Posted on 1 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/essoar.168565419.96137434/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

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

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

[1]Colorado State University

June 1, 2023

Abstract

Stratospheric aerosol injection (SAI) has been proposed as a possible complementary solution to limit global warming and its societal consequences. However, the climate impacts of such intervention remain unclear. Here, we introduce an explainable artificial intelligence (XAI) framework to quantify how distinguishable an SAI climate might be from a pre-deployment climate. A suite of neural networks is trained on Earth system model data to learn to distinguish between pre- and post-deployment periods across a variety of climate variables. The network accuracy is analogous to the "climate distinguishability" between the periods, and the corresponding distinctive patterns are identified using XAI methods to gain insights into the emerging signals from SAI. For many variables, the two periods are less distinguishable under SAI than under a no-SAI scenario, suggesting that the specific intervention modeled decelerates future climatic changes. Other climate variables for which the intervention has negligible effect are also highlighted.

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

by

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

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

Submitted to:

Geophysical Research Letters (AGU)

*email: amamalak@colostate.edu

Abstract

Stratospheric aerosol injection (SAI) has been proposed as a possible complementary solution to limit global warming and its societal consequences. However, the climate impacts of such intervention remain unclear. Here, we introduce an explainable artificial intelligence (XAI) framework to quantify how distinguishable an SAI climate might be from a pre-deployment climate. A suite of neural networks is trained on Earth system model data to learn to distinguish between pre- and post-deployment periods across a variety of climate variables. The network accuracy is analogous to the "climate distinguishability" between the periods, and the corresponding distinctive patterns are identified using XAI methods to gain insights into the emerging signals from SAI. For many variables, the two periods are less distinguishable under SAI than under a no-SAI scenario, suggesting that the specific intervention modeled decelerates future climatic changes. Other climate variables for which the intervention has negligible effect are also highlighted.

Keywords

Solar climate intervention, Stratospheric Aerosols Injection (SAI), eXplainable Artificial Intelligence (XAI), deep learning, climate distinguishability, climatic impacts.

Plain Language Summary

We use Earth system model predictions for two scenarios of the future: one policy-relevant climate change scenario where global temperatures continue rising in the coming decades, and that same scenario but with humans intervening in the climate system to limit warming to 1.5ºC. We then train a machine to learn to classify annual maps of climate variables based on whether they originate from the period before or after the intervention. The more successful the machine is at this task, the more distinguishable the pre- and post-intervention periods are with respect to the variable analyzed. Our results show that for many climate variables, the two periods are less distinguishable under the climate intervention scenario than the no-intervention scenario. In those cases, the intervention ends up decelerating future climate change. However, we also show that there are important climate variables for which the intervention has a negligible effect.

Key points

- An explainable artificial intelligence framework is introduced to quantify the "climate distinguishability" under a climate intervention scenario.
- The distinctive patterns between the pre- and post-intervention climates are not predefined but are learned directly from the data.
- For the Earth system model simulations analyzed, stratospheric aerosol injection is shown to decelerate future changes for some climate variables, while it shows a negligible effect for others.

1. Introduction

In order to limit the adverse impacts of global warming on weather, climate and society, various climate intervention strategies have been proposed as complementary to cutting CO_2 emissions. The two main categories of such strategies are greenhouse gas removal and solar climate intervention (Herzog, 2001; Vaughan and Lenton, 2011; National Research Council, 2015; National Academies of Sciences, Engineering and Medicine, NASEM 2021; Xu et al., 2020). Solar climate intervention consists of technologies that aim to increase the reflection of the incoming solar radiation and cool down the planet. A particularly popular strategy of solar climate intervention is stratospheric aerosol injection (SAI), which involves the deliberate injection of tiny particles (i.e., aerosols) into the stratosphere to reflect incoming solar radiation (Crutzen, 2006; Robock et al., 2009; Niemeier and Tilmes, 2017; MacMartin et al., 2017; Tilmes et al., 2018; 2020; Richter et al., 2022). The natural analog of SAI is large volcanic eruptions (e.g., the Mount Pinatubo eruption in 1991), during which, tiny particles are expelled into the atmosphere, resulting in a temporary (for a handful of years) cooling of the planet (Robock and Mao, 1995; Parker et al., 1996; Robock, 2000; Soden et al., 2002).

Although SAI has been shown to be a relatively inexpensive and effective strategy to limit global warming (Smith and Wagner, 2018; Tilmes et al., 2018; 2020; MacMartin et al., 2018), large uncertainties remain as to how such intervention would affect the climate system *beyond* the global mean temperature. For example, the degree to which the intervened Earth system would exhibit a similar climate to the pre-deployment system, whether ongoing/future climatic changes apart from global warming would be decelerated or halted, and the likelihood that SAI would introduce *new* adverse impacts are all questions of great interest (Jones et al., 2018; MacMartin et al., 2019; Kravitz and MacMartin, 2020; NASEM, 2021). Here, we propose an explainable artificial intelligence (XAI) framework to gain insights into these questions. We consider model simulations from the Community Earth System Model 2 under two future scenarios (spanning the

years 2015-2069): an intermediate climate change scenario where global temperatures continue rising, and an identical climate change scenario except where SAI is deployed to limit warming to 1.5ºC relative to the preindustrial era (Richter et al., 2022). We then focus on quantifying the "climate distinguishability" between the pre- and post-SAI worlds, by tasking an artificial neural network to distinguish between the two across a variety of climate variables. The more successful the network is at this task the more "distinguishable" the pre- and post-SAI worlds are in terms of their climate.

Specifically, to quantify the climate distinguishability after SAI, we train a neural network to distinguish between maps of a variable of interest that originate from the SAI climate (i.e., the SAI climate is defined as the 2040-2059 climate under the SAI scenario; see blue box in Figure 1a) vs maps that originate from the pre-deployment/reference climate (the reference climate is defined as the 2020-2039 climate under the intermediate climate change scenario; O'Neill et al., 2017; see gray box in Figure 1a). Although the prediction itself is not useful in this setting (i.e., we already know which map originates from which set of simulations), the accuracy of the network informs us about the climate distinguishability between the two periods for the variable analyzed. In this way, we quantify the degree of climate distinguishability with a single number: the accuracy of the network. To put this number into context, we compare the network accuracy with its "baseline" value, i.e., the network accuracy in the case where there was no intervention. That is, we repeat the above prediction task but this time the network is trained to distinguish between the reference climate and the future SSP climate with no intervention taking place (i.e., the future SSP climate is defined as the 2040-2059 climate under the intermediate climate change scenario; see magenta box in Figure 1a). The network's accuracy from this second task serves as a "baseline" value of climate distinguishability for the variable analyzed and is compared with the results from the first task to help assess the potential benefits (or risks) of deploying SAI.

We highlight that the main advantages of the proposed framework are that i) it provides a way to quantify with a single number the impact of an intervention on the reference climate, by assessing how much distinguishable the pre- and post-deployment climates would be, and ii) it is purely data-driven, thus, one does not need to predefine the form of change between the two compared climates. Instead, with our framework, we let the data tell us "the ways" that the two climates might be different. To gain insight into these distinctive patterns that make the two climates distinguishable, we use tools of explainable artificial intelligence (XAI). XAI tools aim to elucidate the decision-making process of deep learning models and have been increasingly applied in the geosciences in the recent years (see McGovern et al., 2019; Toms et al., 2020; Mamalakis et al., 2022a-c). Based on the climate simulations analyzed, SAI is shown to decelerate future changes for some of the variables, while showing negligible effect for others, highlighting the diversity in the potential effects of such climate interventions. In section 2, we provide details about the data, the prediction task of our framework and methods used, and in section 3 we present our results. Section 4 discusses our conclusions and future research directions.

2. Data and methodology

2.1. Data

We use data from an ensemble of Earth system model simulations: "Assessing Responses and Impacts of Solar climate intervention on the Earth system with Stratospheric Aerosol Injection" (ARISE-SAI; publicly available at https://www.cesm.ucar.edu/community-projects/arise-sai; Richter et al., 2022). The ARISE-SAI experiment consists of two sets of parallel simulations performed with the Community Earth System Model 2, using the Whole Atmosphere Community Climate Model version 6 as its atmospheric component (CESM2(WACCM6); Gettelman, et al., 2019; Danabasoglu, et al., 2020; Tilmes, et al., 2020; Richter et al., 2022): i) 10 ensemble members from 2015 to 2069 under the Shared Socioeconomic Pathway 2-4.5 (SSP2-4.5; O'Neill et al., 2017), which represents an intermediate climate change scenario; and ii) 10 ensemble members

from 2035 to 2069 under an SAI deployment scenario. In the latter, SO_2 is injected every day at roughly 21 km height at 180º longitude and 30ºS, 15ºS, 15ºN, and 30ºN using a "controller" algorithm (MacMartin et al., 2014; Kravitz et al., 2017). The SAI simulations aim to keep the global-mean surface air temperature near 1.5ºC above the preindustrial temperature. For more detailed information on the ARISE-SAI experiment, the reader is referred to Richter et al. (2022).

We quantify climate distinguishability for a list of 21 climate variables that are provided in Table S1. Prior to training the network, all variables are bi-linearly re-gridded to a 2.5° by 2.5° resolution from an approximate 1° by 1° resolution to reduce the dimensionality of the prediction task. Since this re-gridding is applied to the climate data of both scenarios, it is not expected to affect the conclusions about the impacts of SAI.

2.2. Prediction task

We define the CESM2(WACCM6) output over the period 2020-2039 under the SSP2-4.5 scenario as our reference climate, following the original study of ARISE-SAI (Richter et al., 2022). The reference climate represents the climatic conditions before a potential deployment of SAI. We then train a network to *distinguish* between the reference climate (see gray box in Figure 1a) and the climate under SAI over the period 2040-2059 (see blue box in Figure 1a). Specifically, given a randomly chosen map of a variable of interest as an input (e.g., a map of annual mean surface temperature or annual maximum precipitation, see Table S1), a fully connected network is tasked with estimating the probability that the map originated from the 2040-2059 SAI climate. A probability value less than 0.5 indicates that the map is predicted to belong to the reference climate, while a probability value greater than 0.5 indicates that the map is predicted to belong to the SAI climate; see Figure 1b. Framing the prediction task in this way requires the network to identify patterns that serve as robust and distinctive indicators to separate the pre- and post-deployment periods. The more successful the network is at this task, the more the two periods are "climatically distinguishable" under the SAI scenario. In contrast when the network is not successful (e.g., if it

159 performs similarly to a random chance-based model), the climatic conditions between the two
160 periods are deemed indistinguishable with respect to the variable analyzed and based on the
161 network used. We highlight here that the patterns used by the network could be of any form: local,
162 global or any type of combination of patterns, pointing out to the generic nature of the suggested
163 framework.

164 To place climate distinguishability under SAI into context, we compare it to the climate
165 distinguishability under the scenario of no intervention. We do this by we repeating the same
166 approach, but by tasking the network to distinguish between the reference climate and the climate
167 in the period 2040-2059 under the SSP2-4.5 scenario (see magenta box in Figure 1a). The
168 comparison between the climate distinguishability with and without SAI gives insights into the
169 potential of SAI to counter the impacts of climate change. For instance, in the specific case of the
170 ARISE-SAI simulations, it may be concluded that SAI reduces future climate change if the degree
171 of climate distinguishability is significantly lower under the SAI scenario than under the SSP
172 scenario. For details on the training approach and the architectures of the networks, please see
173 Supplementary Text S1.

