Response to the reviewers: Analysis of the cloud fraction adjustment to aerosols and its dependence on meteorological controls using explainable machine learning

 # EGUSPHERE-2023-1667

Yichen Jia [1, 2], Hendrik Andersen [1, 2], and Jan Cermak [1, 2]

[1]Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research, Karlsruhe, Germany
[2]Karlsruhe Institute of Technology (KIT), Institute of Photogrammetry and Remote Sensing, Karlsruhe, Germany

Correspondence: Yichen Jia (yichen.jia@kit.edu)

We thank the anonymous referee for the new round of review of the revised manuscript. Please see the main points that the reviewer is concerned about. Below, the reviewer's comments and suggestions are incorporated in italics and addressed hereafter, and the authors' responses are coloured in blue. Unless otherwise stated, line numbers in this document refer to the manuscript after the second-round review (before the updates following in this response letter).

Referee 2

Specific comments

1. *Acknowledging the issues that are fundamental to the xgboost based approach, i.e., variable independence and potential artificial correlation is a good first step. But it does not address the important point.*

   Thank you for your feedback, we believe the manuscript has improved with the inclusion of a separate section discussing the method and data limitations.

2. *It remains unphysical. First of all, SHAP value is already kind of a sensitivity of the target value to the dependent variable. That is, SHAP value of CF is equivalent to dCF/dNd. Second of all, the SHAP value for Nd and its dependence on Nd figure the authors showed in the response. even if we forget the first point, are qualitatively different from the figure from the reference. Their shape is similar, which is true. However, the SHAP value turns to strongly negative values, which would be interpreted as CF decreases with Nd at these Nd values. That is unphysical either. I could name other physically inconsistencies if the authors show more details like this. The overaching point remains that we do not have reason to believe such boosted tree models would necessarily give us physical insights. I'd have not issues with authors publishing it as a statistical analysis, but if physical interpretations are involved the authors need to demonstrate them with care first.*

   We thank the reviewer for his/her comments. However, the assertions made by the reviewer concerning SHAP values are incorrect or inaccurate. The use of our method aligns with the design of SHAP values, a similar way of sensitivity estimation was also applied in a Nature Communications paper (Li et al., 2022). To clear up the confusion, our response therefore addresses each point made by the reviewer separately in the following:

(a) It remains unphysical. First of all, SHAP value is already kind of a sensitivity of the target value to the dependent variable. That is, SHAP value of CF is equivalent to dCF/dNd.

Short answer: The interpretation of SHAP values as sensitivity and them being comparable to dCF/dNd is incorrect. A better analogy would be that SHAP values are comparable to $dCF/dN_d \times N_{dj}$, where j denotes the N_d value for a specific data instance x_j. Other studies (e.g., Li et al., 2022) have employed the same sensitivity estimation strategy as we have done in our paper.

Longer answer: SHAP values quantify feature contributions for data instances (feature values), they are not a sensitivity estimate (see e.g. the paper from the developer (Lundberg et al., 2020) or this textbook on explainable machine learning (Molnar, 2022)). Regarding the specific mention of SHAP values for N_d (we assume the reviewer was referring to N_d because there are no "SHAP values of CF"), it should be noted that they are neither equivalent nor directly comparable to dCF/dN_d. SHAP values do not directly quantify the rate of change of the target value with respect to changes in the dependent variables. Instead, as we already explained in the manuscript (in lines 175–181), and our response to the first review by R2, SHAP values represent the contribution of each feature to the CF prediction (compared to a base value) for individual data instances (i.e. local contribution). In fact, SHAP values are more comparable to $dCF/dN_d \times N_{dj}$, where j denotes the N_d value for a specific data instance x_j.

Mathematically, SHAP values are derived from Equation (1), wherein, for a given non-zero subset S of feature values, the prediction is assumed to be equal to the expected value of the function conditioned on S ($f_x(S) = E[f(x)|x_S]$) (Lundberg and Lee, 2017; Lundberg et al., 2018).

