Review of:

„Possible evidence of increased global cloudiness due to aerosol-cloud interactions“

by Alyson Douglas and Tristan L‘Ecuyer

General comment

In the manuscript named „Possible evidence of increased global cloudiness due to aerosol-cloud interactions“, the authors use machine learning (ML) models trained on a low-aerosol subset of the observational data set to predict warm cloud fraction in these pristine conditions on the basis of four meteorological parameters. The ML models are then applied to predict warm cloud fraction in polluted conditions and used as a pristine aerosol counterfactuals, where the difference between these counterfactuals and the observed warm cloud fraction is quantified as ΔCF and assumed to be due to aerosol-cloud interactions (ACI). The authors then use the ΔCF to estimate an effective radiative forcing due to aerosol-driven changes in warm cloud fraction, which is found to be -.42 W m^2, in line with other recent estimates.

The topic of constraining ERFaci and the cloud fraction adjustment from observations is highly relevant for climate research (and for the readers of ACP), and as progress constraining ERFaci has been limited over the past decades, new methods of quantifying ERFaci (e.g. exploiting novel ML techniques) should in my opinion be encouraged. While the underlying idea of the authors falls into this category and clearly has merit, I see a number of critical issues and unclarities with the research presented here, and the manuscript leaves the impression of being put together hastily (unit missing, equation missing etc.). Some of the critical issues are

1) sampling issues: The data sets are divided into a clean and a polluted data set, but only trained on the clean data. This is problematic in regions where the aerosol features e.g. a clear seasonal cycle, as data from the polluted season will not be included or underrepresented in the training data. As CF and its meteorological controls have seasonalities as well, realistic predictions of CF during the polluted season can not be expected if this season is largely omitted during training. This is particularly the case, as ML models completely incapable of extrapolation are used. I will go into detail into issues with data sampling below.

2) If I understand the manuscript correctly, ΔCF is quantified for the polluted conditions only (trained on clean, applied to polluted situations?), however, as the clean conditions are also part of today’s aerosol distribution omitting them in the ΔCF estimation would lead to an overestimation of the overall ERFaci, as ΔCF would be about 0 in the clean conditions.

3) Lack of transparency/details: The manuscript lacks relevant information into how the models are set up (hyperparameters, sampling strategy of the train/test split). This is important, because this affects the reproducibility of the research, but also because it hides possible reasons for the inexplicably high skill of the ML models in the testing with the independent data.

These aspects are likely to influence the results and conclusions drawn in this manuscript, and as such I cannot recommend this manuscript to be published before major revisions and additional tests are completed.

Major points

- Sampling biases
  - Aerosol seasonality: As mentioned in the general comment, I am quite worried about the ML models ability to quantify ΔCF and ERFaci in all regions with a pronounced aerosol seasonality. One example for this would be the Southeast Atlantic, where seasonally overlying aerosol plumes from biomass burning in central Africa are a common feature between July and September/October. During this time of year, CF is highest (stability is highest and SSTs are lowest). If aerosol seasonality is not explicitly considered during the data split, it is likely that training data during the polluted season is sparse, and the highest CF conditions not represented in the training data. It can not be expected that the models perform well in predicting prestine counterfactuals in a season that is underrepresented during training. Considering these aspects, I would expect this issue to be amplified in the Southeast Atlantic, and hence I am not surprised to see this region as an outlier with respect to the estimated ΔCF and ERFaci. I do not agree with the interpretation of the authors given in L.223 – 236 or other text passages in the manuscript interpreting this region.
  - On a similar note, sampling representativeness between clean and polluted data will also be an issue in any coastal region where the aerosol loading is strongly correlated with the outflow of continental aerosols (i.e. dynamics). Low aerosol conditions are unlikely to provide representative environmental conditions to high aerosol conditions in such regions, the ΔCF estimation (and ERFaci) would seem to be less robust here. The authors discuss this a bit in L.195 (f), but discard this issue and it remains unclear to me why they believe this is not a problem. Overall, this study needs to more clearly address and discuss these issues, as they only describe the uncertainties their method reduces but not the ones it introduces.
- Lack of transparency/details, unrealistic model performance: It is really unfortunate that the authors do not provide the necessary methodological details to be able to reproduce and fully understand their workflow and results. Even though the authors consider their machine learning methods to be “simple”, their setup is not trivial and its details can influence the results. It is necessary that the authors provide the exact model settings (i.e. hyperparameters), how they are determined, but also provide details on how training and test data is split. This is especially important, as the models perform much better than one would expect for such a complex problem (predicting CF). 98-99% explained variance seems completely unreasonable, and a clear sign of the models overfitting to the data. For example, the explained variance in Chen et al. (2022, nature), who use random forests to predict large-scale CF with **114** meteorologic parameters is 56%. The authors cite a study with comparable skill, however, in that study a highly nonlinear polynomial fit of the fourth power (known to overfit data) is applied to highly aggregated (binned) data, and no independent test skill is reported, so this is just another example of models overfitting the data. Also, the skill increase from the MVLR (0-40% explained variance in the MVLR and > 98% in ML) is suspect to say the least. In my experience, a much smaller performance increase from MVLR to decision trees is to be expected, and this is backed by e.g. Fuchs et