174 **2.3. Explainable AI method**

175 We use the local attribution method Deep SHAP (Lundberg and Lee, 2017) to explain the
176 predictions of the network. We have chosen this method for two reasons: 1) it allows the user to
177 define the baseline for which the attribution is derived (see Mamalakis et al., (2023) on the
178 importance of baselines); and 2) it satisfies the *completeness* property (Sundararajan et al., 2017),
179 which holds that the attributions add up to the difference between the network output at the current
180 sample and the one at the baseline. For further details on the Deep SHAP algorithm, please see
181 Supplementary Text S2. We note that we have also used the method Integrated Gradients
182 (Sundararajan et al., 2017) to explain the network's predictions, and the results were very similar
183 to those based on Deep SHAP (not shown).

3. Results

We start by presenting the results for the case of annual maximum daily precipitation in Figure 2. We first discuss the results for a future climate with no intervention. The global-mean annual maximum precipitation exhibits an increase throughout the century but with large ensemble spread (magenta lines, Figure 2a). The largest increases occur in the deep tropics, specifically over the tropical Pacific (Figure 2b; see also O'Gorman and Schneider, 2009; Kharin et al., 2013; Pfahl et al., 2017). The network can successfully distinguish between the reference climate and the SSP future climate 85% of the time, which is significant at a 0.01 level (Figure 2d). Moreover, the probability assigned by the network that a map corresponds to the future SSP climate increases linearly with the actual year of the map and maximizes in the out-of-sample years 2060-2069 (Figure 2d). This suggests that there are robust signals of climate change that become more and more evident with time. It also suggests that the learned patterns generalize successfully, since the network is able to correctly classify the years 2060-2069, although those years were not used during training (see Supplementary Text S1). Based on the results from the XAI method Deep SHAP, the network mainly uses precipitation extremes over the tropical eastern Pacific (and to a lesser degree over the Southern Ocean and the tropical Atlantic) to make its predictions (Figure 2f). Interestingly, the network does not use precipitation over the western Pacific or Australia, despite the fact that the corresponding ensemble mean difference between the two periods is of high magnitude (Figure 2b). This implies high internal variability of precipitation extremes over these regions, which does not make them robust indicators from a signal-to-noise perspective.

Under the SAI scenario, the overall accuracy of the network is only 58% (Figure 2e), which is not statistically different from a random chance-based model (at a 0.01 significance level, a random chance-based model would perform with up to 69% accuracy, derived using a binomial distribution). The network-estimated probability that a map corresponds to the SAI climate is almost independent from the year of the map (Figure 2e), which indicates that there are no robust

long-term climate signals under SAI that the network could use for distinguishing from the reference climate. This is also suggested by the XAI results; note the incoherent and noisy attributions in Figure 2g. Generally, the results in Figure 2 indicate that although the CESM2(WACCM6) simulates a robust increase in future extreme daily precipitation under the SSP2-4.5 scenario, possible deployment of SAI could preserve the conditions of the reference (i.e., pre-deployment) climate. This could be an example of a potential positive SAI impact.

Next, we consider the annual mean surface temperature over land (Figure 3). Under the SSP scenario, a clear increase in surface temperature is shown throughout the century that is evident globally (Figure 3a-b). Accordingly, the network accuracy in distinguishing between the reference and the future SSP climate is high, on the order of 93%. Many regions around the globe are highlighted by Deep SHAP as robust distinctive patterns; e.g., Mexico, southern South America, southern Africa, Indonesia, and southern Australia. Under the SAI scenario, although the global mean temperature is similar to the one under the reference climate, there are robust patterns of regional cooling that make the two climates *highly* distinguishable: 91% of the time (Figure 3e). Regional cooling happens mainly over southern South America, eastern Africa, eastern Australia, and Greenland (Figure 3c). These are the regions that the network uses to distinguish between the reference and the SAI climates (see Figure 3g). Overall, these results indicate that the CESM2(WACCM6) projects that a potential SAI deployment would lead to a less warm climate than SSP; however, the annual mean surface temperature over land in an SAI world would also be distinguishable from the reference climate. Importantly, the distinctive patterns in the two scenarios are quite different, with warming being the distinctive difference under the SSP scenario, while regional cooling patterns being the most robust distinctive patterns under SAI.

We have repeated the same analysis as in Figures 2-3 for a list of 21 variables (see Table S1), and we summarize the results in Figure 4. For all variables, the network accuracy under the SSP scenario (magenta circles in Figure 4a) is statistically significant. This means that even under

234 the intermediate climate change scenario SSP2-4.5, the CESM2(WACCM6) projects that the Earth
235 system would exhibit climatic conditions that are distinguishable from the reference climate in the
236 coming decades. However, for the majority of variables examined here, SAI would lead to a less
237 distinguishable climate than the SSP scenario, although (with a few exceptions) one that would
238 also be distinguishable from the reference climate (note that the network accuracy (light blue
239 circles) is higher than the random chance-based accuracy). In particular, SAI would decelerate
240 many future greenhouse-gas driven climate changes, especially for surface temperature extremes,
241 precipitation, drought occurrence, sea level pressure, and Arctic sea ice (see also Xu et al., 2020;
242 Tye et al., 2022; Lee et al., 2020; 2023). It is important to note, however, that there are variables
243 for which SAI is projected to have minimal impact relative to climate change. Examples include
244 soil moisture, evapotranspiration, and ocean acidity.

245 We next explore how distinctive patterns might be modified from SAI; note that the
246 network accuracy alone does not provide this information. For example, as is shown in Figure 3,
247 the climate distinguishability under the SSP and the SAI scenarios is very similar, but the
248 corresponding distinctive patterns are different. To explore this further, the spatial correlation
249 between the XAI heatmaps under the SSP and SAI scenarios are presented in Figure 4b. In most
250 cases, the correlation is not statistically different from zero, which means that SAI is projected to
251 introduce different distinctive patterns relative to those from the SSP scenario. Exceptions are for
252 cases where the correlation is high, such as for ocean acidity and ocean heat, which means that the
253 anticipated SSP-driven distinctive patterns are projected to remain almost unchanged under SAI.

254 The results in Figure 4 indicate the diverse impacts of SAI on different components of the
255 climate system, which highlights the need for systematic and thorough investigations into the
256 possible impacts of SAI on the Earth system beyond only the global-mean temperature response.
257 Such research is needed for a well-informed policy making regarding potential deployment of
258 climate intervention approaches (NASEM, 2021). The framework introduced here allows for such

data-driven and generic investigations to uncover the ways in which an SAI climate would be different from a pre-deployment one.

4. Conclusions

In this study, a new framework was used that allows quantification (with a single number) of the degree of climate distinguishability between a reference climate and future climate states from both SAI and no-SAI worlds. The framework is based on the use of machine learning and leverages XAI tools to identify robust distinctive patterns under the intervention and the no-intervention scenarios. The framework is purely data driven, nonlinear, nonlocal, and it accounts for underlying uncertainties in the data that may originate from internal stochastic variability or uncertainties in Earth system model physics.

We applied this framework to data from ensembles of simulations that were developed to examine the potential impacts of stratospheric aerosol injection; namely, the ARISE-SAI project (Richter et al., 2022). In these simulations, SAI was shown to have diverse impacts on the simulated climate. These include minimizing changes due to greenhouse gas forcing in temperature and precipitation extremes, while having negligible effect on ocean acidification. Also, for the majority of variables examined here, the simulated deployment of SAI led to new patterns of change with respect to the reference climate that were different from the SSP patterns. This raises the possibility of SAI leading to *new* (and perhaps unwanted) changes in specific components of the Earth system or in certain regions of the world.

We do note some potential limitations of the presented framework. One is the dependence of the results on the amount of data. Neural networks are known to be "data-thirsty" models (LeCun et al., 2015), so it is possible that certain patterns that were not identified as robust indicators during training could become robust with more data. However, the dependence on the amount of data is present in virtually all climate settings involving questions of signal-to-noise and statistical significance. Another limitation is the possible dependence of the results on the network

architecture. In order to address this issue here, we searched over many different architectures and combinations of hyperparameters before training the network, as described in Supplementary Text S1. That way, we let the data guide us as to what architecture we should use for each climate variable. Yet, we acknowledge that it is possible that some of these results depend on the adopted architectures.

Our work highlights the need to further research the impacts of possible intervention approaches *beyond* just global mean temperatures, as has been done in other studies, examining ARISE-SAI data in particular (Keys et al., 2022; Labe et al., 2023; Hueholt et al., 2023). In doing so, we envision that the notion of "quantifiable climate distinguishability" will be a relevant and informative metric to assess impacts and to expand the design space of possible interventions (Lee et al., 2020), as illustrated by the presented results. Further investigation could include further assessing the climate distinguishability by considering multiple variables at the same time (i.e., the network input consists of many channels each of which refers to a different variable), to assess potential impacts on the dependence structure of different components of the Earth system and the occurrence of compound events. Future work could also focus on analyzing the output of more than one model and of more than one climate intervention strategy to establish a more holistic picture of the potential impacts of proposed climate intervention strategies.

Acknowledgments

This work was supported by Defense Advanced Research Projects Agency (DARPA) Grant No. HR00112290071. The views expressed here do not necessarily reflect the positions of the U.S. government. The authors would also like to thank the efforts of the ARISE-SAI team for making their data publicly available.

Data availability

The ARISE-SAI data is publicly available at https://www.cesm.ucar.edu/community-projects/arise-sai. The code to reproduce the presented results is publicly available at https://github.com/amamalak/Quantify_SAI_impacts.

References

Crutzen, P.J. (2006) Albedo Enhancement by Stratospheric Sulfur Injections: A contribution to Resolve a Policy Dilemma? *Clim. Change*, **77**(3-4), 211-220, doi:10.1007/s10584-006-9101-y.

Danabasoglu, G., et al. (2020) The Community Earth System Model Version 2 (CESM2), *J. Adv. Model. Earth Sy.*, **12**, e2019MS001916, https://doi.org/10.1029/2019MS001916.

Gettelman, A., et al. (2019) The whole atmosphere community climate model version 6 (WACCM6), *J. Geophys. Res.-Atmos*., **124**, 12380–12403, https://doi.org/10.1029/2019JD030943.

Herzog, H.J., (2001) What Future for Carbon Capture and Sequestreation?, *Environ. Sci. Technol*., **35**(7), 148A-153A.

Hueholt, D.M., E.A. Barnes, J.W. Hurrell, J.H. Richter, and L. Sun (2023) Assessing Outcomes in Stratospheric Aerosol Injection scenarios shortly after deployment, authorea preprints.

Jones, A.C., M.K. Hawcroft, J.M. Haywood,A. Jones, X. Guo, and J.C. Moore (2018) Regional climate impacts of stabilizing global warming at 1.5 K using solar geoengineering, *Earth's Future*, **6**, 230-251.

Keys, P.W., E.A. Barnes, N.S. Diffenbaugh, J.W. Hurrell, and M.B. Curtis (2022) Potential for perceived failure of Stratospheric Aerosol Injection deployment, *PNAS*, **119**(40), e2210036119.

Kharin, V.V., F.W. Zwiers, X. Zhang, et al. (2013) Changes in temperature and precipitation extremes in the CMIP5 ensemble, *Climatic Change*, **119**, 345-357.

Kravitz, B., et al. (2017) First simulations of designing stratospheric sulfate aerosol geoengineering to meet multiple simultaneous climate objectives, *J. Geophys. Res.-Atmos*., **122**, 12616-12634.