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\} - f_x(S) \right] \tag{1}$$

Equation (1) combines the traditional equation for Shapley values with conditional expectations. N is the set of all features, i denotes the i-th feature, M is the number of input features and f is the ML model function.

Because SHAP values do not provide a sensitivity estimate, we quantify the sensitivity by fitting a linear regression to the feature values and their respective SHAP values. This method has been used before, e.g. by Li et al. (2022) in their Nature Communications paper. In this study, the authors estimate the sensitivity of leaf area index (LAI) to sub– or near–surface soil moisture (SMsurf) as the slope of the regression between feature values of SMsurf anomalies and their SHAP values (Fig. 1 of this letter). The authors further argue that this approach facilitates the robustness of the sensitivity estimation compared to traditional statistical methods because it "combines the advantages of bootstrap aggregating and non-distribution-assumption by random forest modeling, as well as advantages of global interpretations being consistent with the local explanations in the SHAP algorithm (Li et al., 2022, page 7, section on overall sensitivity)." Our study also benefits from similar advantages and enhanced robustness. We have referenced this study and included relevant discussion in Sect. 2.3.2 of the manuscript at line 212.

(b) Second of all, the SHAP value for Nd and its dependence on Nd figure the authors showed in the response. even if we forget the first point, are qualitatively different from the figure from the reference. Their shape is similar, which

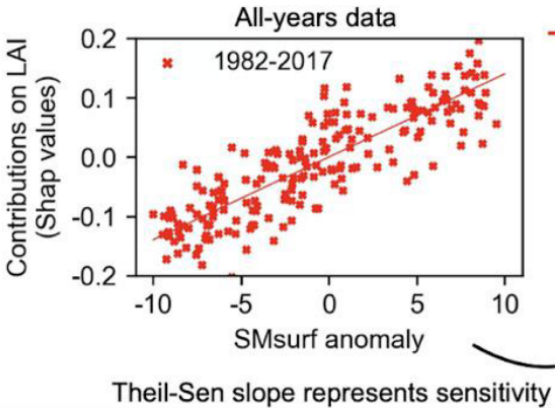

Figure 1. Figure adapted from the supplementary information of (Li et al., 2022), showing the sensitivity of leaf area index (LAI) to sub– or near–surface soil moisture (SMsurf) as the slope of the Theil-Sen linear regression between SHAP values and feature values.

is true. However, the SHAP value turns to strongly negative values, which would be interpreted as CF decreases with Nd at these Nd values.

Short answer: Again, the interpretation of the SHAP values by R2 is not correct, and thus only the acknowledgement of the similarities between our results and the findings by Yuan et al. (2023) remains.

Longer answer: It is incorrect to interpret negative SHAP values for N_d at low N_d values as CF decreases with N_d (i.e. negative sensitivity). As we already explained in Sect. 2.3.1 of the manuscript in lines 177–180, as well as in our first response to the first review of R2, SHAP values indicate that the specific feature value increases/decreases the prediction compared to the "base value" (average of all predictions). SHAP values therefore quantify the extent to which each feature contributes to a prediction deviating from the model's average prediction/baseline. Fig. 2 from Lundberg et al. (2018) shows an example plot of how the sum of the baseline value and all feature contributions (positive and negative SHAP values) is equal to the individual model prediction. These feature contributions depend on the feature value. In the example $E[f(x)]$ is the base value (the prediction if we did not know any information on the input features), and $f(x)$ is the current model output. This plot illustrates how positive SHAP values (red) push the model prediction higher and negative SHAP values (blue) push the model prediction lower. This also indicates an important internal consistency of SHAP: $\sum_{i=0}^{M} \phi_i = f(x)$, where $i = 0$ denotes what would be predicted in the absence of any feature information (base value).

