Response to the reviewer #3: 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 third anonymous referee for the new round of review of the revised manuscript. 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 third-round review (before the updates following in this response letter).

Referee 2

Specific comments

1. *In this work, the authors use a novel method to identify the sensitivity of cloud fraction to aerosol variations. They find that Nd is strongly correlated to CF, after accounting for the impact of other cloud controlling factors. This work is clearly in scope for ACP and would be of interest to its readers. The work is of a good standard and I think would be suitable for publication with a few additions/changes.*

   I appreciate the authors have already done a considerable amount of work responding to other reviewers, I would only have a few small points to add here.

   Thank you for your positive evaluation of our manuscript. We have addressed your specific comments individually below.

2. *It seems that the interpretation of the SHAP values are not straightforward. This is not a fault of the author or the reader, but given this method is still fairly new, a little extra explanation could be useful. I am not sure directing the reader to read a textbook targeted at computer scientists is useful - with a small amount of extra text, this paper could provide an explanation of the SHAP-based techniques accessible at a broader range of atmospheric scientists and help encourage the use of this technique.*

   As I undersatnd it, the SHAP value does not represent a sensitivity (as the authors say),. However, this was not immediately clear to me (having not looked at this in much detail before). Phrases like 'way to measure the relative contributions of Nd (as a surrogate for aerosols) and meteorological factors to CLF changes' initially suggested to me that we were looking at a sensitivity-like measure. Assuming I am understanding this correctly, the 'base value' here is effectively the

climatological CLF at a given location (what would be predicted with no extra information. The SHAP values of Nd then show the CLF that would be predicted, given that the Nd is known (or at least the difference from the climatological CLF), such that $CLF|Nd = SHAP(Nd) * sigma_C LF + CLF_c lim$. *To me, this framing then makes it clearer that a more traditional sensitivity (dCF/dlnNd) would have to be calculated using dSHAP(Nd)/dlnNd and CLF scaling.*

Thank you for your insightful comments. We agree that providing a clearer explanation of this method will enhance the comprehensibility of the manuscript within the field of atmospheric sciences. Thus, we have rephrased Sect. 2.3.1 as follows:

(a) We agree the sentence the reviewer mentioned at lines 179 and 180 "It provides a novel ... to CLF changes" might lead to an incorrect first impression of SHAP values. Therefore, we have removed it.

(b) We have rephrased the sentence "The contribution of ... besides global feature importance "from line 182 to line 185 to distinguish between local and global explanations: "The contribution of a predictor value to a specific model prediction is calculated as the difference between the predictions of the model in the presence and absence of this particular predictor for all possible combinations of predictor values. Since this is performed at a "local" level (i.e. for this specific instance's prediction), it allows for insights into how a certain model outcome is achieved, thereby complementing more traditional "global" (considering all instances) feature importance measures (e.g. partial dependence plot)."

(c) Line 186: We have included an explanation of the base value "what would be predicted in the absence of any feature information".

(d) Line 190: We find explaining the base value as the climatological CLF to be appropriate and inspiring, and it helps the atmospheric science community more quickly understand the SHAP approach. We have incorporated it into the manuscript: "The base value could be analogous to the climatological CLF for a given geographical window assuming no information about the input parameters is known. In this context, the SHAP values of input features indicate the extent to which knowing information about each feature value would deviate the prediction from the climatological CLF (base value)." The structure of this section has also been altered accordingly.

3. *I am a little concerned by the very linear relationship in log space shown in Fig. 1b. Given CLF is capped at 0 and 100 %, many previous studies have found a non-linear relationship even in lnNd-space (e.g. Gryspeerdt et al, JGR, 2016). It is not clear to me how such a linear relationship here can be achieved, especially when using instantaneous daily cloud property data?*

Thank you for highlighting this important point for discussion. In fact, the very linear relationship shown in Fig. 1 (b) does not hold for all $5° \times 5°$grid cells. Here, we showcase some additional SHAP dependence plots where the points on the scatter plot do not form a distinctly linear relationship (see Fig. 1 in this document). Naturally, if we plot N_d without taking the logarithm, the nonlinearity will be much more apparent. The reason appears to be that, in certain $5° \times 5°$grid

boxes, the XGB + SHAP framework is unable to capture the nonlinearity of the relationship, despite the models being tuned for daily data at these specific grid boxes.