342 Kravitz, B., and D.G. MacMartin (2020) Uncertainty and the basis for confidence in solar
343 geoengineering research, *Nat Rev Earth Environ*, **1**, 64-75.
344 Labe, Z.M., E.A. Barnes, and J.W. Hurrell (2023) Identifying the regional emergence of climate
345 patterns in the ARISE-SAI-1.5 simulations, *Environmental Research Letters*,
346 https://doi.org/10.1088/1748-9326/acc81a.
347 LeCun, Y., Y. Bengio, and G. Hinton (2015) Deep learning, *Nature*, **521**, 436-444,
348 https://doi.org/10.1038/nature14539.
349 Lee, W., D. MacMartin, D. Visioni, and B. Kravitz (2020) Expanding the design space of
350 stratospheric aerosol geoengineering to include precipitation-based objectives and explore
351 trade-offs, *Earth Syst. Dynam*., **11**(4), 1051-1072.
352 Lee, W.R., et al., (2023) High-latitude stratospheric aerosol injection to preserve the Arctic,
353 *Earth's Future*, **11**(1), e2022EF003052.
354 Lundberg, S. M. and S. I. Lee (2017) A unified approach to interpreting model predictions," *Proc.*
355 *Adv. Neural Inf. Process. Syst.*, pp. 4768-4777.
356 MacMartin, D.G., B. Kravitz, D.W. Keith, and A. Jarvis (2014) Dynamics of the coupled human-
357 climate system resulting from closed-loop control of solar geoengineering, *Clim. Dynam*., **43**,
358 243-258.
359 MacMartin, D.G., et al. (2017) The Climate Response to Stratospheric Aerosol Geoengineering
360 Can Be Tailored Using Multiple Injection Locations, *J. Geophys. Res. Atmos*., **122**(23),
361 12,574-12,590, doi:10.1002/2017JD026868.
362 MacMartin, D.G., K.L. Ricke, and D.W. Keith (2018) Solar geoengineering as part of an overall
363 strategy for meeting 1.5∘C Paris target, *Philosophical Transactions of the Royal Society A*,
364 **376**, 20160454.

MacMartin, D.G., W. Wang, B. Kravitz, S. Tilmes, J.H. Richter, and M.J. Mills (2019) Timescale for detecting the climate response to stratospheric aerosol geoengineering. *Journal of Geophysical Research: Atmospheres*, **124**, 1233-1247.

Mamalakis, A., I. Ebert-Uphoff, E.A. Barnes (2022a) Explainable Artificial Intelligence in Meteorology and Climate Science: Model fine-tuning, calibrating trust and learning new science, in *Beyond explainable Artificial Intelligence* by Holzinger et al. (Editors), Springer Lecture Notes on Artificial Intelligence (LNAI).

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

Mamalakis, A, E.A. Barnes and I. Ebert-Uphoff (2022c) Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience, *Artificial Intelligence for the Earth Systems*, **1**(4), e220012.

Mamalakis, A, E.A. Barnes and I. Ebert-Uphoff (2023) Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience, *Artificial Intelligence for the Earth Systems*, **2**(1), e220058.

McGovern, A., *et al.,* (2019) Making the black box more transparent: Understanding the physical implications of machine learning," *Bulletin of the American Meteorological Society*, vol. **100**, no. 11, pp. 2175-2199.

National Academies pf Sciences, Engineering, and Medicine (2021) Reflecting Sunlight: Recommendations for Solar Geoengineering Research and Research Governance. Washington, DC: The National Academies Press. https://doi.org/10.17226/25762.

National Research Council (2015) Climate Intervention: Carbon Dioxide Removal and Reliable Sequestration. Washington, CD: The National Academies Press. https://doi.org/10.17226/18805.

Niemeier, U., and S. Tilmes (2017) Sulfur injections for a cooler planet, *Science*, **357**(6348), 246-248, doi:10.1126/science.aan3317.

O'Gormanm, O.A., and T. Schneider (2009) The physical basis for increases in precipitation extremes in simulations of 21st-century climate change, *PNAS*, **106**(35), 14773-14777.

O'Neill, et al. (2017) The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century, *Global Environ. Change*, **42**, 169-180.

Parker, D.E., H. Wilson, P.D. Jones, J.R. Christy, C.K. Folland (1996) The impact of mount Pinatubo on world-wide temperatures, *Int. J. Climatol*., **16**, 487-497.

Pfahl, S., P. O'Gorman, and E. Fischer (2017) Understanding the regional pattern of projected future changes in extreme precipitation, *Nature Climate Change*, **7**, 423-427.

Richter, J.H., et al. (2022) Assessing Responses and Impacts of Solar climate Intervention on the Earth system with stratospheric aerosol injection (ARISE-SAI): protocol and initial results from the first simulations, *Geosci. Model Dev*., **15**, 8221-8243.

Robock, A. (2000) Volcanic eruptions and climate, *Rev. Geophys*., **38**(2), 191-219, doi:10.1029/1998RG000054.

Robock, A., and J. Mao (1995) The Volcanic Signal in Surface Temperature Observations, *J. Climate*, **8**(5), 1086-1103.

Robock, A., A. Marquardt, B. Kravitz, and G. Stenchikov (2009) Benefits, risks, and costs of stratospheric geoengineering, *Geophys. Res. Lett.*, **36**(19), L19703, doi:10.1029/2009GL039209.

Smith, W., and G. Wagner (2018) Stratospheric aerosol injection tactics and costs in the first 15 years of deployment, *Environ. Res. Lett.*, **13**, 124001.

Soden, B.J., et al. (2002) Global cooling after the eruption of mount Pinatubo: A test of climate feedback by water vapor, *Science*, **296**, 727-730, doi:10.1126/science.296.5568.727.

Sundararajan, M., A. Taly, Q. Yan, (2017) Axiomatic attribution for deep networks," arXiv preprint, https://arxiv.org/abs/1703.01365.

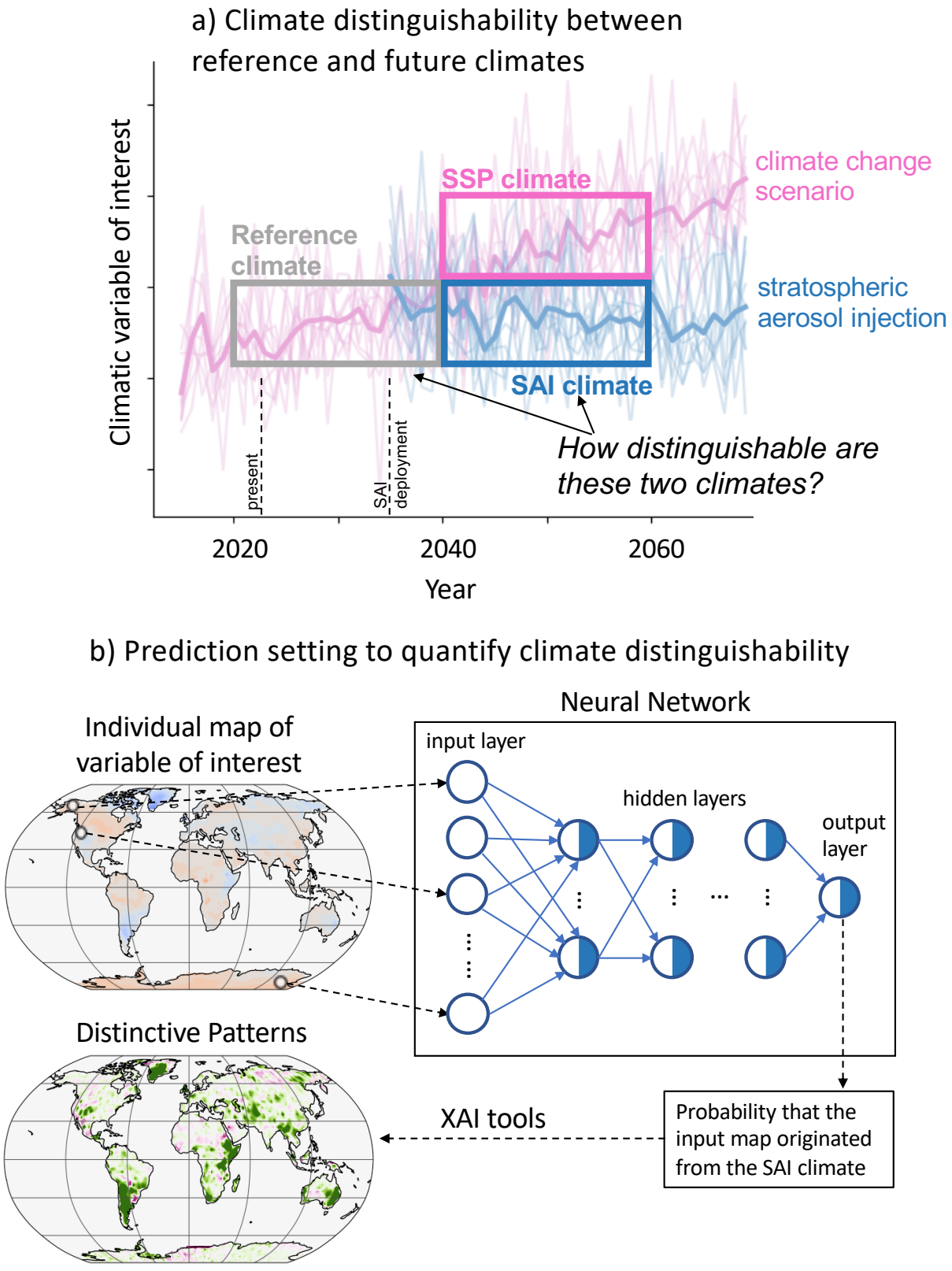
Tilmes, S., et al. (2018) CESM1(WACCM) Stratospheric Aerosol Geoengineering Large Ensemble Project, *Bull. Am. Meteorol. Soc*., **99**(11), 2361-2371, doi:10.1175/BAMS-D-17-0267.1.

Tilmes, S., et al. (2020) Reaching 1.5 and 2.0 ∘C global surface temperature targets using stratospheric aerosol geoengineering, *Earth Syst. Dynam*., **11**, 579–601, https://doi.org/10.5194/esd-11-579-2020.

Toms, B.A., E. A. Barnes, I. Ebert-Uphoff, "Physically interpretable neural networks for the geosciences: Applications to Earth system variability," *Journal of Advances in Modeling Earth Systems*, vol. **12**, e2019MS002002, 2020.

Tye, M.R., K. Dagon, M.J. Molina, J.H. Richter, D. Visioni, B. Kravitz, and S. Tilmes (2022) Indices of extremes: geographic patterns of change in extremes and associated vegetation impacts under climate intervention, *Earth Syst. Dynam.*, **13**, 1233-1257.

Vaughan, N.E. and T.M. Lenton (2011) A review of climate geoengineering proposals, *Clim. Change*, **109**(3-4), 745-790.