al. (2018, ACP), who report an R^2 between 0.45 and 0.7 for the decision trees and 0.3-0.5 for an MVLR, and Dadashazar et al. (2021, ACP), who report an R^2 in predicting cloud droplet number concentration between 0.43 and 0.47 for the decision trees and 0.25-0.28 for an MVLR. One should note that the MVLR results are fairly close to what is reported in this study, however, the ML results are obviously not.

I can speculate that the reason for these high model skills could be that splitting training and test data is done randomly (and not in two completely separated time periods (e.g. 2 years training, 1 year testing)), and that the models then learn the training data by heart. Because training is done on 80% and testing on 20% of the data, and autocorrelation of CF and the predictors is high at the daily time scale, the task becomes very easy for the decision trees, as the test data is not independent from the training data and they are very powerful in learning data by heart. However, this remains speculation, because the authors have not added the necessary information on the model setup in the manuscript.

This (very likely) wrong skill estimate is then used to confirm the hypothesis that "a majority of the variation in clean, "pristine" aerosol scenes are due to the environment" as "the explained variance scores would be lower as variation due to aerosol would not be learned by the models" (L.304 and L.308-309), which in general I would agree with, but not based on the research presented here. It would be much easier to trust the results if the model setup were described completely, and the training and test split done as described above (and considered to be good practice for temporally structured data; Roberts et al. (2017, ecography, doi: 10.1111/ecog.02881)). Additionally, it would be appreciated if the authors would do an additional test on how well the models can predict polluted situations in the year left out for testing.

- It is unfortunate that the authors do not provide any detailed information on the aerosol (MODIS AI) thresholds derived from the SPRINTARS model. The authors only vaguely state that "We choose a selection of scenes with a distribution of AI similar to the pre-industrial values for that region from SPRINTARS [...] (L.103-104)". As the definition of AI values interpreted as pre-industrial proxy is critical for the results, it is necessary that the authors provide a precise statement on how this is determined and ideally a map of this threshold, and information on how much data is in the prestine and polluted groups.
- If I understand the method described in L.150 correctly, ΔCF is only estimated for the "polluted" situations that should represent present-day conditions. However, I don't think that is reasonable, as the „prestine" conditions used as a proxy for pre-industrial conditions are still part of the aerosol distribution in the present-day. Neglecting this leads to an overestimation of ΔCF, because ΔCF should be close to 0 for the „prestine" conditions and thus decrease the overall average ΔCF. Will this make a significant difference in the end? There is no way for the reader to know, as the AI thresholds used are not provided, and hence it remains unclear how much data is assigned to the prestine and polluted groups in each region.
- I see the interpretation of the ML models as „True" and the MVLR as "False" (e.g. L. 182) to be problematic. It is also unfair to compare sorted regional ΔCF values of the 3 ML models to the MVLR to show how similar the ML models are compared to the MVLR (Fig. 11). Clearly the 3 ML models chosen in this study are closely related in the way the learn

(all are ensembles of decision trees), and hence they are expected to more or less lead to the same learned patterns. It would be more interesting to use e.g. a simple neural network as an additional comparison, as this represents a different way of mapping the model input to the output than the decision trees. Also, a simple neural network does not have similar overfitting issues, as it cannot learn training data by heart as easily as the decision trees.