Figure 3 (which we have already shown in our first response to R2) shows the simplest case of SHAP values when they are applied to a linear model. The slopes of linear regressions fitted to the original data and the SHAP values are equal. The only difference is between the y-axes of $E[f(x)]$, as the base value is subtracted in the case of the SHAP values. This reaffirms the validity of using the linear regression slope of feature values and SHAP values for sensitivity estimation, but also clearly shows how R2 misinterprets the negative SHAP values at low N_d as a negative sensitivity. In the case of a positive linear sensitivity low feature values (in Fig. 3 MedInc) will lead to

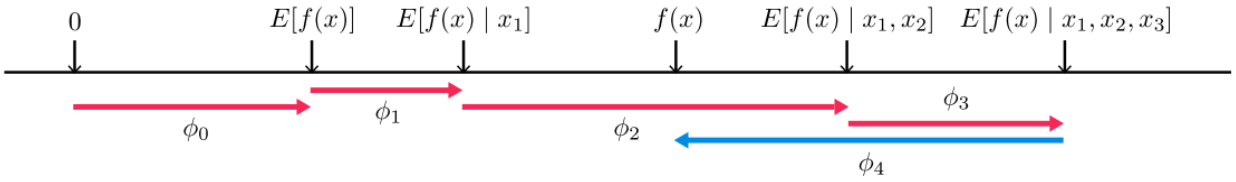

Figure 2. Illustration of how SHAP values explain the output of a function f as a sum of the effects ϕ_i of each feature being introduced into a conditional expectation (Lundberg et al., 2018).

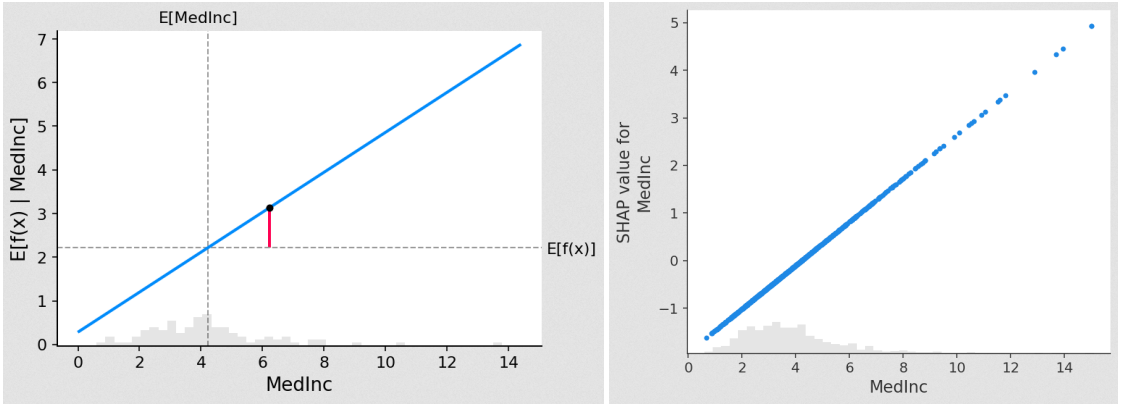

Figure 3. Example figures showing the application a linear regression (left) and its SHAP values (right) - the only difference in this case is the subtraction of the base value (horizontal dashed line of $E[f(x)]$ in the left panel). Figures are from the SHAP documentation page, section "An introduction to explainable AI with Shapley values" on: https://shap.readthedocs.io/en/latest/overviews.html, last accessed 03 March 2024.

a below-average prediction of y (left panel), which in the case of SHAP values is a negative value (right panel). Therefore, negative SHAP values "decrease" the prediction with respect to the base value (i.e. at very low N_d values a below-average CF is expected). In our study, the correct interpretation of local/individual SHAP values is provided in lines 177-179: "Positive (negative) SHAP values indicate that the specific feature value increases (decreases) the prediction compared to this base value.", and for the example of the global interpretation of N_d SHAP values in Fig. 1a) is given in line 201: "...increased N_d values lead to an increase in the predicted CLF, while the rate of the increase ($dSHAP/dN_d$) drops with N_d as shown by the orange line." This is physically expected and agrees well with the cited literature.

