General Comments

The authors trained the models with 4 parameters: boundary layer stability, relative humidity, vertical motion, and sea surface temperature. Why did they use these 4 parameters? Are they relevant for all the different regions? I think that there are too few parameters to determine the cloud fraction.

To help clarify to the reader before the methods section, our particular set of environmental predictors is introduced, we have added to the introduction:

"In particular, the stability of the atmosphere, often indicated by the estimated inversion strength or lower tropospheric stability, and the humidity of the free atmosphere that the boundary layer clouds entrain strongly control the sensitivity of the cloud fraction to aerosol perturbations (Gryspeerdt et al. 2016). Further, others have indicated the presence non-linear, sudden transitions of warm clouds due to their dependence on a strong inversion and cool surface temperatures to maintain equilibrium (Schneider et al. 2019)."

These parameters have been shown to be highly correlated with cloud fraction in past studies, such as EIS explaining 70% of the variance in major stratocumulus regions globally (Wood & Bretherton, 2006). As mentioned within the Methods section, all four parameters are known to be cloud controlling factor (Klein et al. 2017). Given the models show high explained variances and low mean squared errors on the test data, it is unnecessary to add additional parameters to our dataset. Further, the more environmental variables the models are trained on, the higher the likelihood of violating the positivity assumption, whereby for every possible environmental regime, there must exist a corresponding counterfactual representation of that regime. By limiting our parameters to the strongest cloud controlling factors, we are reducing the likelihood of violating this assumption.

Do they consider all clear sky pixels around the clouds to do the interpolation, or the closest clear-sky pixels to the considered cloudy pixels? If the clouds are horizontally extended, is there a maximum distance between the cloudy pixel and the clear-sky pixels?

We remove the AI nearest the clouds (within 2 km) as mentioned within the Methods section. This reduces the chances of hygroscopic growth from influencing measurements of AOD or the Angstrom exponent. Assuming the small likelihood that clouds are affecting AOD measurements beyond 2 km however would not substantially change our results as we do not use AI as a quantitative metric beyond delineating "clean" from "polluted" environments, and those affected scenes would then more likely be delineated as "polluted," perhaps incorrectly. The predicted cloud fraction for these scenes would most likely match the observed cloud fraction, as our cloud fraction estimates are based on active retrevials from CloudSat and will not be biased by near cloud hygroscopic growth of aerosol. Hygroscopic effects are most likely to alter results when AI is used as a direct, quantitative proxy for CCN to define sensitivities, as this can bias the high AI conditions.

All the considered results are shown averaging over time and space, are the models still good with a single prediction? There is no measure to determine if the results are statistically robust or not even on the figures. Are all pixels show a statistically significant change in CF?

Decision tree methods are meant to be used for tabular data, such as our satellite snap shots. Our data is not cross-sectional, as that would imply we use only a single time for each point. Instead, inherently within our data, best captured by the environment, are the fluctations over time. As the satellites pass over each point, they add a new shap shot of warm clouds for us to analyze. By including four years of satellite datea (2007 - 2010), we can then more robustly understand these variations as a function of the environments they occur in, environments which change over time.

All of the simple ML models showed statistically significant differences, however that is not the type of metric that should be used to validate a study like this. If the predictions were incorrect, then the results could still be statistically significant as determined by a single p-value. We chose to show the explained variance scores and mean squared errors to justify that our models were correctly capturing the behavior of clean clouds and that the error in those predictions is smaller than the predicted changes in cloud fraction.

The authors performed the training of the models for different regions. Some regions are constantly under the influence of anthropogenic aerosols or maybe not after specific events (e.g., precipitation) but is it safe to consider this as pre-industrial cases? Also, the regions are never properly defined and I do not know what they refer to.

We have added additonal clarification of how our assumptions may leade to uncertainty in the methods and results section.

The effects of aerosol on precipitation occurrence, duration, or severity remain uncertainty; to remove any uncertainty of aerosol-cloud-precipitation interactions we only consider non-precipitating scenes within our analysis. We consider our predicted, clean cloud fractions as a proxy for pre-industrial, however we acknowledge several times within our manuscript that this has limitations. We do not consider what we are predicting as the true pre-industrial cloud fraction for each region, as the global environment, including large scale dynamics determining local meteorologies, has changed due to global warming. However, our proxies are appropriate to create a base estimate of how cloudy the Earth may have been. As stated within the discussion, we believe our technique is a more accurate method than what is currently being done, i.e. using the product of linear sensitivities and the assumed, regional change in AI (or other CCN proxy) to find the overall change due to aerosol-cloud interactions.

The regions are defined in the Methods section under Data and under Methodology. Additional references to what the regions refer to (the 15° x 15° aggregation grid of observations) have been added throughout the manuscript to clarify their resolution.

The authors state that they assumed a constant cloud albedo and therefore cannot account for the Twomey effect. Therefore I am wondering why estimating the radiative forcing at all?

This is correct that we do not calculate a Twomey effect, as we do not compare the albedo of the polluted vs. clean or predicted clean to find how the albedo has increased due to aerosol interactions. The Twomey effect is only one portion of aerosol-cloud interactions that leads to an increase in forcing. An increase in cloud amount will also leading a change in forcing, as more cloud will then reflect more incoming sunlight, cooling the Earth.