It seems that there is no relationship between CLF and lnN_d explicitly reported in log space in Gryspeerdt et al. (2016) that we can directly compare with. They only presented CLF–AOD relationships, mediated by N_d, in log space for their global patterns. However, it appears that their regional CLF–AOD, N_d–AOD and CLF–N_d relationships using joint histograms and probability distributions are not based on log-transformed N_d and AOD. We are unsure if the reviewer precisely referred to the nonlinear regional relationships by conditional probabilities and joint histograms, or the lnAOD–N_d–CLF sensitivity using linear regression in this study.

That being said, generally, the nonlinear relationships in Gryspeerdt et al. (2016) would suggest that the nonlinearity in this system might be better captured using joint histograms and conditional probabilities in the regions where the relationship is notably linear in our study. This makes sense because their method is designed to retain more information about the nonlinearity and report nonlinear relationships across different regions and cloud regimes. It is worth mentioning that the relationships in Gryspeerdt et al. (2016) based on conditional probability distributions were analysed for three $20^\circ \times 20^\circ$regions and one $25^\circ \times 25^\circ$region, which are larger than our $5^\circ \times 5^\circ$geographical windows. We could expect similarly nonlinear relationships by training XGB models for the same four regions. This is an interesting comparison for our future work.

After presenting these regional results, Gryspeerdt et al. (2016) subsequently showed the spatial patterns of the sensitivities as we do in our study. The sensitivity values on their maps are also calculated as the slope of linear regressions, meaning that some nonlinear and "convolved" relationships are disregarded as well. The main goal of our study is also to show the spatial patterns of sensitivity. The philosophy of our study is to use linear regression to capture as much of the relationship revealed by the explainable machine learning framework as possible. Therefore, in the grid boxes where the XGB models may not fully capture the nonlinearity, we can at least be confident that using simple linear regression reduces the loss of information from the XGB model.

We have now summarised and incorporated the discussion into Sect. 2.3.2 in the manuscript. Figure 1 in this document has also been included in the supplementary material as Fig. S1. The numbering of the figures in the supplementary material has been updated accordingly:

"It should be noted that the notably linear relationship in Fig. 1 (b) does not hold across all geographical windows. Fig. S1 displays additional exemplary windows where the relationships exhibit less linearity. Our approach also captures nonlinearity in the system; in these cases, the linear regression helps decrease the convolved relationships as in Gryspeerdt et al. (2016)."