- The authors hypothesize that a “good” ML model will be 1) physically realistic, and 2) observational scale independent. In my opinion, the authors do not show convincing results to support either hypothesis. 1): The only results that can be interpreted as a sign of the models being physically realistic is presented in Fig. 10. However, in the text this is treated as an outlook, and it is not even clear in which region the results were produced (“example”). It is certainly not physically realistic that 4 large-scale environmental controls explain more than 90% of daily CF variability. 2): The authors use two different initial cloud fraction scales (12 and 96 km) which are then both aggregated to 15°x15°. The analysis is then conducted on the same scale, and all predictor data seems to be identical as well. While this is an interesting experiment, it does not seem to be a convincing argument that the results are scale independent.
- While this manuscript is focused on the cloud fraction adjustment to aerosols, this is not reflected by the introduction, where the cloud fraction adjustement and current approaches to quantify it and its radiative forcing are not discussed. Actually, the LWP adjustment is discussed in much more detail. I recommend that the authors provide a better overview of past observational research done on the cloud fraction adjustment and provide sources for the statement in L. 25.

Minor points

- L.8: units missing for the cooling estimate
- L.28-31: Cloud fraction retrieval is also not straight forward, as it depends on an optical depth threshold and is affected by surface reflectance etc.. A change in optical depth due to the Twomey effect could thus also lead to an artificial increase in CF for thin clouds (Mieslinger et al. 2022, ACP).
- L157: Equation is missing
- L.171-174: Sentence is quite hard to read.
- L.176-178: Confusing sentence
- L.229-231: I don’t quite understand why only cloudy scenes are used, and don’t understand the authors reasoning given here: “we do not want to possibly conflate cloud feedbacks (how cloud formation has changed due to global warming) with aerosol-cloud interactions (how aerosol loading has increased or decreased cloud extent.)[sic]”, as cloud feedbacks are not constrained to changes in cloud formation, but rather describe the change in overall cloudiness (could also be due to a longer lifetime or larger spatial extent of the clouds).
- L.231-232: Is there a source for the statement that the observational record includes a reduction in warm cloud cover in the south Atlantic? If not, I don’t understand how the authors can claim that “our models have *succeeded in identifying* a reduction in warm cloud cover in the south Atlantic”. The authors then go on with this claim a few lines further

down: “It is possible in future climates that more regions will experience the same conditions that have led to a reduction of cloud fraction seen at this south Atlantic region”

- L.233-234: Technically, this is not true for all machine learning models, as some do have the capability for a limited extrapolation (e.g. MVLR, simple neural networks), but this is true for tree-based ML models which can only interpolate on the training data.
- L.319: I don’t undestand what the authors mean with this sentence.
- L.348-351: The authors use the average model predicted cloud fraction of the prestine situations used for training (Fig. 8 and 9) and interpret this as a sign that the models are consistent with each other and represent physics internally (in the conclusion). This is not related to physics or a sign of consistency of what the models represent at all though, but just the fact that any statistical or ML model will quickly learn the average predictand value during training. I believe that these Figures and the discussion on it (also in L. 308-310 and the conclusions) add nothing to this paper and should be removed.
- Figures 2-5: I would recommend that the range on the colorbars should be equally far positive/negative, as one naturally compares the hue of the colors, but in these figures the change of the hue for positive/negative values is different by a factor of 4.
- The difference between ΔCF of the Northern and Southern Hemispheres is interesting, and I think this should be discussed with more detail. I would especially be interested if there is a hemispheric difference between the AI threshold used in this study.

Specific points

- L.37: „environment“ is used as a term for meteorological factors (excluding aerosol), however this is not clear as aerosols are also a component of the environment
- L.52: to constrain
- L.54: Adjust citation style
- L.75: 2 Methods 3 Data
- L.134: Grammar
- L.146: „split into by 20%“
- L.235: „until our we“
- L.321: EIS has not been introduced so far