In summary, SHAP values are not a sensitivity measure. Sensitivity should be interpreted from a global perspective, whereas SHAP values should be initially interpreted from a local viewpoint, and then aggregated and summarized by a global sensitivity (here the slopes of the linear regressions, as also done in e.g. Li et al. (2022)), or global feature importance (the mean of absolute local SHAP values e.g. Fig. 3 of the manuscript).

This is well summarized in the explainable machine learning textbook by Molnar (2022): "The global interpretation methods include feature importance, feature dependence, interactions, clustering and summary plots. With SHAP, global interpretations are consistent with the local explanations, since the Shapley values are the "atomic unit" of the global interpretations (section 9.6.10)."

(c) *That is unphysical either. I could name other physically inconsistencies if the authors show more details like this. The overaching point remains that we do not have reason to believe such boosted tree models would necessarily give us physical insights. I'd have not issues with authors publishing it as a statistical analysis, but if physical interpretations are involved the authors need to demonstrate them with care first.*

The concerns raised regarding unphysical N_d–CF relationships or physically inconsistent ones are based on the reviewer's misinterpretation of SHAP values, but are shown to be physically consistent and in agreement with the literature. We certainly agree with the reviewer that the interpretation regarding physical processes should always be done with care, (no matter if using a linear regression, a neural network or boosted trees) which we have done by:

i. In the abstract, now explicitly stating that the results are based on a statistical/data-driven method and mentioning that limitations are discussed in the main text.

ii. Openly discussing the potentials and limitations of the method and the data (e.g. in Sect. 2.3.3).

iii. Directly stating in line 273 that the interpretation should be done with the limitations in mind.

iv. Careful wording when physical interpretations are made (e.g. "These marked positive N_d–CLF sensitivities *may be* caused by high N_d delaying the transition from stratocumulus to cumulus clouds (Gryspeerdt et al., 2016; Christensen et al., 2020)", line 332).

v. We now additionally rephrase the manuscript including: add "seems to" before "indicates" at line 310.
Line 324: "leading to" to "which could lead to".
Line 361 and 362: "suggests that" to "may be a hint that"; add "presumably" after "increase of CLF".
Line 417: add "seem to" between "factors" and "have".

vi. Directly following up such interpretations by detailing possible unphysical alternative explanations (e.g. "As N_d retrievals tend to negatively bias at lower CLF and positively bias at higher CLF, the N_d–CLF sensitivity may be overestimated, and at the scales considered here, should be interpreted as an upper bound to the physical N_d–CLF sensitivity.", lines 337–339).

vii. Emphasizing again that the sensitivities and IAIs are subject to aforementioned limitations: "should be noted that the quantification of the dependence of the Nd–CLF relationship on meteorological factors (EIS, SST discussed in this section) is also likely subject to the biases in the Nd–CLF sensitivity caused by the Nd retrieval biases as a function of CLF. This would potentially contribute to the non-causal facets of the relationships and interactive effects quantified by SHAP values.", lines 441–444.

viii. We have also rephrased Sect. 3.3.2 to underscore the caution in making physical interpretations.

ix. In the conclusion section, it has been mentioned that "The statistical sensitivities and interactive effects are interpreted with the guidance of hypothesised causal pathways and the state-of-the-art physical understanding of the system." (lines 451–453); we also mentioned incorporating causal setups for SHAP would be a promising way to go (lines 476–478). Furthermore, we revised the conclusion part to exercise caution and to serve as a reminder to readers.