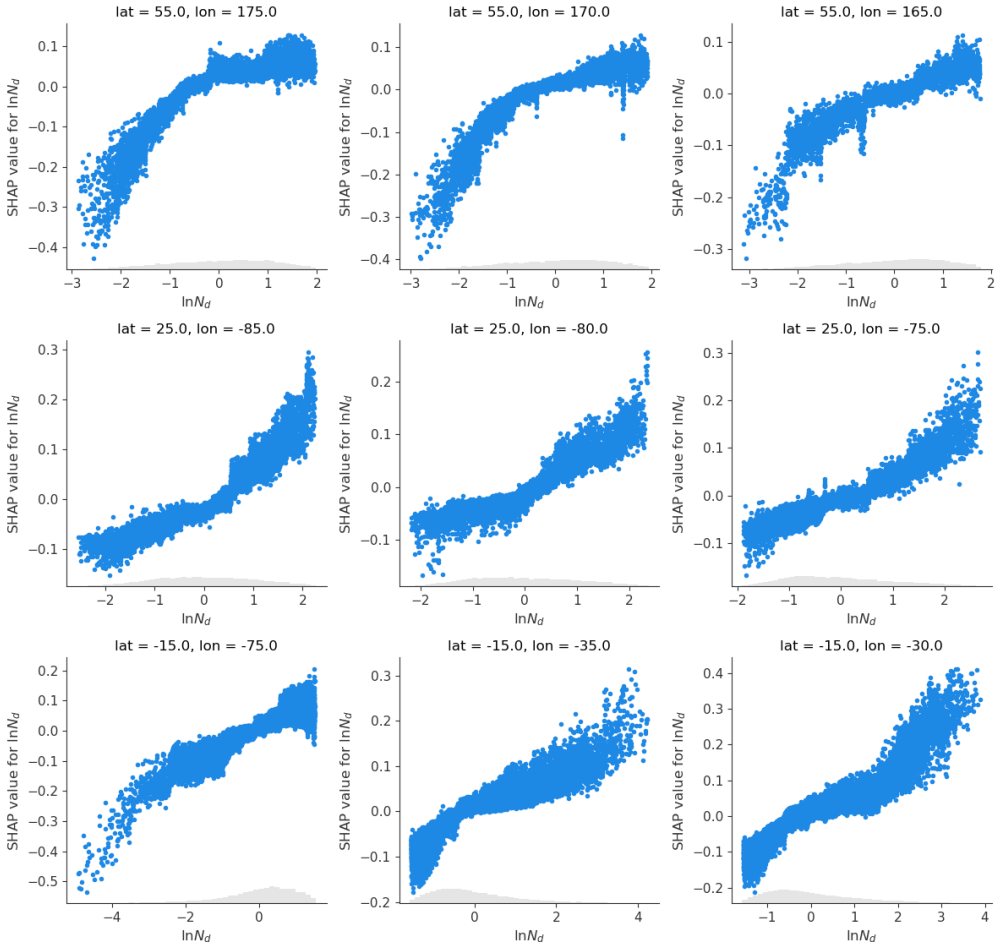


Figure 1. SHAP dependence plots similar to Fig. 1 (b) illustrating relatively nonlinear relationships between lnN_d SHAP values and feature values (standardized). The latitude and longitude values for each subplot represent the midpoint of each $5^\circ \times 5^\circ$ geographical window.

4. *There are good arguments for using the normalised values, particularly when comparing the different controls against each other. However, it is also important to be able to compare the results of this work against other studies, where the use of normalised values are less common. It would be good to have either a beta value that can be compared to the range in Bellouin et al., Rev. Geophys. (2020), or a forcing estimate that could be compared to those from other studies (e.g. Gryspeerdt et al, ACP, 2020).*

Thank you very much for your suggestions. We agree that there are studies where the CLF response to aerosol (proxies) is not directly compared with the quantification of the effects of meteorological cloud-controlling factors. Comparison with these studies will improve the scientific significance of this manuscript.

In the supplementary material, Fig. S4 shows the spatial patterns of the CLF-N_d sensitivity without standardization. In other words, the global average of these sensitivity values is comparable to the β value in Bellouin et al. (2020). We have added this comparison in Sect. 3.2.2 of the manuscript:

"The global weighted average of the CLF-lnN_d sensitivity without standardization is 0.112 (unitless), and its spatial pattern is shown in Fig. S4. This value is higher than the upper bound of 0.1 reported by Bellouin et al. (2020), which is based on global climate models and large eddy simulations. This may be partly due to the aforementioned bias. However, it is important to note that our non-standardized CLF-N_d sensitivity, shown in Fig. 1 (a), closely mirrors that from Yuan et al. (2023) with a similar range. In addition, the high lnCLF–lnN_d values estimated in Chen et al. (2022, 2024) suggest that values exceeding the upper bound of 0.1 might be plausible. These recent observational studies, including quantifying cloud fraction adjustment based on ship tracks Yuan et al. (2023), volcano aerosol perturbations (Chen et al., 2022, 2024), and our SHAP approach using global satellite observations, indicate that the 0.1 upper bound may be extended. In future work, estimating a radiative forcing using the SHAP-based sensitivities will make our study more comparable with other research on cloud fraction adjustment.