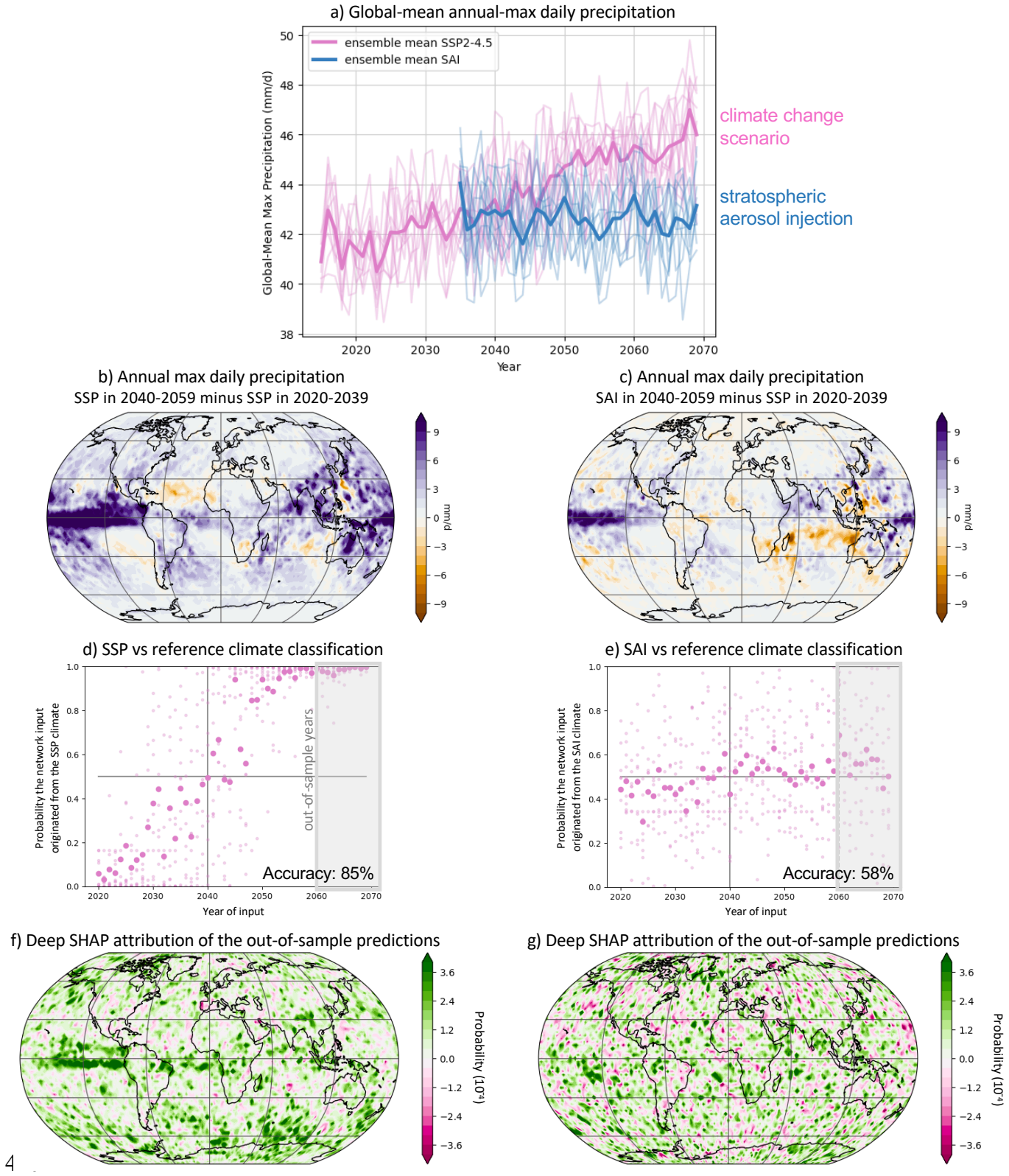
433

434 **Figure 1:** Schematic of our framework to quantify SAI impacts using XAI. a) Assessing climate

435 distinguishability between reference and future climates. Note that the pre-2040 period under an

intermediate climate change scenario is used as the refence climate, in accordance to Richer et al (2022). b) Schematic of the prediction task to quantify climate distinguishability after SAI and the use of XAI to derive the distinctive patterns between the reference and SAI climates.

439

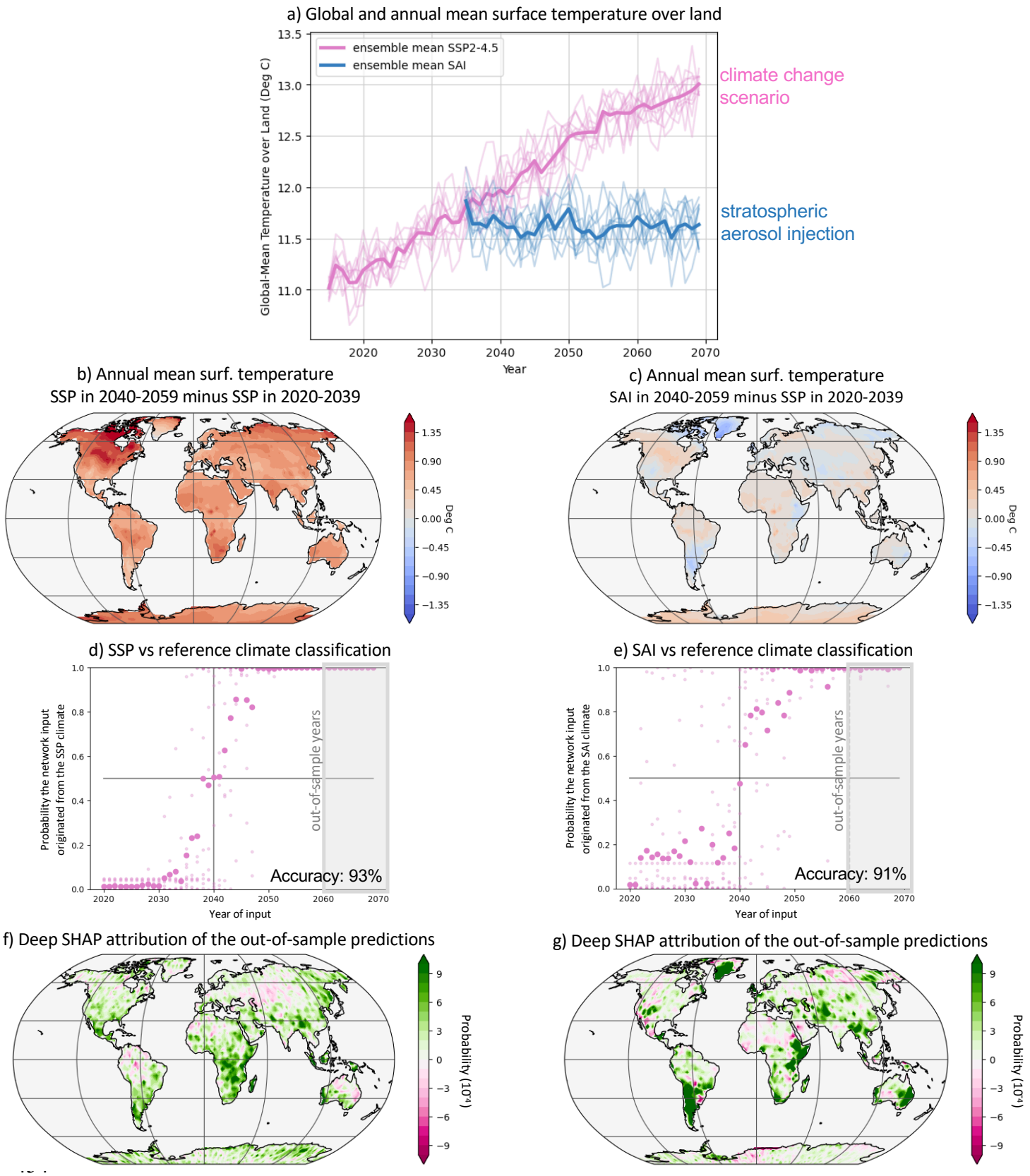


441 **Figure 2.** Results of our framework for annual maximum daily precipitation. a) Series of global-mean

442 annual maximum precipitation (in mm/d) under the SSP2-4.5 scenario and the ARISE-SAI scenario. All

10 ensemble members and the ensemble mean are shown. b) Ensemble mean difference between the annual maximum precipitation in the 2040-2059 SSP2-4.5 climate and the reference climate. d) Network-generated probability that different annual maximum precipitation maps originated from the 2040-2059 SSP2-4.5 climate. The actual year of each map is provided in the horizontal axis. The overall accuracy of the network is shown on the bottom right corner. f) Distinctive patterns that were used by the network to separate the reference climate from the 2040-2059 SSP2-4.5 climate, as estimated using the method Deep SHAP. The presented attributions correspond to the average attributions across the 2060-2069 network predictions and all testing members, using the years 2035-2044 as baseline. c,e,g) Same as (b,d,f), but the network is trained to separate the reference climate from the 2040-2059 ARISE-SAI climate.

453



455 **Figure 3.** Same as in Figure 2, but results are for the annual mean surface temperature over land.

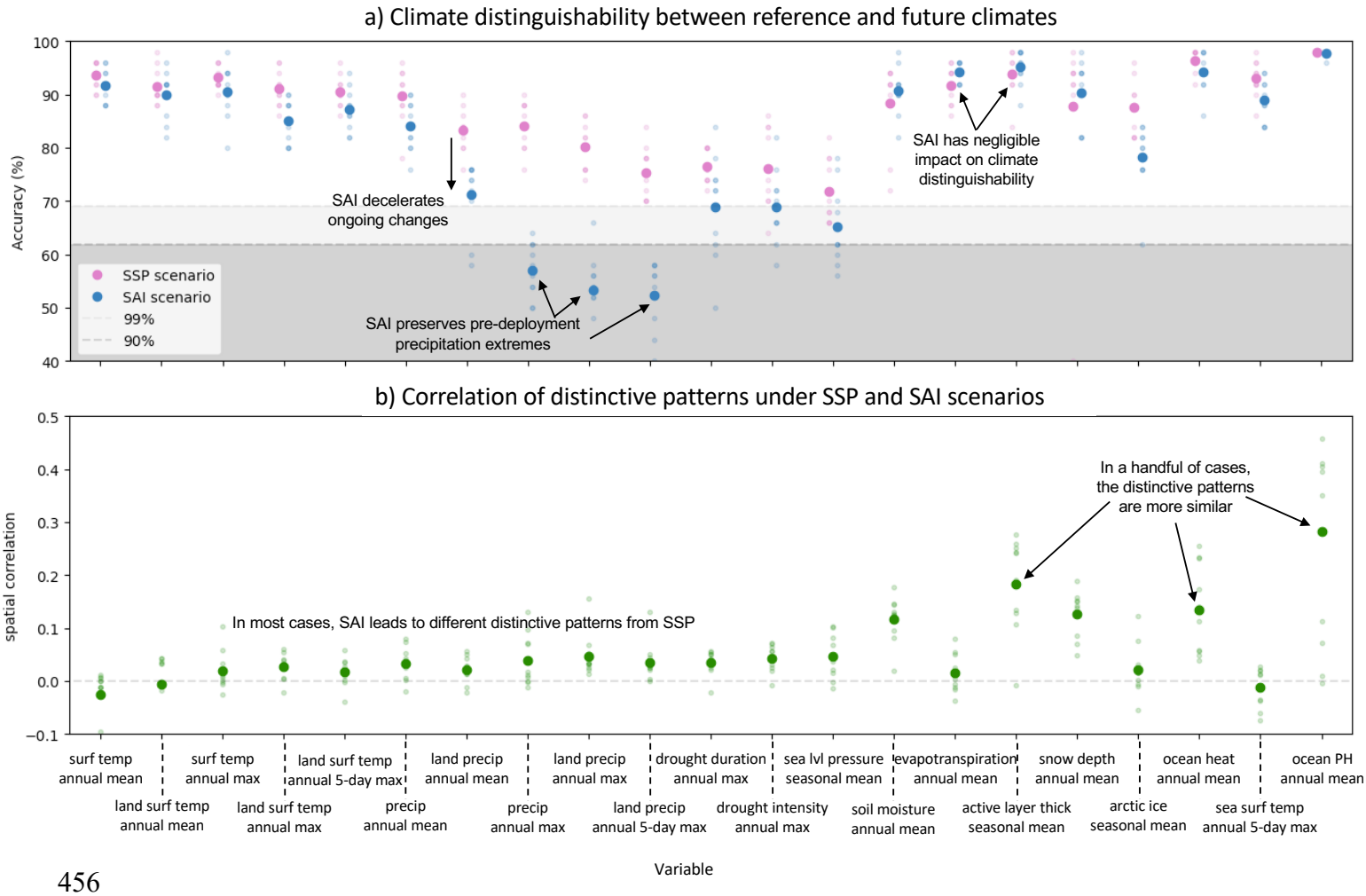


Figure 4. a) Accuracy of the network in distinguishing between the reference climate and the future SSP 2-4.5 climate (magenta) or the future ARISE-SAI climate (light blue), for all variables considered in the study (see Supplementary Table S1). Results from individual testing members (smaller circles) and the ensemble mean (bigger circles) are presented. The critical values for the 10% and 1% significance levels are derived using a binomial distribution. b) Correlation coefficient between attribution heatmaps that correspond to predicting in the two scenarios. Results from individual testing members (smaller circles) and the ensemble mean (bigger circles) are presented.