3. *This proves my point. The transformation basically doesn't make sense. Nearly all figures and results in this paper are about gross statistics and map distributions. When the underlying statistics are not consistent, it has to be corrected IMHO.*

The use of standardized regression coefficients is a standard and common practice when aiming for comparability of sensitivity estimates among predictors. It is described and recommended as the standard procedure to eliminate the effect of units and place the predictors on the same scale in this paper: "A review of techniques for parameter sensitivity analysis of environmental models" (Hamby, 1994). As it is a general strategy to compare sensitivities across predictors, there are plenty of environmental studies that employ this strategy. We realize that the reviewer probably takes an issue with the resulting maps (standard deviations are not the same everywhere), rather than with the technique itself. We have revised the content introducing the standardization in lines 132–138 and have included a more direct discussion about the trade-off between comparability between predictors, vs. comparability in space into the manuscript. One should note though, that even in this context (maps of sensitivities) this is still a standard method. Examples can be found e.g. in this nature paper (Seddon et al., 2016), which quantifies and shows maps of standardized sensitivities for multiple predictors. In the cloud community it is also commonly done when sensitivities are compared among predictors, and in similar settings (papers showing standardized sensitivity maps). Here is an incomplete list of references that have opted for this strategy:

- (Scott et al., 2020): Journal of Climate paper that quantifies standardized low-cloud sensitivities for a range of cloud-controlling factors, please see Fig. 4 of this letter for examplary sensitivity maps (note the unit of the color bar).
- (Myers et al., 2021): Nature Climate Change paper that uses the standardized low-cloud sensitivities from Scott et al. (2020).
- (Ceppi and Nowack, 2021): PNAS paper that quantifies standardized cloud sensitivities for a range of cloud-controlling factors, as shown by Fig. 5 of this letter for examplary sensitivity maps.
- (Andersen et al., 2023): ACP paper that quantifies standardized sensitivities of cloud radiative effects to a range of cloud-controlling factors including aerosols.
- (Grise and Kelleher, 2021): Journal of Climate paper that quantifies standardized sensitivities of cloud radiative effects in the midlatitudes for a range of cloud-controlling factors, Fig. 6 of this letter for examplary sensitivity maps.

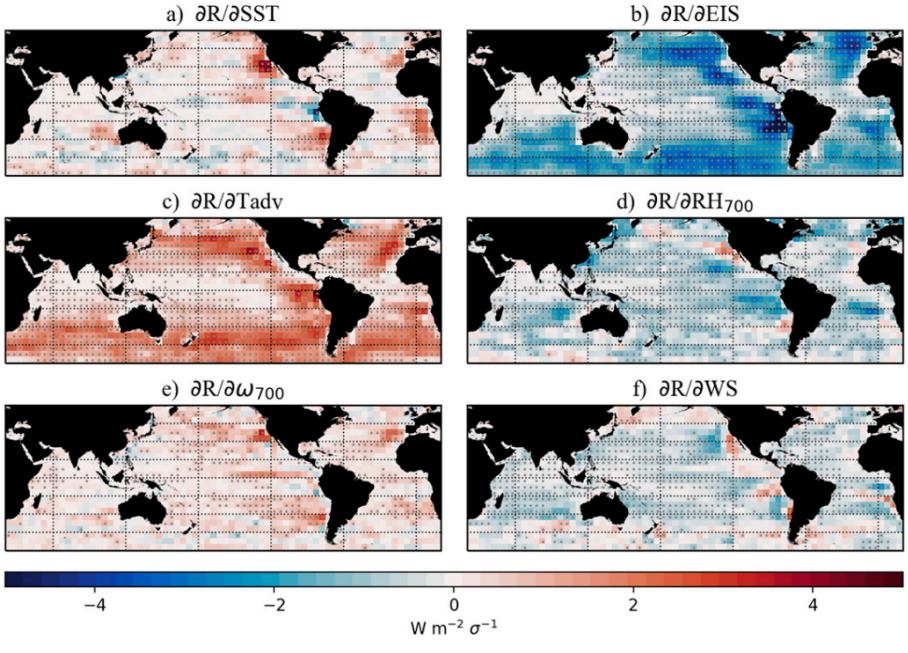

Figure 4. Figure taken from Scott et al. (2020) showing standardized low-cloud sensitivities to different cloud-controlling factors.

– (Wilson Kemsley et al., 2024): EGUsphere manuscript that quantifies standardized high-cloud sensitivities for a range of cloud-controlling factors.