5. *I also found the mention/discussion of the potential non-causal Nd-CF link to be a bit lacking, especially around the headline results in the abstract and conclusions. While it is mentioned that there are potential retrieval biases in Nd as a function of CF, the impact of these is somewhat downplayed in the interpretation of the results. There are significant uncertainties in the retrieval of Nd in broken-cloud regimes, exactly where the observed Nd-CF relationship is strongest. While these results are clearly consistent with the theoretical behaviour of stratocumulus clouds, statements that CLF is 'sensitive to' Nd and that aerosol has a 'considerable impact on MBL cloudiness' might be over-attributing the causality of the results. This doesn't require a large change in the paper, just a little bit of rewording of some of the results.*

Thank you for your feedback. Although we have included a separate section to discuss the limitations, including the non-causal N_d–CLF relationship, and have rephrased sentences throughout the manuscript, we acknowledge that some statements may still be overly assertive. Therefore, we have revised the manuscript to provide a more cautious interpretation of the results, including updates to the abstract, method, conclusion, and results sections:

(a) Abstract, lines 10–11: We have revised the summary of the results concerning the N_d–CLF sensitivity: "Based on our statistical approach, global patterns of CLF sensitivity suggest that CLF is positively associated with N_d, particularly in the stratocumulus-to-cumulus transition regions and the southern hemispheric midlatitudes. However, N_d retrieval bias may contribute to non-causality in these positive sensitivities, and hence they should be considered as upper-bound estimates."

(b) Sect. 2.3.3, line 258: We have added "For example, the subpixel effect can introduce more bias in the N_d retrieval process within broken-cloud regimes due to increased heterogeneity. The N_d retrieval biases are ..."

(c) Sect. 3.2.2, line 342: "CLF is particularly sensitive to Nd in the regions of ..." has been reworded as "The relationship between CLF and N_d is found particularly strong in the regions of ..."

(d) Line 346: We have updated the sentence as: "However, as this cloud regime transition involves clouds shifting from more overcast to more broken, the strong relationships in these regions may be more affected by N_d retrieval errors."

(e) Sect. 4 Conclusions, line 466: The last sentence of this paragraph has been revised as: "The main findings of this study, which should be interpreted in light of the data and methodology limitations discussed in Sect. 2.3.3, are summarized as follows:"

(f) Sect. 4 Conclusions, point 1, lines 470 to 472: reworded as "The estimated N_d–CLF sensitivity and its magnitude suggest that aerosols likely have a considerable impact on MBL cloudiness, although this may partially result from an overestimation caused by the effect of a positive retrieval bias of N_d at high CLF."

Other minor suggestions

- *Is the temporal split of train-test data suitable, given the underlying change in climate? I doubt it would affect the results much, but could be something to consider for future work.*

  Thank you for bringing this up, we think this is a valuable point for discussion. We acknowledge that even within the relatively short period we studied (2011-2019), climate data from different periods may exhibit differences in data characteristics due to factors such as year-to-year variations and extreme weather events. We agree with the reviewer that different methods of partitioning data into training and test sets can influence the model, which is an interesting point for future investigations. Besides the simple chronological non-random split, strategies like the one proposed by Salazar et al. (2022), which handles spatial autocorrelation, could also be implemented in future work.

  The temporal train-test split has been commonly applied in many studies using machine learning to analyze aerosol-cloud-climate interactions (e.g. Fuchs et al., 2018; Dadashazar et al., 2021; Bender et al., 2024; Chen et al., 2024). However, some of these studies split the data sets by random shuffling, which may not be the best practice.

