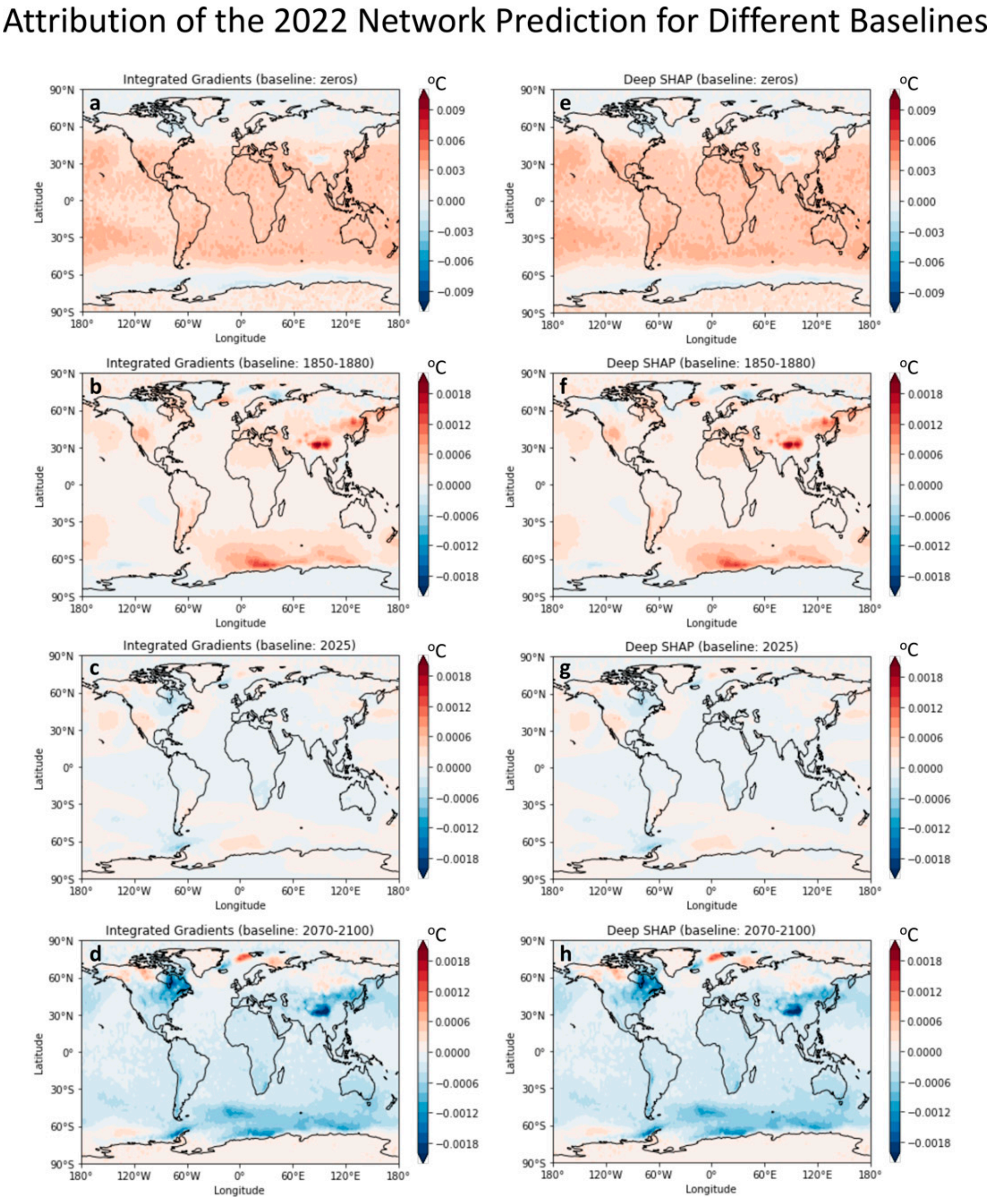
Attribution of the 2022 Network Prediction for Different Baselines
Integrated Gradients (baseline: zeros)
Deep SHAP (baseline: zeros)
Integrated Gradients (baseline: 1850-1880)
Deep SHAP (baseline: 1850-1880)
Integrated Gradients (baseline: 2025)
Deep SHAP (baseline: 2025)
Integrated Gradients (baseline: 2070-2100)
Deep SHAP (baseline: 2070-2100)
a
b
c
d
e
f
g
h
°C
0.009
0.006
0.003
0.000
−0.003
−0.006
−0.009
0.0018
0.0012
0.0006
0.0000
−0.0006
−0.0012
−0.0018
Latitude
90°N
60°N
30°N
0°
30°S
60°S
90°S
Longitude
180°
120°W
60°W
0°
60°E
120°E
180°