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

by

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

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

Submitted to:

Geophysical Research Letters (AGU)

*email: amamalak@colostate.edu

Abstract

Stratospheric aerosol injection (SAI) has been proposed as a possible complementary solution to limit global warming and its societal consequences. However, the climate impacts of such intervention remain unclear. Here, we introduce an explainable artificial intelligence (XAI) framework to quantify how distinguishable an SAI climate might be from a pre-deployment climate. A suite of neural networks is trained on Earth system model data to learn to distinguish between pre- and post-deployment periods across a variety of climate variables. The network accuracy is analogous to the "climate distinguishability" between the periods, and the corresponding distinctive patterns are identified using XAI methods to gain insights into the emerging signals from SAI. For many variables, the two periods are less distinguishable under SAI than under a no-SAI scenario, suggesting that the specific intervention modeled decelerates future climatic changes. Other climate variables for which the intervention has negligible effect are also highlighted.

Keywords

Solar climate intervention, Stratospheric Aerosols Injection (SAI), eXplainable Artificial Intelligence (XAI), deep learning, climate distinguishability, climatic impacts.

Plain Language Summary

We use Earth system model predictions for two scenarios of the future: one policy-relevant climate change scenario where global temperatures continue rising in the coming decades, and that same scenario but with humans intervening in the climate system to limit warming to 1.5ºC. We then train a machine to learn to classify annual maps of climate variables based on whether they originate from the period before or after the intervention. The more successful the machine is at this task, the more distinguishable the pre- and post-intervention periods are with respect to the variable analyzed. Our results show that for many climate variables, the two periods are less distinguishable under the climate intervention scenario than the no-intervention scenario. In those cases, the intervention ends up decelerating future climate change. However, we also show that there are important climate variables for which the intervention has a negligible effect.

Key points

- An explainable artificial intelligence framework is introduced to quantify the "climate distinguishability" under a climate intervention scenario.
- The distinctive patterns between the pre- and post-intervention climates are not predefined but are learned directly from the data.
- For the Earth system model simulations analyzed, stratospheric aerosol injection is shown to decelerate future changes for some climate variables, while it shows a negligible effect for others.

1. Introduction

In order to limit the adverse impacts of global warming on weather, climate and society, various climate intervention strategies have been proposed as complementary to cutting CO_2 emissions. The two main categories of such strategies are greenhouse gas removal and solar climate intervention (Herzog, 2001; Vaughan and Lenton, 2011; National Research Council, 2015; National Academies of Sciences, Engineering and Medicine, NASEM 2021; Xu et al., 2020). Solar climate intervention consists of technologies that aim to increase the reflection of the incoming solar radiation and cool down the planet. A particularly popular strategy of solar climate intervention is stratospheric aerosol injection (SAI), which involves the deliberate injection of tiny particles (i.e., aerosols) into the stratosphere to reflect incoming solar radiation (Crutzen, 2006; Robock et al., 2009; Niemeier and Tilmes, 2017; MacMartin et al., 2017; Tilmes et al., 2018; 2020; Richter et al., 2022). The natural analog of SAI is large volcanic eruptions (e.g., the Mount Pinatubo eruption in 1991), during which, tiny particles are expelled into the atmosphere, resulting in a temporary (for a handful of years) cooling of the planet (Robock and Mao, 1995; Parker et al., 1996; Robock, 2000; Soden et al., 2002).

Although SAI has been shown to be a relatively inexpensive and effective strategy to limit global warming (Smith and Wagner, 2018; Tilmes et al., 2018; 2020; MacMartin et al., 2018), large uncertainties remain as to how such intervention would affect the climate system *beyond* the global mean temperature. For example, the degree to which the intervened Earth system would exhibit a similar climate to the pre-deployment system, whether ongoing/future climatic changes apart from global warming would be decelerated or halted, and the likelihood that SAI would introduce *new* adverse impacts are all questions of great interest (Jones et al., 2018; MacMartin et al., 2019; Kravitz and MacMartin, 2020; NASEM, 2021). Here, we propose an explainable artificial intelligence (XAI) framework to gain insights into these questions. We consider model simulations from the Community Earth System Model 2 under two future scenarios (spanning the

years 2015-2069): an intermediate climate change scenario where global temperatures continue rising, and an identical climate change scenario except where SAI is deployed to limit warming to 1.5ºC relative to the preindustrial era (Richter et al., 2022). We then focus on quantifying the "climate distinguishability" between the pre- and post-SAI worlds, by tasking an artificial neural network to distinguish between the two across a variety of climate variables. The more successful the network is at this task the more "distinguishable" the pre- and post-SAI worlds are in terms of their climate.

Specifically, to quantify the climate distinguishability after SAI, we train a neural network to distinguish between maps of a variable of interest that originate from the SAI climate (i.e., the SAI climate is defined as the 2040-2059 climate under the SAI scenario; see blue box in Figure 1a) vs maps that originate from the pre-deployment/reference climate (the reference climate is defined as the 2020-2039 climate under the intermediate climate change scenario; O'Neill et al., 2017; see gray box in Figure 1a). Although the prediction itself is not useful in this setting (i.e., we already know which map originates from which set of simulations), the accuracy of the network informs us about the climate distinguishability between the two periods for the variable analyzed. In this way, we quantify the degree of climate distinguishability with a single number: the accuracy of the network. To put this number into context, we compare the network accuracy with its "baseline" value, i.e., the network accuracy in the case where there was no intervention. That is, we repeat the above prediction task but this time the network is trained to distinguish between the reference climate and the future SSP climate with no intervention taking place (i.e., the future SSP climate is defined as the 2040-2059 climate under the intermediate climate change scenario; see magenta box in Figure 1a). The network's accuracy from this second task serves as a "baseline" value of climate distinguishability for the variable analyzed and is compared with the results from the first task to help assess the potential benefits (or risks) of deploying SAI.

We highlight that the main advantages of the proposed framework are that i) it provides a way to quantify with a single number the impact of an intervention on the reference climate, by assessing how much distinguishable the pre- and post-deployment climates would be, and ii) it is purely data-driven, thus, one does not need to predefine the form of change between the two compared climates. Instead, with our framework, we let the data tell us "the ways" that the two climates might be different. To gain insight into these distinctive patterns that make the two climates distinguishable, we use tools of explainable artificial intelligence (XAI). XAI tools aim to elucidate the decision-making process of deep learning models and have been increasingly applied in the geosciences in the recent years (see McGovern et al., 2019; Toms et al., 2020; Mamalakis et al., 2022a-c). Based on the climate simulations analyzed, SAI is shown to decelerate future changes for some of the variables, while showing negligible effect for others, highlighting the diversity in the potential effects of such climate interventions. In section 2, we provide details about the data, the prediction task of our framework and methods used, and in section 3 we present our results. Section 4 discusses our conclusions and future research directions.

2. Data and methodology

2.1. Data

We use data from an ensemble of Earth system model simulations: "Assessing Responses and Impacts of Solar climate intervention on the Earth system with Stratospheric Aerosol Injection" (ARISE-SAI; publicly available at https://www.cesm.ucar.edu/community-projects/arise-sai; Richter et al., 2022). The ARISE-SAI experiment consists of two sets of parallel simulations performed with the Community Earth System Model 2, using the Whole Atmosphere Community Climate Model version 6 as its atmospheric component (CESM2(WACCM6); Gettelman, et al., 2019; Danabasoglu, et al., 2020; Tilmes, et al., 2020; Richter et al., 2022): i) 10 ensemble members from 2015 to 2069 under the Shared Socioeconomic Pathway 2-4.5 (SSP2-4.5; O'Neill et al., 2017), which represents an intermediate climate change scenario; and ii) 10 ensemble members

from 2035 to 2069 under an SAI deployment scenario. In the latter, SO_2 is injected every day at roughly 21 km height at 180º longitude and 30ºS, 15ºS, 15ºN, and 30ºN using a "controller" algorithm (MacMartin et al., 2014; Kravitz et al., 2017). The SAI simulations aim to keep the global-mean surface air temperature near 1.5ºC above the preindustrial temperature. For more detailed information on the ARISE-SAI experiment, the reader is referred to Richter et al. (2022).

We quantify climate distinguishability for a list of 21 climate variables that are provided in Table S1. Prior to training the network, all variables are bi-linearly re-gridded to a 2.5° by 2.5° resolution from an approximate 1° by 1° resolution to reduce the dimensionality of the prediction task. Since this re-gridding is applied to the climate data of both scenarios, it is not expected to affect the conclusions about the impacts of SAI.

2.2. Prediction task

We define the CESM2(WACCM6) output over the period 2020-2039 under the SSP2-4.5 scenario as our reference climate, following the original study of ARISE-SAI (Richter et al., 2022). The reference climate represents the climatic conditions before a potential deployment of SAI. We then train a network to *distinguish* between the reference climate (see gray box in Figure 1a) and the climate under SAI over the period 2040-2059 (see blue box in Figure 1a). Specifically, given a randomly chosen map of a variable of interest as an input (e.g., a map of annual mean surface temperature or annual maximum precipitation, see Table S1), a fully connected network is tasked with estimating the probability that the map originated from the 2040-2059 SAI climate. A probability value less than 0.5 indicates that the map is predicted to belong to the reference climate, while a probability value greater than 0.5 indicates that the map is predicted to belong to the SAI climate; see Figure 1b. Framing the prediction task in this way requires the network to identify patterns that serve as robust and distinctive indicators to separate the pre- and post-deployment periods. The more successful the network is at this task, the more the two periods are "climatically distinguishable" under the SAI scenario. In contrast when the network is not successful (e.g., if it

159 performs similarly to a random chance-based model), the climatic conditions between the two
160 periods are deemed indistinguishable with respect to the variable analyzed and based on the
161 network used. We highlight here that the patterns used by the network could be of any form: local,
162 global or any type of combination of patterns, pointing out to the generic nature of the suggested
163 framework.

164 To place climate distinguishability under SAI into context, we compare it to the climate
165 distinguishability under the scenario of no intervention. We do this by we repeating the same
166 approach, but by tasking the network to distinguish between the reference climate and the climate
167 in the period 2040-2059 under the SSP2-4.5 scenario (see magenta box in Figure 1a). The
168 comparison between the climate distinguishability with and without SAI gives insights into the
169 potential of SAI to counter the impacts of climate change. For instance, in the specific case of the
170 ARISE-SAI simulations, it may be concluded that SAI reduces future climate change if the degree
171 of climate distinguishability is significantly lower under the SAI scenario than under the SSP
172 scenario. For details on the training approach and the architectures of the networks, please see
173 Supplementary Text S1.

174 **2.3. Explainable AI method**

175 We use the local attribution method Deep SHAP (Lundberg and Lee, 2017) to explain the
176 predictions of the network. We have chosen this method for two reasons: 1) it allows the user to
177 define the baseline for which the attribution is derived (see Mamalakis et al., (2023) on the
178 importance of baselines); and 2) it satisfies the *completeness* property (Sundararajan et al., 2017),
179 which holds that the attributions add up to the difference between the network output at the current
180 sample and the one at the baseline. For further details on the Deep SHAP algorithm, please see
181 Supplementary Text S2. We note that we have also used the method Integrated Gradients
182 (Sundararajan et al., 2017) to explain the network's predictions, and the results were very similar
183 to those based on Deep SHAP (not shown).