  In Earth sciences, data tend to be structured spatiotemporally and may have dependencies on nearby data points (Karpatne et al., 2017). A random train-test split can lead to temporally autocorrelated training and testing data from neighbouring

time steps, resulting in overoptimistic model performance evaluations (Roberts et al., 2017; Beucler et al., 2023). In addition, this random partitioning breaks the natural sequential order and results in data leakage, as the model already has information from the future while being trained (Malik, 2020; Kapoor et al., 2023). Since this "time travelling" does not align with real-world prediction scenarios, the test score cannot accurately reflect how well the model generalizes to unseen data. Therefore, the chronological temporal train-test split in our study, as done in Andersen et al. (e.g. 2023), is preferable to a random split for more realistic and reliable model performance. Furthermore, the 5-fold cross-validation used for hyperparameter tuning in our study is also done non-randomly to avoid this problem.

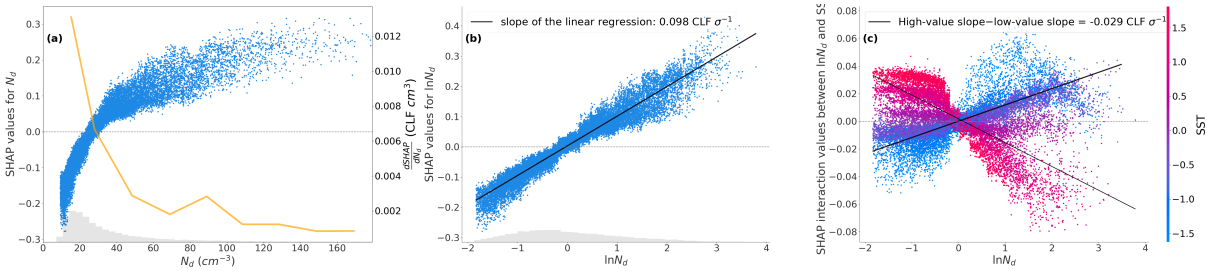
In Fig. 1, we examined how different training-test splits (80/20 %, 75/25 %, 50/50 %) affect the sensitivity and IAI for the exemplary window shown in Fig. 1 of the manuscript. Although we were unable to assess the global impact of different split ratios on the final sensitivities and interaction effects due to limitations in time and computational resources, our results from the example window indicate that varying the split ratios only slightly changes model performance (R^2). Importantly, the final results based on the SHAP approach remain consistent.

Accordingly, we have revised the sentence at lines 158-159 as "By chronologically splitting the training and test sets without random shuffling, we ensure that the training data will not see future information and the autocorrelation in data will not lead to overopstimic evaluation of the model's performance Beucler et al. (2023); Kapoor et al. (2023).

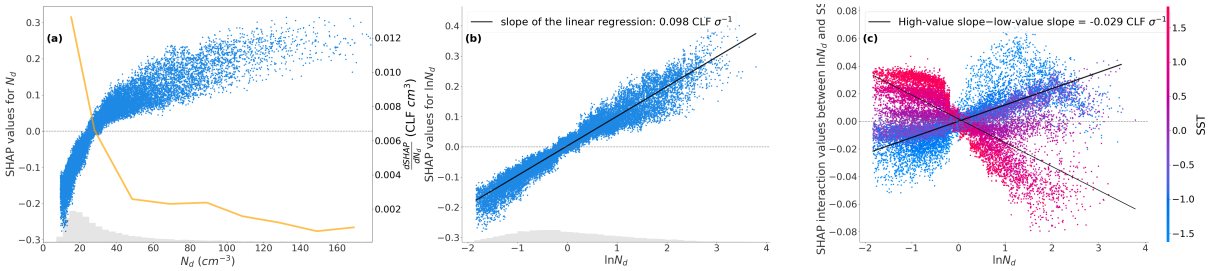
- *Are the Nd values calculated from the 1x1 degree tau-re values, or from the underlying joint histograms? The use of the large-scale mean values may lead to biases, due to the non-linearity of the Nd calculation. If it is useful, there is pre-filtered Nd data available for the period of study as a companion to Gryspeerdt et al., ACP, 2022.*