FIG. 2. Attribution heat maps (°C) derived by the methods (left) integrated gradients and (right) deep SHAP for the NN prediction for 2022 (80th member), using four different baselines: (a),(e) the zero input; (b),(f) the average temperature over 1850–80 (from the 80th member); (c),(g) the temperature in 2025 (from the 80th member); and (d),(h) the average temperature over 2070–2100 (from the 80th member).

temperature map from the 80th ensemble member as the input of interest. The attributions based on the integrated gradients and the deep SHAP are shown in Fig. 2 for four different baselines. Figure 3 shows the difference between the considered input to the NN and the four baselines. We note that for the considered input, the NN estimates the forced global-mean temperature to be 15.4°C, which is almost identical to the true ensemble- and global-mean temperature for 2022.

We first consider the baseline of zero input (see Figs. 2a,e and 3a); this is a very common choice made by scientists for XAI applications and the default option for many XAI methods. Reasonably, the NN output that corresponds to a zero input (although such an input was not present in the training or the testing set) is almost 0°C. Thus, with this choice of baseline, the question that we will gain insight into here is as follows: "Which patterns in the 2022 map made the network

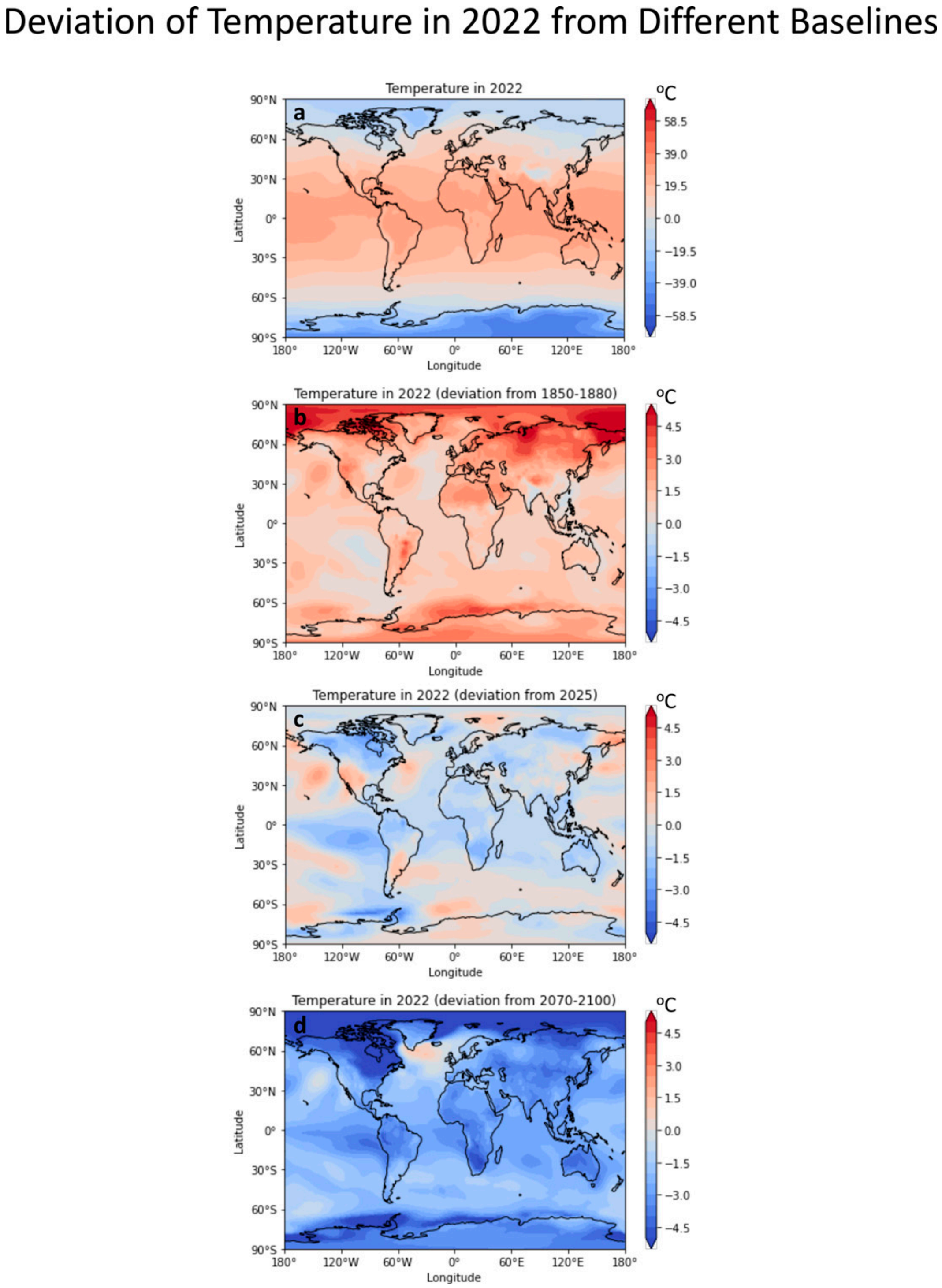
Deviation of Temperature in 2022 from Different Baselines
Temperature in 2022
a
°C
58.5
39.0
19.5
0.0
−19.5
−39.0
−58.5
Temperature in 2022 (deviation from 1850-1880)
b
Temperature in 2022 (deviation from 2025)
c
Temperature in 2022 (deviation from 2070-2100)
d
°C
4.5
3.0
1.5
0.0
−1.5
−3.0
−4.5
Latitude
90°N
60°N
30°N
0°
30°S
60°S
90°S
Longitude
180°
120°W
60°W
0°
60°E
120°E
180°

