AGU PUBLICATIONS

Geophysical Research Letters

Supporting Information for

Quantifying "climate distinguishability" after stratospheric aerosol injection using explainable artificial intelligence

Antonios Mamalakis[1], Elizabeth A. Barnes[1] and James W. Hurrell[1]

[1] Department of Atmospheric Science, Colorado State University, Fort Collins, CO

Contents of this file

Introduction

In this document, supporting information for the manuscript entitled *Quantifying "climate distinguishability" after stratospheric aerosol injection using explainable artificial intelligence* is provided. Specifically, Text S1 discusses the details of the training approach of our neural networks and the strategy of how we determine the corresponding architectures (i.e., the choices of hyperparameter values). Text S2 provides details on the algorithm of Deep SHAP, which is used to gain insights on the decision-making process of our networks. Moreover, in Table S1 we present a list of all the variables used in our study, together with the corresponding temporal scales and domains of focus.

Text S1: Network training and architectures

For each of the two considered tasks (i.e., distinguishability under the SAI or under the SSP scenario) and for each variable of interest, we train a fully-connected neural network using a cross-validation approach: we use 8 simulation members out of the 10 that are available for training (i.e., to estimate the network's parameters), 1 member for validation (to estimate the network's hyperparameters; see below) and the remaining 1 member for testing (to assess performance and interpret the predictions). We repeat the above 10 times, each time using a different member as the testing one and different validation and training members accordingly. The presented results in the main text and the conclusions are based *only* on the testing results. We use the 40-year period 2020-2059 for our training and validation, whereas for testing, we additionally use the "out-of-sample" years 2060-2069 from the testing member to assess the generalizability of the distinctive patterns learned by the network.

Regarding the architecture of the network, for each task, for each variable, and for each iteration in the cross-validation sequence, we search across many combinations of hyperparameters. Specifically, we consider the following hyperparameters and corresponding search spaces: learning rate: [0.00001, 0.0001, 0.001, 0.01]; dropout probability in the input layer: [0.1, 0.25, 0.5, 0.75]; number of hidden layers: [0, 1, 2, 4]; number of neurons per hidden layer: [3, 5, 10, 25]. We quantify the validation loss (after 50 epochs of training) for each of the combinations of hyperparameters and we choose the one with the lowest loss. We then train the network using the chosen architecture for 10,000 epochs and using an early stopping approach with a patience parameter equal to 30 and a batch size equal to 32. We use ReLU activation functions for all hidden layers. The output layer consists of a single neuron with a sigmoid activation function.

The same training approach as described above is used for both tasks and for all variables. Thus, the difference in the network's performance across different cases signifies the diversity of SAI impacts and the degree to which distinctive patterns exist in the data or not. Indeed, in some cases the network performs with almost 100% classification accuracy, while in other cases, it performs no better than random chance, as we show in section 3 of the main text.

Text S2: Deep SHAP

Deep SHAP is an attribution method that aims to identify the relative contribution of each of the input variables (features) to a specific model output (local attribution method). It 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 to explainable artificial intelligence, 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 methods like Layer-wise Relevance Propagation, Bach et al., 2015, DeepLIFT, Shrikumar et al., 2016, etc.), 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:

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

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 $f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S)$ 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 $S \subseteq M \setminus \{i\}$. Due to computational constraints, Deep SHAP approximates the contribution of each feature in the input to the network's prediction by computing the Shapley values for small components of the network and propagating them backwards until the input layer is reached and the input attributions are computed. For more details on Deep SHAP, the reader is referred to the original study by Lundberg and Lee (2017).

Supplementary Table S1. List of variables used in our study together with their corresponding temporal scales and domains of focus.

VARIABLE	TEMPORAL FOCUS	DOMAIN OF FOCUS
surface temperature	annual mean	global
surface temperature	annual mean	global land
surface temperature	annual max	global
surface temperature	annual max	global land
surface temperature	annual 5-day max	global land
precipitation	annual mean	global
precipitation	annual mean	global land
precipitation	annual max	global
precipitation	annual max	global land
precipitation	annual 5-day max	global land
drought duration (precipitation based)	annual max	global land
drought intensity (precipitation based)	annual max	global land
sea level pressure	hemispheric winter mean	latitudes 30-70 in each hemisphere
soil moisture (top ~50 cm of soil)	annual mean	global land
evapotranspiration	annual mean	global land
active layer thickness	Jun-Nov mean	latitudes 10N-90N
snow depth	annual mean	global land
sea ice extent	Jun-Nov mean	latitudes 50N-90N
ocean heat content (top ~400 m)	annual mean	global ocean
sea surface temperature	annual 5-day max	latitudes 55S-55N; ocean
ocean PH	annual mean	global ocean

References

Bach, S., *et al.,* “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PLoS One*, vol. 10, no. 7, e0130140, 2015.

Lundberg, S. M. and S. I. Lee, “A unified approach to interpreting model predictions,” *Proc. Adv. Neural Inf. Process. Syst.*, pp. 4768-4777, 2017.

Shapley, L.S. “A value for n-person games”. In: Contributions to the Theory of Games 2.28, pp. 307–317, 1953.

Shrikumar, A., *et al.,* “Not just a black box: Learning important features through propagating activation differences,” arXiv preprint, https://arxiv.org/abs/1605.01713, 2016.