Thank you for raising this point. The N_d values are calculated from the $1^\circ \times 1^\circ r_e$ and τ_c values. We have acknowledged a caveat in the manuscript that using large-scale mean values in our study may bias our results. The existing N_d data sets by Gryspeerdt et al. (2022) appear to be good for our use. However, in addition to filtering criteria for solar zenith and satellite viewing angles, the "G18" sampling strategy also filters out pixels with a 5 km CLF smaller than 0.9 for more homogeneous clouds. Since we aim to analyse the relationship between CLF and the predictors, including only CLF values larger than 0.9 may not be appropriate because it would limit our analysis to a small portion of these relationships. In future work, N_d calculated based on underlying joint histograms could be a better choice. Exploring the impact of using mean values versus joint histograms on our quantification of sensitivities and interaction effects is another interesting point for the follow-up research.

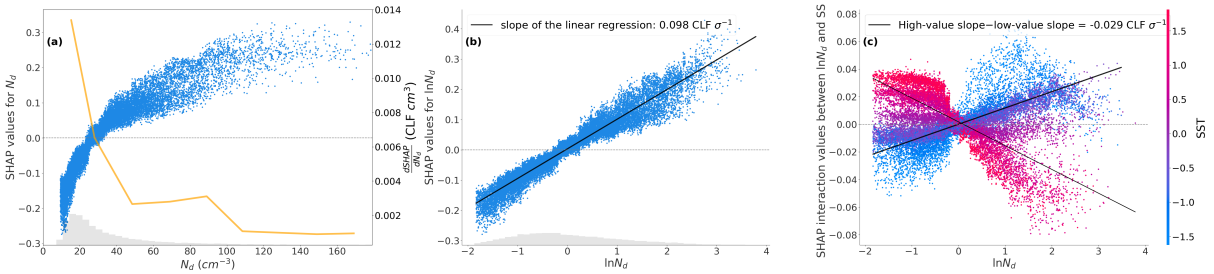
We have added into Sect. 2.3.3 Data limitation in the manuscript: "Another caveat in our data is that N_d values in our study are computed using MODIS level-3 large-scale mean r_e and τ_c values instead of joint histograms as in Gryspeerdt et al. (2016). This may introduce additional biases considering the nonlinearity of the N_d calculation. In future work, N_d data calculated from underlying joint histograms or pre-filtered data by Gryspeerdt et al. (2022) could be applied to be compared with the results in this study."



(a) Test size = 20 %, R^2 = 0.49



(b) Test size = 25 %, R^2 = 0.50



(c) Test size = 50 %, R^2 = 0.44

Figure 2. Examples illustrate that varying train-test split ratios affects only the test score R^2, but does not alter the quantification of sensitivities.

Minor modifications independent of the reviewer comments

Line 161: “As suggested by Karpatne et al. (2017), a single ML model may not perform well across all regions due to the heterogeneity of relevant processes. Therefore, data ...”

Line 161: “regionally-specific relationships” to “regional relationships”.

Line 187: “data points” to “data set”

Line 228: “Figure. S1” to “Fig. S2”

Line 339: “the” has been inserted between “of” and “MBLC”.

Line 348: “AOD” to “aerosol optical depth”

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.

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., Boucher, O., Carslaw, K. S., Christensen, M., Daniau, A. L., Dufresne, J. L., Feingold, G., Fiedler, S., Forster, P., Gettelman, A., Haywood, J. M., Lohmann, U., Malavelle, F., Mauritsen, T., McCoy, D. T., Myhre, G., Mülmenstädt, J., Neubauer, D., Possner, A., Rugenstein, M., Sato, Y., Schulz, M., Schwartz, S. E., Sourdeval, O., Storelvmo, T., Toll, V., Winker, D., and Stevens, B.: Bounding Global Aerosol Radiative Forcing of Climate Change, Reviews of Geophysics, 58, 1–45, https://doi.org/10.1029/2019RG000660, 2020.