3. Results

We start by presenting the results for the case of annual maximum daily precipitation in Figure 2. We first discuss the results for a future climate with no intervention. The global-mean annual maximum precipitation exhibits an increase throughout the century but with large ensemble spread (magenta lines, Figure 2a). The largest increases occur in the deep tropics, specifically over the tropical Pacific (Figure 2b; see also O'Gorman and Schneider, 2009; Kharin et al., 2013; Pfahl et al., 2017). The network can successfully distinguish between the reference climate and the SSP future climate 85% of the time, which is significant at a 0.01 level (Figure 2d). Moreover, the probability assigned by the network that a map corresponds to the future SSP climate increases linearly with the actual year of the map and maximizes in the out-of-sample years 2060-2069 (Figure 2d). This suggests that there are robust signals of climate change that become more and more evident with time. It also suggests that the learned patterns generalize successfully, since the network is able to correctly classify the years 2060-2069, although those years were not used during training (see Supplementary Text S1). Based on the results from the XAI method Deep SHAP, the network mainly uses precipitation extremes over the tropical eastern Pacific (and to a lesser degree over the Southern Ocean and the tropical Atlantic) to make its predictions (Figure 2f). Interestingly, the network does not use precipitation over the western Pacific or Australia, despite the fact that the corresponding ensemble mean difference between the two periods is of high magnitude (Figure 2b). This implies high internal variability of precipitation extremes over these regions, which does not make them robust indicators from a signal-to-noise perspective.

Under the SAI scenario, the overall accuracy of the network is only 58% (Figure 2e), which is not statistically different from a random chance-based model (at a 0.01 significance level, a random chance-based model would perform with up to 69% accuracy, derived using a binomial distribution). The network-estimated probability that a map corresponds to the SAI climate is almost independent from the year of the map (Figure 2e), which indicates that there are no robust

long-term climate signals under SAI that the network could use for distinguishing from the reference climate. This is also suggested by the XAI results; note the incoherent and noisy attributions in Figure 2g. Generally, the results in Figure 2 indicate that although the CESM2(WACCM6) simulates a robust increase in future extreme daily precipitation under the SSP2-4.5 scenario, possible deployment of SAI could preserve the conditions of the reference (i.e., pre-deployment) climate. This could be an example of a potential positive SAI impact.

Next, we consider the annual mean surface temperature over land (Figure 3). Under the SSP scenario, a clear increase in surface temperature is shown throughout the century that is evident globally (Figure 3a-b). Accordingly, the network accuracy in distinguishing between the reference and the future SSP climate is high, on the order of 93%. Many regions around the globe are highlighted by Deep SHAP as robust distinctive patterns; e.g., Mexico, southern South America, southern Africa, Indonesia, and southern Australia. Under the SAI scenario, although the global mean temperature is similar to the one under the reference climate, there are robust patterns of regional cooling that make the two climates *highly* distinguishable: 91% of the time (Figure 3e). Regional cooling happens mainly over southern South America, eastern Africa, eastern Australia, and Greenland (Figure 3c). These are the regions that the network uses to distinguish between the reference and the SAI climates (see Figure 3g). Overall, these results indicate that the CESM2(WACCM6) projects that a potential SAI deployment would lead to a less warm climate than SSP; however, the annual mean surface temperature over land in an SAI world would also be distinguishable from the reference climate. Importantly, the distinctive patterns in the two scenarios are quite different, with warming being the distinctive difference under the SSP scenario, while regional cooling patterns being the most robust distinctive patterns under SAI.

We have repeated the same analysis as in Figures 2-3 for a list of 21 variables (see Table S1), and we summarize the results in Figure 4. For all variables, the network accuracy under the SSP scenario (magenta circles in Figure 4a) is statistically significant. This means that even under

234 the intermediate climate change scenario SSP2-4.5, the CESM2(WACCM6) projects that the Earth
235 system would exhibit climatic conditions that are distinguishable from the reference climate in the
236 coming decades. However, for the majority of variables examined here, SAI would lead to a less
237 distinguishable climate than the SSP scenario, although (with a few exceptions) one that would
238 also be distinguishable from the reference climate (note that the network accuracy (light blue
239 circles) is higher than the random chance-based accuracy). In particular, SAI would decelerate
240 many future greenhouse-gas driven climate changes, especially for surface temperature extremes,
241 precipitation, drought occurrence, sea level pressure, and Arctic sea ice (see also Xu et al., 2020;
242 Tye et al., 2022; Lee et al., 2020; 2023). It is important to note, however, that there are variables
243 for which SAI is projected to have minimal impact relative to climate change. Examples include
244 soil moisture, evapotranspiration, and ocean acidity.

245 We next explore how distinctive patterns might be modified from SAI; note that the
246 network accuracy alone does not provide this information. For example, as is shown in Figure 3,
247 the climate distinguishability under the SSP and the SAI scenarios is very similar, but the
248 corresponding distinctive patterns are different. To explore this further, the spatial correlation
249 between the XAI heatmaps under the SSP and SAI scenarios are presented in Figure 4b. In most
250 cases, the correlation is not statistically different from zero, which means that SAI is projected to
251 introduce different distinctive patterns relative to those from the SSP scenario. Exceptions are for
252 cases where the correlation is high, such as for ocean acidity and ocean heat, which means that the
253 anticipated SSP-driven distinctive patterns are projected to remain almost unchanged under SAI.

254 The results in Figure 4 indicate the diverse impacts of SAI on different components of the
255 climate system, which highlights the need for systematic and thorough investigations into the
256 possible impacts of SAI on the Earth system beyond only the global-mean temperature response.
257 Such research is needed for a well-informed policy making regarding potential deployment of
258 climate intervention approaches (NASEM, 2021). The framework introduced here allows for such

data-driven and generic investigations to uncover the ways in which an SAI climate would be different from a pre-deployment one.

4. Conclusions

In this study, a new framework was used that allows quantification (with a single number) of the degree of climate distinguishability between a reference climate and future climate states from both SAI and no-SAI worlds. The framework is based on the use of machine learning and leverages XAI tools to identify robust distinctive patterns under the intervention and the no-intervention scenarios. The framework is purely data driven, nonlinear, nonlocal, and it accounts for underlying uncertainties in the data that may originate from internal stochastic variability or uncertainties in Earth system model physics.

We applied this framework to data from ensembles of simulations that were developed to examine the potential impacts of stratospheric aerosol injection; namely, the ARISE-SAI project (Richter et al., 2022). In these simulations, SAI was shown to have diverse impacts on the simulated climate. These include minimizing changes due to greenhouse gas forcing in temperature and precipitation extremes, while having negligible effect on ocean acidification. Also, for the majority of variables examined here, the simulated deployment of SAI led to new patterns of change with respect to the reference climate that were different from the SSP patterns. This raises the possibility of SAI leading to *new* (and perhaps unwanted) changes in specific components of the Earth system or in certain regions of the world.

We do note some potential limitations of the presented framework. One is the dependence of the results on the amount of data. Neural networks are known to be "data-thirsty" models (LeCun et al., 2015), so it is possible that certain patterns that were not identified as robust indicators during training could become robust with more data. However, the dependence on the amount of data is present in virtually all climate settings involving questions of signal-to-noise and statistical significance. Another limitation is the possible dependence of the results on the network

architecture. In order to address this issue here, we searched over many different architectures and combinations of hyperparameters before training the network, as described in Supplementary Text S1. That way, we let the data guide us as to what architecture we should use for each climate variable. Yet, we acknowledge that it is possible that some of these results depend on the adopted architectures.

Our work highlights the need to further research the impacts of possible intervention approaches *beyond* just global mean temperatures, as has been done in other studies, examining ARISE-SAI data in particular (Keys et al., 2022; Labe et al., 2023; Hueholt et al., 2023). In doing so, we envision that the notion of "quantifiable climate distinguishability" will be a relevant and informative metric to assess impacts and to expand the design space of possible interventions (Lee et al., 2020), as illustrated by the presented results. Further investigation could include further assessing the climate distinguishability by considering multiple variables at the same time (i.e., the network input consists of many channels each of which refers to a different variable), to assess potential impacts on the dependence structure of different components of the Earth system and the occurrence of compound events. Future work could also focus on analyzing the output of more than one model and of more than one climate intervention strategy to establish a more holistic picture of the potential impacts of proposed climate intervention strategies.

Acknowledgments

This work was supported by Defense Advanced Research Projects Agency (DARPA) Grant No. HR00112290071. The views expressed here do not necessarily reflect the positions of the U.S. government. The authors would also like to thank the efforts of the ARISE-SAI team for making their data publicly available.

Data availability

The ARISE-SAI data is publicly available at https://www.cesm.ucar.edu/community-projects/arise-sai. The code to reproduce the presented results is publicly available at https://github.com/amamalak/Quantify_SAI_impacts.

References

Crutzen, P.J. (2006) Albedo Enhancement by Stratospheric Sulfur Injections: A contribution to Resolve a Policy Dilemma? *Clim. Change*, **77**(3-4), 211-220, doi:10.1007/s10584-006-9101-y.

Danabasoglu, G., et al. (2020) The Community Earth System Model Version 2 (CESM2), *J. Adv. Model. Earth Sy.*, **12**, e2019MS001916, https://doi.org/10.1029/2019MS001916.

Gettelman, A., et al. (2019) The whole atmosphere community climate model version 6 (WACCM6), *J. Geophys. Res.-Atmos*., **124**, 12380–12403, https://doi.org/10.1029/2019JD030943.

Herzog, H.J., (2001) What Future for Carbon Capture and Sequestreation?, *Environ. Sci. Technol*., **35**(7), 148A-153A.

Hueholt, D.M., E.A. Barnes, J.W. Hurrell, J.H. Richter, and L. Sun (2023) Assessing Outcomes in Stratospheric Aerosol Injection scenarios shortly after deployment, authorea preprints.

Jones, A.C., M.K. Hawcroft, J.M. Haywood,A. Jones, X. Guo, and J.C. Moore (2018) Regional climate impacts of stabilizing global warming at 1.5 K using solar geoengineering, *Earth's Future*, **6**, 230-251.

Keys, P.W., E.A. Barnes, N.S. Diffenbaugh, J.W. Hurrell, and M.B. Curtis (2022) Potential for perceived failure of Stratospheric Aerosol Injection deployment, *PNAS*, **119**(40), e2210036119.

Kharin, V.V., F.W. Zwiers, X. Zhang, et al. (2013) Changes in temperature and precipitation extremes in the CMIP5 ensemble, *Climatic Change*, **119**, 345-357.

Kravitz, B., et al. (2017) First simulations of designing stratospheric sulfate aerosol geoengineering to meet multiple simultaneous climate objectives, *J. Geophys. Res.-Atmos*., **122**, 12616-12634.

342 Kravitz, B., and D.G. MacMartin (2020) Uncertainty and the basis for confidence in solar
343 geoengineering research, *Nat Rev Earth Environ*, **1**, 64-75.
344 Labe, Z.M., E.A. Barnes, and J.W. Hurrell (2023) Identifying the regional emergence of climate
345 patterns in the ARISE-SAI-1.5 simulations, *Environmental Research Letters*,
346 https://doi.org/10.1088/1748-9326/acc81a.
347 LeCun, Y., Y. Bengio, and G. Hinton (2015) Deep learning, *Nature*, **521**, 436-444,
348 https://doi.org/10.1038/nature14539.
349 Lee, W., D. MacMartin, D. Visioni, and B. Kravitz (2020) Expanding the design space of
350 stratospheric aerosol geoengineering to include precipitation-based objectives and explore
351 trade-offs, *Earth Syst. Dynam*., **11**(4), 1051-1072.
352 Lee, W.R., et al., (2023) High-latitude stratospheric aerosol injection to preserve the Arctic,
353 *Earth's Future*, **11**(1), e2022EF003052.
354 Lundberg, S. M. and S. I. Lee (2017) A unified approach to interpreting model predictions," *Proc.*
355 *Adv. Neural Inf. Process. Syst.*, pp. 4768-4777.
356 MacMartin, D.G., B. Kravitz, D.W. Keith, and A. Jarvis (2014) Dynamics of the coupled human-
357 climate system resulting from closed-loop control of solar geoengineering, *Clim. Dynam*., **43**,
358 243-258.
359 MacMartin, D.G., et al. (2017) The Climate Response to Stratospheric Aerosol Geoengineering
360 Can Be Tailored Using Multiple Injection Locations, *J. Geophys. Res. Atmos*., **122**(23),
361 12,574-12,590, doi:10.1002/2017JD026868.
362 MacMartin, D.G., K.L. Ricke, and D.W. Keith (2018) Solar geoengineering as part of an overall
363 strategy for meeting 1.5∘C Paris target, *Philosophical Transactions of the Royal Society A*,
364 **376**, 20160454.