We do not share the opinion of the reviewer that this method "does not make sense", and the published literature in the cloud community (but also more broadly in environmental sciences) supports this assessment. As we believe the added value of the comparability between the sensitivities of the different predictors is of interest to the readership of ACP, and outweighs the discussed downside of marginally reduced comparability in space (as shown by the supplementary material), we would like to keep the figures in the manuscript to show standardized sensitivities with the non-standardized sensitivities in the supplement (as done in Grise and Kelleher (2021)).

Minor modifications independent of the reviewer comments

Abstract, line 20: "ACIs" to "the CLF adjustment".

Line 20: "-" between "CLF" and "sensitivities" has been removed.

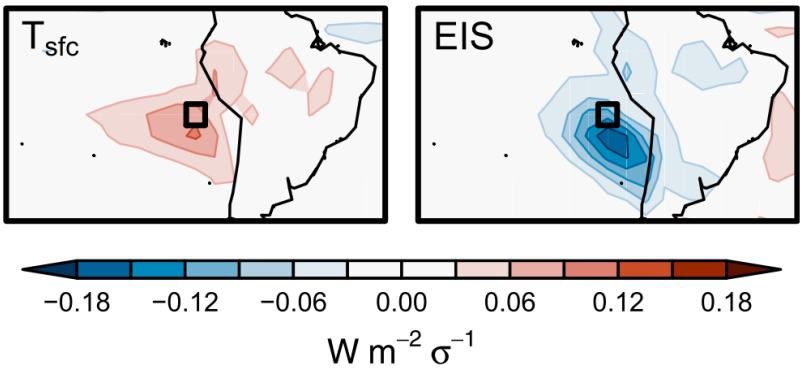

Figure 5. Figure from Ceppi and Nowack (2021) showing shortwave cloud-radiative sensitivities to standardized surface temperature and EIS.

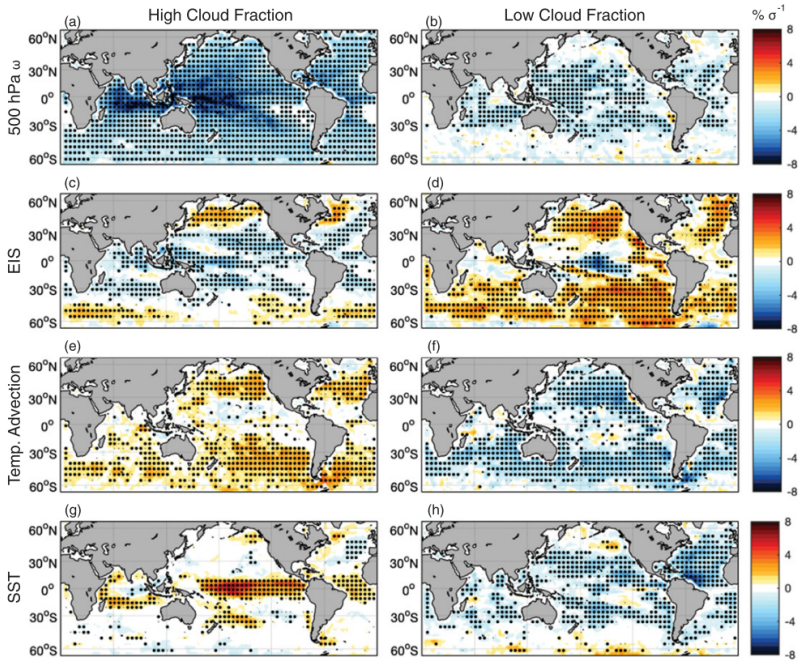

Figure 6. Figure from Grise and Kelleher (2021) showing sensitivities of low and high cloud fraction to four different standardized cloud-controlling factors.

References

Andersen, H., Cermak, J., Douglas, A., Myers, T. A., Nowack, P., Stier, P., Wall, C. J., and Wilson Kemsley, S.: Sensitivities of cloud radiative effects to large-scale meteorology and aerosols from global observations, Atmospheric Chemistry and Physics, 23, 10 775–10 794, https://doi.org/10.5194/acp-23-10775-2023, 2023.