FIG. 3. Deviation (°C) of the 2022 surface air temperature from the four baselines used in Fig. 2.

predict the forced global-mean temperature to be 15.4°C as opposed to 0°C?" According to both XAI methods (the two methods provide almost identical results for all of the baselines), the features that determined this difference mainly occur over the zone of 55°S–55°N and partially over Antarctica. Thus, this attribution indicates that the NN has learned the basic concept that the temperature of the globe cannot be 0°C because of the heat stored in the tropics and subtropics originating from the solar radiation and the greenhouse effect. Also, note that the majority of the attributions are positive, since all attributions must add up to a positive value of 15.4°C; that is, recall that the attributions need to add up to the difference of the NN output at the current 2022 map (15.4°C) minus the NN output at the baseline (0°C).

As a second baseline, we consider the temperature map averaged over the 1850–80 period from the 80th ensemble member. The NN output that corresponds to this baseline is 14.1°C, and thus the question that we aim to answer here is the following: "Which patterns in the 2022 map made the network predict the forced global-mean temperature to be 15.4°C as opposed to 14.1°C?" or, alternatively, "Which regions made 2022 warmer than the period 1850–80 by 1.3°C?" The XAI methods highlight positive attributions mostly over land and oceanic regions in the midlatitudes and specifically over the Himalayas, eastern Asia, North and South America, the Southern Ocean, and the northern Pacific Ocean. The high latitudes show negligible contribution, despite the high degree of warming that occurs locally (see Fig. 3b). We hypothesize that this is due to the high internal variability that is associated with the warming of these regions, and thus, they may not constitute robust predictors for the NN to determine the forced global warming.

As a third baseline, we next consider the year 2025 (from the 80th ensemble member). The NN output for the baseline is 15.7°C; thus, the attributions need to add up to only 15.4° − 15.7°C = −0.3°C. Consequently, attributions are of a lower magnitude in this case. The question answered here is "which patterns in the 2022 map made the network predict the forced global-mean temperature to be 15.4°C as opposed to 15.7°C?" or alternatively "which regions made 2022 cooler than 2025 by 0.3°C?" The regions of the North Pacific and Southern Oceans are highlighted the most. We note that the strong pattern of an active ENSO over the eastern tropical Pacific in 2025 (see Fig. 3c) is not highlighted by the XAI methods. This suggests that the network has learned that ENSO variability constitutes internal variability, thus, it is not an appropriate predictor of the forced global-mean temperature.