MacMartin, D.G., W. Wang, B. Kravitz, S. Tilmes, J.H. Richter, and M.J. Mills (2019) Timescale for detecting the climate response to stratospheric aerosol geoengineering. *Journal of Geophysical Research: Atmospheres*, **124**, 1233-1247.

Mamalakis, A., I. Ebert-Uphoff, E.A. Barnes (2022a) Explainable Artificial Intelligence in Meteorology and Climate Science: Model fine-tuning, calibrating trust and learning new science, in *Beyond explainable Artificial Intelligence* by Holzinger et al. (Editors), Springer Lecture Notes on Artificial Intelligence (LNAI).

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

Mamalakis, A, E.A. Barnes and I. Ebert-Uphoff (2022c) Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience, *Artificial Intelligence for the Earth Systems*, **1**(4), e220012.

Mamalakis, A, E.A. Barnes and I. Ebert-Uphoff (2023) Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience, *Artificial Intelligence for the Earth Systems*, **2**(1), e220058.

McGovern, A., *et al.,* (2019) Making the black box more transparent: Understanding the physical implications of machine learning," *Bulletin of the American Meteorological Society*, vol. **100**, no. 11, pp. 2175-2199.

National Academies pf Sciences, Engineering, and Medicine (2021) Reflecting Sunlight: Recommendations for Solar Geoengineering Research and Research Governance. Washington, DC: The National Academies Press. https://doi.org/10.17226/25762.

National Research Council (2015) Climate Intervention: Carbon Dioxide Removal and Reliable Sequestration. Washington, CD: The National Academies Press. https://doi.org/10.17226/18805.

Niemeier, U., and S. Tilmes (2017) Sulfur injections for a cooler planet, *Science*, **357**(6348), 246-248, doi:10.1126/science.aan3317.

O'Gormanm, O.A., and T. Schneider (2009) The physical basis for increases in precipitation extremes in simulations of 21st-century climate change, *PNAS*, **106**(35), 14773-14777.

O'Neill, et al. (2017) The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century, *Global Environ. Change*, **42**, 169-180.

Parker, D.E., H. Wilson, P.D. Jones, J.R. Christy, C.K. Folland (1996) The impact of mount Pinatubo on world-wide temperatures, *Int. J. Climatol*., **16**, 487-497.

Pfahl, S., P. O'Gorman, and E. Fischer (2017) Understanding the regional pattern of projected future changes in extreme precipitation, *Nature Climate Change*, **7**, 423-427.

Richter, J.H., et al. (2022) Assessing Responses and Impacts of Solar climate Intervention on the Earth system with stratospheric aerosol injection (ARISE-SAI): protocol and initial results from the first simulations, *Geosci. Model Dev*., **15**, 8221-8243.

Robock, A. (2000) Volcanic eruptions and climate, *Rev. Geophys*., **38**(2), 191-219, doi:10.1029/1998RG000054.

Robock, A., and J. Mao (1995) The Volcanic Signal in Surface Temperature Observations, *J. Climate*, **8**(5), 1086-1103.

Robock, A., A. Marquardt, B. Kravitz, and G. Stenchikov (2009) Benefits, risks, and costs of stratospheric geoengineering, *Geophys. Res. Lett.*, **36**(19), L19703, doi:10.1029/2009GL039209.

Smith, W., and G. Wagner (2018) Stratospheric aerosol injection tactics and costs in the first 15 years of deployment, *Environ. Res. Lett.*, **13**, 124001.

Soden, B.J., et al. (2002) Global cooling after the eruption of mount Pinatubo: A test of climate feedback by water vapor, *Science*, **296**, 727-730, doi:10.1126/science.296.5568.727.

Sundararajan, M., A. Taly, Q. Yan, (2017) Axiomatic attribution for deep networks," arXiv preprint, https://arxiv.org/abs/1703.01365.

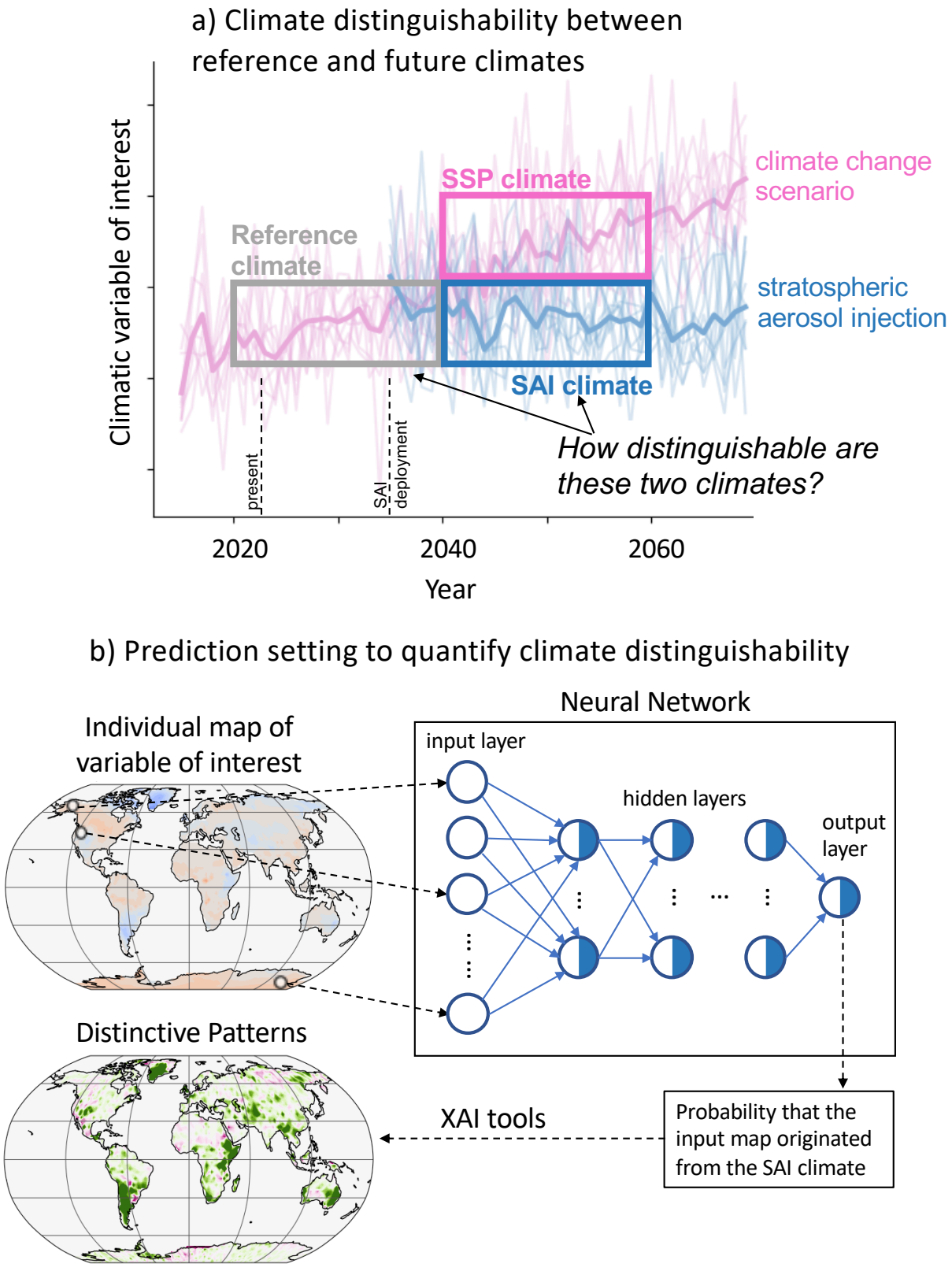
Tilmes, S., et al. (2018) CESM1(WACCM) Stratospheric Aerosol Geoengineering Large Ensemble Project, *Bull. Am. Meteorol. Soc*., **99**(11), 2361-2371, doi:10.1175/BAMS-D-17-0267.1.

Tilmes, S., et al. (2020) Reaching 1.5 and 2.0 ∘C global surface temperature targets using stratospheric aerosol geoengineering, *Earth Syst. Dynam*., **11**, 579–601, https://doi.org/10.5194/esd-11-579-2020.

Toms, B.A., E. A. Barnes, I. Ebert-Uphoff, "Physically interpretable neural networks for the geosciences: Applications to Earth system variability," *Journal of Advances in Modeling Earth Systems*, vol. **12**, e2019MS002002, 2020.

Tye, M.R., K. Dagon, M.J. Molina, J.H. Richter, D. Visioni, B. Kravitz, and S. Tilmes (2022) Indices of extremes: geographic patterns of change in extremes and associated vegetation impacts under climate intervention, *Earth Syst. Dynam.*, **13**, 1233-1257.

Vaughan, N.E. and T.M. Lenton (2011) A review of climate geoengineering proposals, *Clim. Change*, **109**(3-4), 745-790.