Bender, F. A., Lord, T., Staffansdotter, A., Jung, V., and Undorf, S.: Machine Learning Approach to Investigating the Relative Importance of Meteorological and Aerosol-Related Parameters in Determining Cloud Microphysical Properties, Tellus B: Chemical and Physical Meteorology, 76, 1–18, https://doi.org/10.16993/tellusb.1868, 2024.

Beucler, T., Ebert-Uphoff, I., Rasp, S., Pritchard, M., and Gentine, P.: Machine Learning for Clouds and Climate, in: Clouds and Their Climatic Impacts, Geophysical Monograph Series, pp. 325–345, https://doi.org/https://doi.org/10.1002/9781119700357.ch16, 2023.

Chen, Y., Haywood, J., Wang, Y., Malavelle, F., Jordan, G., Partridge, D., Fieldsend, J., De Leeuw, J., Schmidt, A., Cho, N., Oreopoulos, L., Platnick, S., Grosvenor, D., Field, P., and Lohmann, U.: Machine learning reveals climate forcing from aerosols is dominated by increased cloud cover, Nature Geoscience, 15, 609–614, https://doi.org/10.1038/s41561-022-00991-6, 2022.

Chen, Y., Haywood, J., Wang, Y., Malavelle, F., Jordan, G., Peace, A., Partridge, D. G., Cho, N., Oreopoulos, L., Grosvenor, D., Field, P., Allan, R. P., and Lohmann, U.: Substantial cooling effect from aerosol-induced increase in tropical marine cloud cover, Nature Geoscience, https://doi.org/10.1038/s41561-024-01427-z, 2024.

Dadashazar, H., Painemal, D., Alipanah, M., Brunke, M., Chellappan, S., Corral, A. F., Crosbie, E., Kirschler, S., Liu, H., Moore, R. H., Robinson, C., Scarino, A. J., Shook, M., Sinclair, K., Thornhill, K. L., Voigt, C., Wang, H., Winstead, E., Zeng, X., Ziemba, L., Zuidema, P., and Sorooshian, A.: Cloud drop number concentrations over the western North Atlantic Ocean: seasonal cycle, aerosol interrelationships, and other influential factors, Atmospheric Chemistry and Physics, 21, 10 499–10 526, https://doi.org/10.5194/acp-21-10499-2021, 2021.

Fuchs, J., Cermak, J., and Andersen, H.: Building a cloud in the southeast Atlantic: Understanding low-cloud controls based on satellite observations with machine learning, Atmospheric Chemistry and Physics, 18, 16 537–16 552, https://doi.org/10.5194/acp-18-16537-2018, 2018.

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.

Gryspeerdt, E., Mccoy, D. T., Crosbie, E., Moore, R. H., Nott, G. J., Painemal, D., Small-griswold, J., Sorooshian, A., and Ziemba, L.: The impact of sampling strategy on the cloud droplet number concentration estimated from satellite data, pp. 3875–3892, 2022.

Kapoor, S., Cantrell, E., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G., and Narayanan, A.: REFORMS: Reporting Standards for Machine Learning Based Science, https://arxiv.org/abs/2308.07832, 2023.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V.: Machine Learning for the Geosciences: Challenges and Opportunities, https://arxiv.org/abs/1711.04708, 2017.

Malik, M. M.: A Hierarchy of Limitations in Machine Learning, https://arxiv.org/abs/2002.05193, 2020.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/https://doi.org/10.1111/ecog.02881, 2017.

Salazar, J. J., Garland, L., Ochoa, J., and Pyrcz, M. J.: Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy, Journal of Petroleum Science and Engineering, 209, 109 885, https://doi.org/https://doi.org/10.1016/j.petrol.2021.109885, 2022.

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, https://doi.org/10.1126/sciadv.adh7716, 2023.