Last, we consider the average temperature over the period 2070–2100 (from the 80th ensemble member) as the baseline. In this case, we are interested in gaining insights about the regions that made 2022 cooler than the end of the century. Both XAI methods highlight the majority of the globe with negative attribution, since the attributions need to add up to 15.4° − 18°C = −2.6°C. The most important contributors are shown to be the Himalayas, eastern Asia, North America, the Southern Ocean, and the northern Pacific Ocean. Similar to the remarks in the case of the 1850–80 baseline, the high latitudes are shown to contribute only slightly.

The above results make clear that the attribution task depends substantially on the considered baseline. This is true both in terms of the magnitude of the attributions, because of the completeness property (i.e., note the fivefold difference in the color scale between the top panels and the rest of the panels in Fig. 2), but also in terms of spatial patterns. For example, although the Himalayas is shown to be a very important region for distinguishing the forced global warming between the year 2022 and, for example, the 1850–80 period (see Figs. 2b,f), it is not a strong determining factor in the case of a zero baseline (see Figs. 2a,e; the same may apply for other regions, e.g., the Southern Ocean). In contrast, many regions in the deep tropics that are highlighted in the attribution when considering a zero baseline are not highlighted for the other three baselines. This highlights the importance of explicitly declaring the baseline, to avoid misidentifying or overlooking predictive (or causal) factors. At the same time, the above results illustrate that the ill-defined nature of the attribution task is beneficial in that different baselines can be considered to gain insights into different science questions of varying complexity. After repeating the analysis with a deeper network (using 12 hidden layers instead of 2), we obtained very similar results (not shown), which confirms that the baseline choice affects any attribution task independently from what model architecture (or type of XAI) is being used.

4. Implications for the use of XAI methods in geoscientific research

The dependency of the attribution task on the baseline highlights a few important considerations for the use of XAI methods in geoscience. First, it means that when comparing explanations using a variety of attribution methods, one must ensure that the same baseline is used for every method to avoid introducing artificial intermethod discrepancies that might be misinterpreted as intermethod variability. For example, in Fig. 2, we show that the considered methods integrated gradients and deep SHAP provide almost identical results for the same baseline (likely because of the simple, semilinear nature of the prediction task). However, it would be incorrect to compare their results for different baselines (e.g., comparing Fig. 2a with Fig. 2f), as they refer to different questions. We highlight this since in current geoscientific research, the baseline is typically not discussed at all when XAI attribution methods are applied.

Second, one needs to keep in mind that although many XAI methods allow the user to choose the baseline (e.g., as the ones used herein), some XAI methods assume specific types of baselines by construction, thus they should be used with caution. For example, unless extra modifications are implemented (Letzgus et al. 2021), the methods layerwise relevance propagation (Bach et al. 2015; Samek et al. 2016) and input*gradient (Shrikumar et al. 2016, 2017) provide attributions using a zero baseline. This implies that a zero-input value is automatically assigned a zero attribution, although the presence of a zero value might be important for the prediction when viewed from a different baseline that might be of interest (this was discussed as the "ignorant to zero input" issue in Mamalakis et al. 2022c).

Last, we note that XAI baselines are also very relevant and impactful in classification settings, but they should be used differently than in regression settings. We showed here that, in regression settings, different baselines form the necessary decision boundaries (i.e., the questions "X as opposed to Y") for the user to understand why certain decisions were made. In classification settings, these decision boundaries are already existing and predefined by the prediction classes. Thus, in classification settings, there is less need to consider multiple baselines to answer different questions. In fact, for classification tasks, a single baseline should suffice. The choice of this baseline is quite important, and as Sundararajan et al. (2017) suggested, it should be chosen so that it corresponds to a uniform distribution of baseline likelihoods for all classes, or in simple words, it contains no signal or information. In this way, the attribution for any class will be a function only of the input, without the presence of artifacts originating from considering a baseline that is informative (Sundararajan et al. 2017).

5. Summary

In this study, we highlight our "lesson learned" that the attribution task, that is, attributing a model's certain output to the corresponding input, does not have a single solution. Rather, the attributions and their interpretation depend greatly on the choice of the baseline; an issue often overlooked in the geoscientific literature. We illustrated this in a climate prediction task, where a fully connected network was trained to predict the ensemble- and global-mean temperature of annual temperature maps from a large ensemble of climate simulations. Our results make clear that when considering different baselines, attributions differ substantially both in magnitude and in spatial patterns.