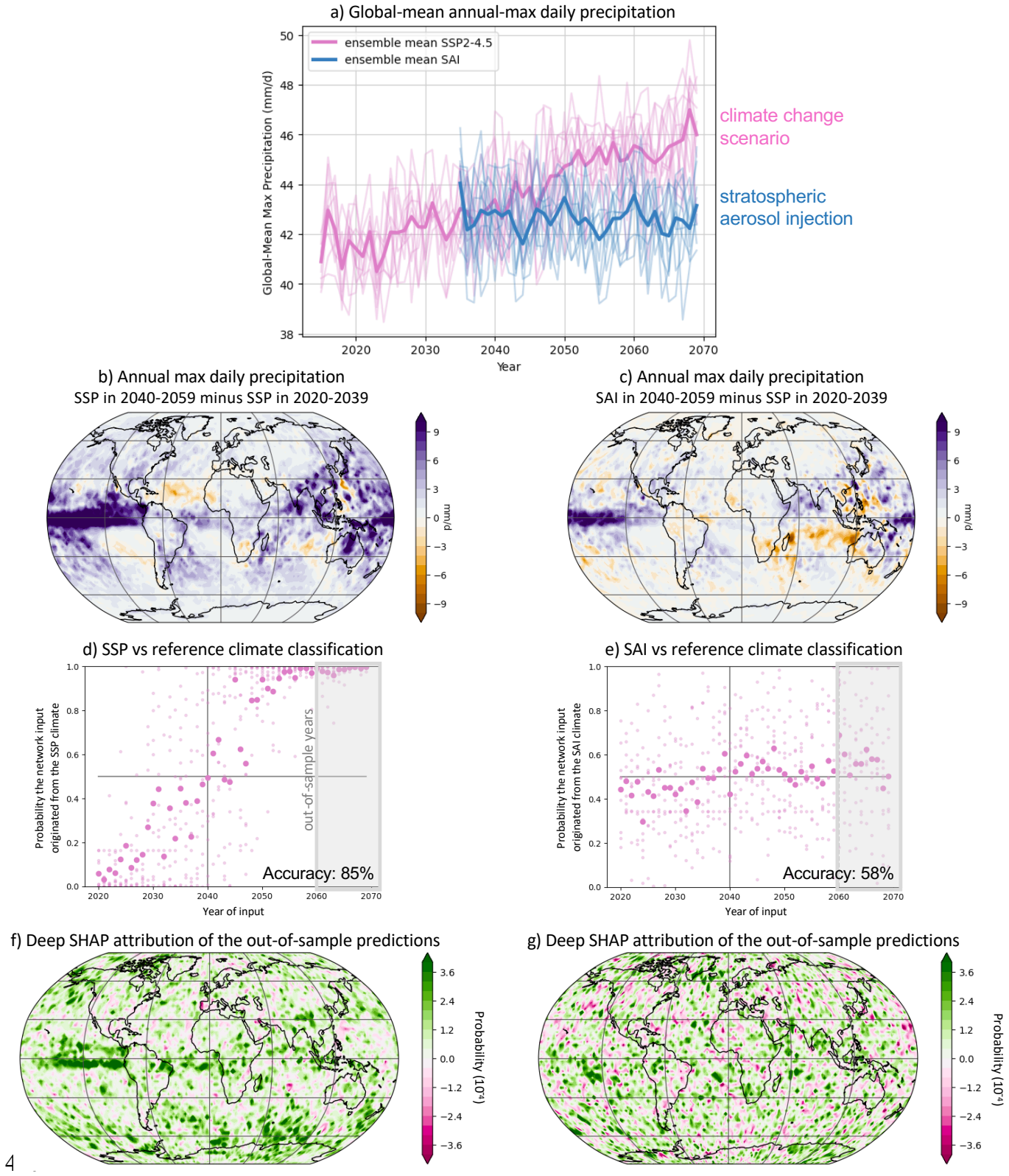
433

434 **Figure 1:** Schematic of our framework to quantify SAI impacts using XAI. a) Assessing climate

435 distinguishability between reference and future climates. Note that the pre-2040 period under an

intermediate climate change scenario is used as the refence climate, in accordance to Richer et al (2022). b) Schematic of the prediction task to quantify climate distinguishability after SAI and the use of XAI to derive the distinctive patterns between the reference and SAI climates.

439

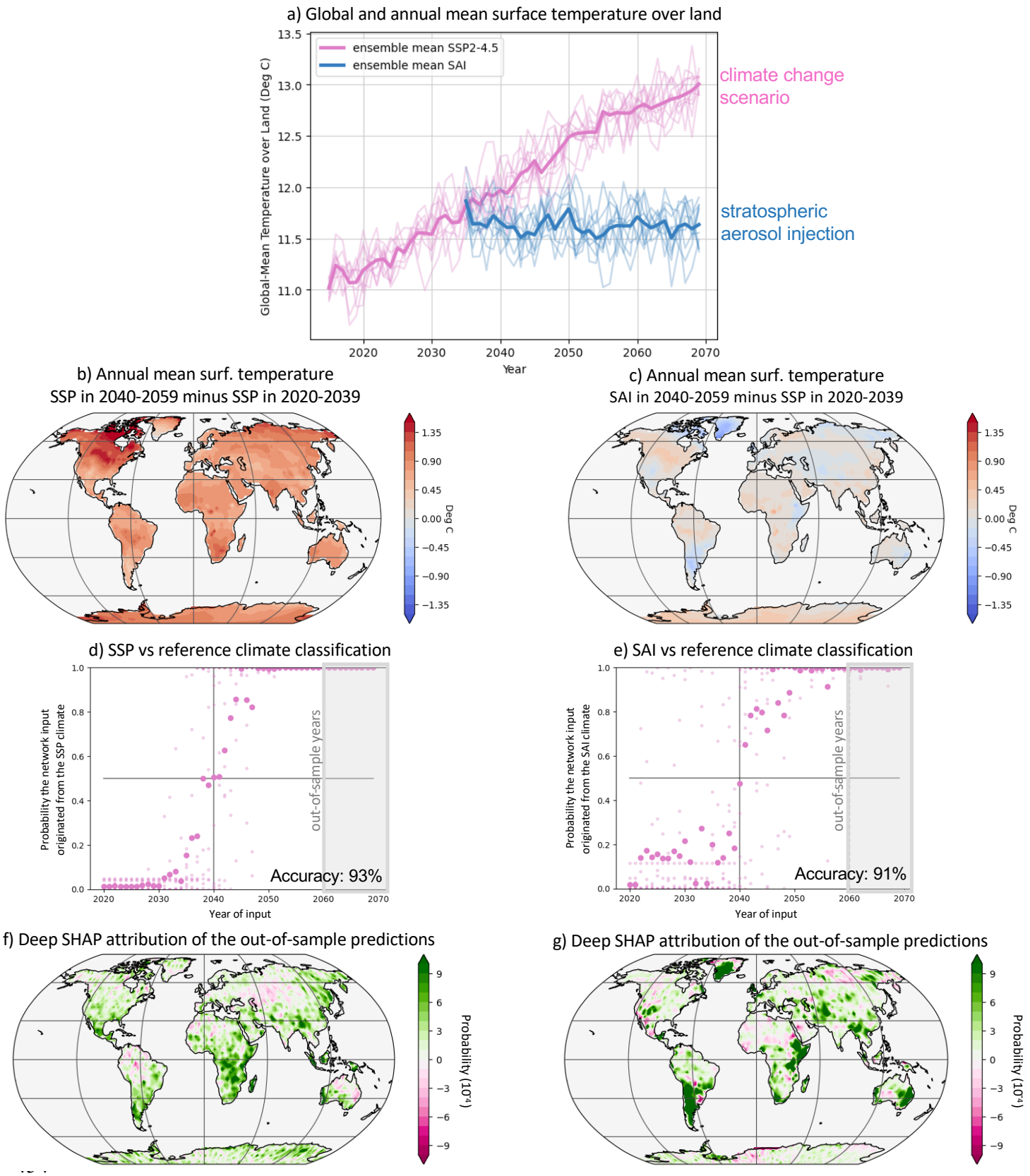


441 **Figure 2.** Results of our framework for annual maximum daily precipitation. a) Series of global-mean

442 annual maximum precipitation (in mm/d) under the SSP2-4.5 scenario and the ARISE-SAI scenario. All

10 ensemble members and the ensemble mean are shown. b) Ensemble mean difference between the annual maximum precipitation in the 2040-2059 SSP2-4.5 climate and the reference climate. d) Network-generated probability that different annual maximum precipitation maps originated from the 2040-2059 SSP2-4.5 climate. The actual year of each map is provided in the horizontal axis. The overall accuracy of the network is shown on the bottom right corner. f) Distinctive patterns that were used by the network to separate the reference climate from the 2040-2059 SSP2-4.5 climate, as estimated using the method Deep SHAP. The presented attributions correspond to the average attributions across the 2060-2069 network predictions and all testing members, using the years 2035-2044 as baseline. c,e,g) Same as (b,d,f), but the network is trained to separate the reference climate from the 2040-2059 ARISE-SAI climate.

453



455 **Figure 3.** Same as in Figure 2, but results are for the annual mean surface temperature over land.

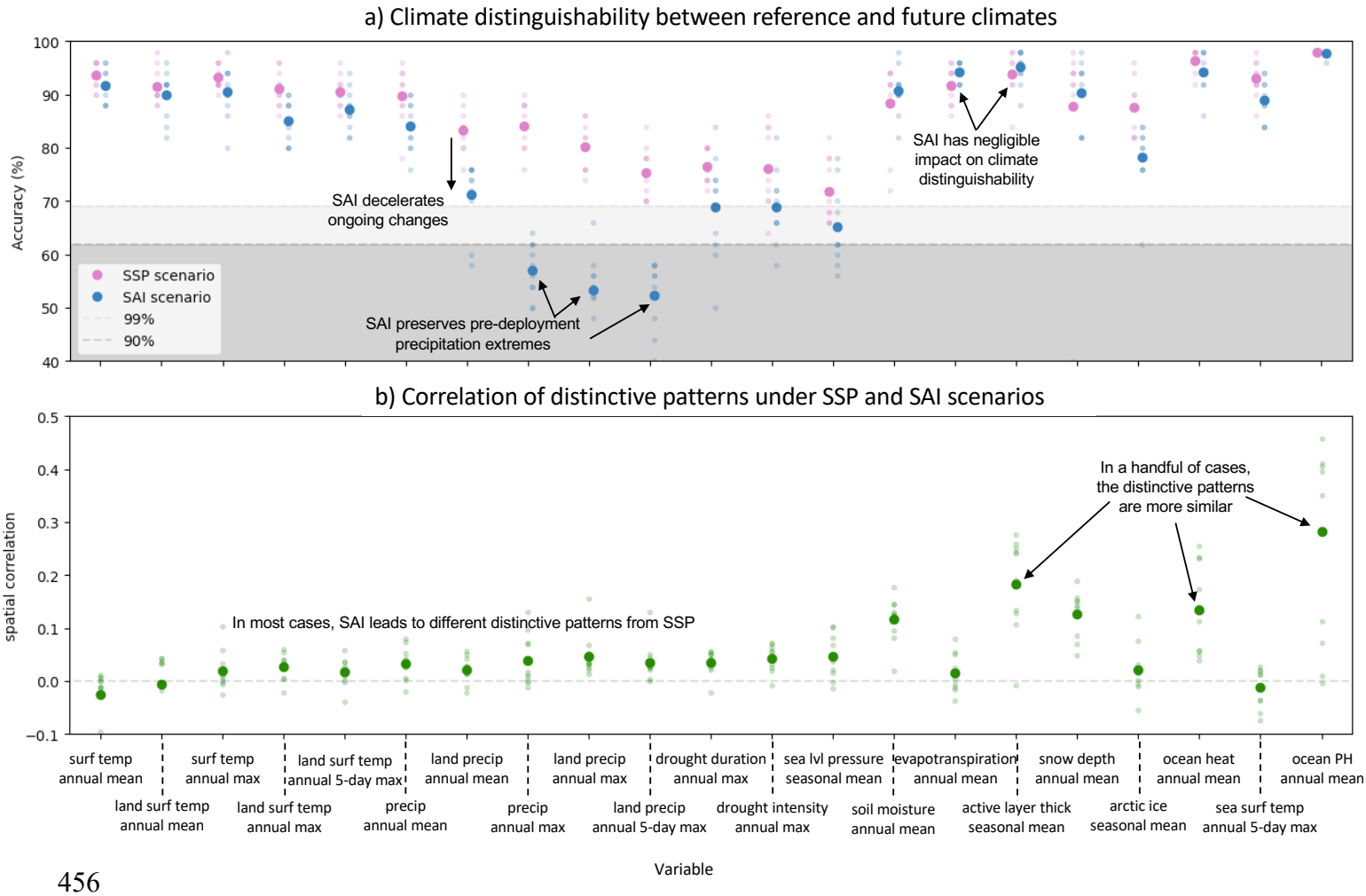


Figure 4. a) Accuracy of the network in distinguishing between the reference climate and the future SSP 2-4.5 climate (magenta) or the future ARISE-SAI climate (light blue), for all variables considered in the study (see Supplementary Table S1). Results from individual testing members (smaller circles) and the ensemble mean (bigger circles) are presented. The critical values for the 10% and 1% significance levels are derived using a binomial distribution. b) Correlation coefficient between attribution heatmaps that correspond to predicting in the two scenarios. Results from individual testing members (smaller circles) and the ensemble mean (bigger circles) are presented.

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:

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

where M is the set of all input features, $M \backslash \{i\}$ 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 $S \subseteq M \backslash \{i\}$. 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.