Ceppi, P. and Nowack, P.: Observational evidence that cloud feedback amplifies global warming, Proceedings of the National Academy of Sciences, 118, e2026290 118, https://doi.org/10.1073/pnas.2026290118, 2021.

Christensen, M. W., Jones, W. K., and Stier, P.: Aerosols enhance cloud lifetime and brightness along the stratus-to-cumulus transition, Proceedings of the National Academy of Sciences of the United States of America, 117, 17 591–17 598, https://doi.org/10.1073/pnas.1921231117, 2020.

Grise, K. M. and Kelleher, M. K.: Midlatitude Cloud Radiative Effect Sensitivity to Cloud Controlling Factors in Observations and Models: Relationship with Southern Hemisphere Jet Shifts and Climate Sensitivity, Journal of Climate, 34, 5869–5886, https://doi.org/https://doi.org/10.1175/JCLI-D-20-0986.1, 2021.

Gryspeerdt, E., Quaas, J., and Bellouin, N.: Constraining the aerosol influence on cloud fraction, Journal of Geophysical Research, 121, 3566–3583, https://doi.org/10.1002/2015JD023744, 2016.

Hamby, D. M.: A review of techniques for parameter sensitivity analysis of environmental models, Environmental Monitoring and Assessment, 32, 135–154, https://doi.org/10.1007/BF00547132, 1994.

Li, W., Migliavacca, M., Forkel, M., Denissen, J. M. C., Reichstein, M., Yang, H., Duveiller, G., Weber, U., and Orth, R.: Widespread increasing vegetation sensitivity to soil moisture, Nature Communications, 13, 3959, https://doi.org/10.1038/s41467-022-31667-9, 2022.

Lundberg, S. M. and Lee, S. I.: A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems, 2017-Decem, 4766–4775, 2017.

Lundberg, S. M., Erion, G. G., and Lee, S.-I.: Consistent Individualized Feature Attribution for Tree Ensembles, ArXiv, abs/1802.0, http://arxiv.org/abs/1802.03888, 2018.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence, 2, 56–67, https://doi.org/10.1038/s42256-019-0138-9, 2020.

Molnar, C.: Interpretable Machine Learning, 2 edn., https://christophm.github.io/interpretable-ml-book, 2022.

Myers, T. A., Scott, R. C., Zelinka, M. D., Klein, S. A., Norris, J. R., and Caldwell, P. M.: Observational constraints on low cloud feedback reduce uncertainty of climate sensitivity, Nature Climate Change, 11, 501–507, https://doi.org/10.1038/s41558-021-01039-0, 2021.

Scott, R. C., Myers, T. A., Norris, J. R., Zelinka, M. D., Klein, S. A., Sun, M., and Doelling, D. R.: Observed Sensitivity of Low-Cloud Radiative Effects to Meteorological Perturbations over the Global Oceans, Journal of Climate, 33, 7717–7734, https://doi.org/https://doi.org/10.1175/JCLI-D-19-1028.1, 2020.

Seddon, A. W. R., Macias-Fauria, M., Long, P. R., Benz, D., and Willis, K. J.: Sensitivity of global terrestrial ecosystems to climate variability, Nature, 531, 229–232, https://doi.org/10.1038/nature16986, 2016.

Wilson Kemsley, S., Ceppi, P., Andersen, H., Cermak, J., Stier, P., and Nowack, P.: A systematic evaluation of high-cloud controlling factors, EGUsphere, 2024, 1–32, https://doi.org/10.5194/egusphere-2024-226, 2024.

Yuan, T., Song, H., Wood, R., Oreopoulos, L., Platnick, S., Wang, C., Yu, H., Meyer, K., and Wilcox, E.: Observational evidence of strong forcing from aerosol effect on low cloud coverage, Science Advances, 9, eadh7716, https://doi.org/10.1126/sciadv.adh7716, 2023.