We suggest that the dependence of the attribution task on the baseline choice is actually beneficial, since we can use different baselines to gain insights on different science questions of varying complexity. We also highlight that in regression settings, the issue of the baseline needs to be cautiously considered to avoid misidentifying sources of predictability and/or artificially introducing intermethod discrepancies in XAI applications. In classification settings, a single, noninformative baseline should suffice, since the decision boundaries are predefined by the prediction classes. Nevertheless, in all XAI applications, the baseline needs to be carefully chosen and explicitly stated.

Appropriate use and/or experimentation with multiple baselines will be advantageous for many XAI-pursued goals in geoscientific applications. These include deciphering the decision-making process of the network better (e.g., McGovern et al. 2019; Toms et al. 2020), accelerating the learning of new science (e.g., Sonnewald and Lguensat 2021; Clare et al. 2022; Mamalakis et al. 2022a) and potentially helping to identify problems in the training dataset (Sun et al. 2022).

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 AI Institute Grant ICER-2019758. The authors also thank the the CESM2 Large Ensemble team for their efforts in making their data publicly available and the IBM Center for Climate Physics in South Korea for providing supercomputing resources.

Data availability statement. The CESM2-LE model output is publicly available online (https://www.cesm.ucar.edu/projects/community-projects/LENS2/data-sets.html). The code to reproduce the presented results is available online (https://github.com/amamalak/XAI-baselines).

REFERENCES

Ancona, M., E. Ceolini, C. Öztireli, and M. Gross, 2018: Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv, 1711.06104v4, https://doi.org/10.48550/arXiv.1711.06104.

——, ——, ——, 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.

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.

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.

Clare, M. C. A., M. Sonnewald, R. Lguensat, J. Deshayes, and V. Balaji, 2022: Explainable artificial intelligence for Bayesian neural networks: Towards trustworthy predictions of ocean dynamics. arXiv, 2205.00202v1, https://doi.org/10.48550/arXiv.2205.00202.

Dombrowski, A.-K., C. J. Anders, K.-R. Müller, and P. Kessel, 2020: Towards robust explanations for deep neural networks. arXiv, 2012.10425v1, https://doi.org/10.48550/arXiv.2012.10425.

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.

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.

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., S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, 2017: The (un)reliability of saliency methods. arXiv, 1711.00867v1, https://doi.org/10.48550/arXiv.1711.00867.

Kingma, P. D., and J. L. Ba, 2017: ADAM: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

Letzgus, S., P. Wagner, J. Lederer, W. Samek, K.-R. Müller, and G. Montavon, 2021: Toward explainable AI for regression models. arXiv, 2112.11407v1, https://doi.org/10.48550/arXiv.2112.11407.

Lundberg, S. M., and S. I. Lee, 2017: A unified approach to interpreting model predictions. *Proc. 31st Conf. on Neural Information Processing Systems*, Long Beach, CA, Association for Computing Machinery, 4768–4777, https://dl.acm.org/doi/pdf/10.5555/3295222.3295230.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2022a: 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.

——, ——, and ——, 2022b: 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.

——, E. A. Barnes, and I. Ebert-Uphoff, 2022c: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. arXiv, 2202.03407v2, https://doi.org/10.1175/AIES-D-22-0012.1.

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.

Rodgers, K. B., and Coauthors, 2021: Ubiquity of human-induced changes in climate variability. *Earth Syst. Dyn.*, **12**, 1393–1411, https://doi.org/10.5194/esd-12-1393-2021.

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 2.28*, H. A. Kuhn and A. W. Tucker, Ed., Vol. 2, Princeton University Press, 307–318.

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. arXiv, 1704.02685v2, https://doi.org/10.48550/arXiv.1704.02685.

Sonnewald, M., and R. Lguensat, 2021: Revealing the impact of global heating on North Atlantic circulation using transparent machine learning. *Adv. Model. Earth Syst.*, **13**, e2021MS002496, https://doi.org/10.1029/2021MS002496.

Sun, Z., and Coauthors, 2022: A review of Earth artificial intelligence. *Comput. Geosci.*, **159**, 105034, https://doi.org/10.1016/j.cageo.2022.105034.

Sundararajan, M., A. Taly, and Q. Yan, 2017: Axiomatic attribution for deep networks. arXiv, 1703.01365v2, https://doi.org/10.48550/arXiv.1703.01365.

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, 2104.14403v3, https://doi.org/10.48550/arXiv.2104.14403.