We are able to accurately calculate this forcing by using the CERES fluxes. The radiative forcing at the top-of-atmoshpere from CERES inherently includes cloud albedo as brighter scenes -> a higher SW TOA.

```
SW TOA = SW CRE + SW CLR
whereby when fully expanded
SW CRE = Incoming Sunlight x Cloud Albedo x Cloud Fraction
```

Using the SW CRE from CERES includes any albedo changes due to aerosol.

In the figures, I do not understand why the resolution of 12km and 96 km have not been plotted instead of the 30 degree x 30 degree boxes. Also the changes in cloud fraction are shown but I am wondering if the changes are statistically significant for every box (see point 5).

The changes are statistically significant for every region (p << .05). CloudSat is an active sensor with a native resolution of ~ 1 km x 1 km. In order to define a "cloud fraction," you must choose a resolution to average cloud occurrence over. Therefore, to understand how this choice of resolution (aka aggregation of observations over a chosen scale) influences either the training/predictions of the ML models or the estimated change in cloud fraction, we compare our observations averaged over 12 km along track and 96 km along track. These along track observations are then re-aggregated over 15° x 15° regions in order to train regional models and create regional predictions of the "clean" cloud fractions (at both 12 km and 96 km resolutions).

I think the ML techniques are useful for a deeper analysis as presented in Figure 10 to understand the correlation between the different parameters. Otherwise I feel that this paper is presenting a promising method (then ACP might not be the best journal) with interesting results on CF but I really would like to see an explanation on the reasons for the CF changes, if one parameter is more important than others, if it is the same for all regions...

We agree that understanding the different cloud controlling factors and their interactions with each other is important and have already written up our additional, related work to JGR: Atmospheres (*in review*). However, this work we believe does fit within the bounds of ACP as other similar articles (such as Fuchs et al. 2018) and new methodologies (Gryspeerdt et al. 2021) that deal with understanding clouds, aerosols, and the envrionment have been published before.

I am not sure the study mentions exactly what a region is, is it a 30x30 degrees region ?

This is detailed in the Methodolgy. We have added additional reminders of the gridding resolution to the results section to help clarify our regional resolution.

Machine learning models require some parameters as input, for example the depth of the tree, boosting iterations, learning rate for random forest regressor. The current study does not mention the parameters used and how they are decided. Some details are required.

We have added the details of the hyperparameter tuning for each model. As these are all decision tree based models, we tried to keep the depth consistent. Some models (SGB, XGBoost) allowed for sub-sampling, which we utilized in order to help evaluate if the Random Forest was overfitting (as the SGB or XGB would then have shown different results). We have added to the Methods section: "The RF hyperparameters include having bootstrapped training, meaning the RF is trained on a random subset of observations over each iteration, the number of iterations was limited to 125, and the depth limited to 30. The SGB hyperparameters are a max depth of 30, a learning rate of 0.1, a minimum number of samples per leaf as 3, subsampling at 0.8, and the number of iterations as 200. The XG hyperparameters are max depth of 30, subsampling of 0.8, 'GBTree' as the booster type, a histogram tree method, a learning rate of 0.8, and the max number of iterations to train as 300. These hyperparameters were chosen by using a grid search over each model using as subset of the total, global data. "

Minor comments

Specifics:

Line 1: Uncertainty is also gramatically correct.

Line 15: Aerosol can be used in place of aerosols in this context.

Line 15: Sentence now reads "Aerosol enters a cloud and acts as cloud condensation nuclei (CCN), increasing the total number of cloud droplets."

Line 25: Aerosol can be used as is in this context.

Line 54: Fixed, removed paranthesis.

Line 106: "in" added to sentence.

Line 122: Changed to affect.

Line 145: Removed "a."

Line 254: Fixed, changed wording to "proceed with."

Added new IPCC report reference, specifically as ACI are now contained within a "short lived forcers" section.

Fixed section headings between Introduction and Methods.

Added additional citations for data used, specifically the MODIS Deep Blue Aerosol product and AMSR-E Sea Surface Temperature validation.

Yes, we omit multilayer scenes and focus only on single layer, marine warm clouds. We have added a more clarifications of this to Data section.

line 156: We have removed warm from the equation and added multiple clarifications that our cloud fractions are the single level warm cloud fractions within each region.

Equation 2: Fixed.

line 169: We agree and have rephrased that statement to read: "To a first degree, warm clouds dominate the cooling due to ACI globally; however polar low clouds could potentially also have some small cooling effect (Christensen et al. 2016, Rosenfeld et al. 2014)."

L.190 You are correct, this was an error. We have added when specifying the hyperparameters of each model: "Subsampling over each iteration, as the XG and SGB allow, reduces the chances of overfitting as this forces the trees to generalize; the RF uses random sampling over each iteration for similar purposes."

The part “Validation of ML Model Results” arrives at the end of the paper. Should it not be more fitted at the beginning of the result section? Or it should be stated that a discussion on the validation is later when presenting the method.

We have rephrased the section to "Validation of ML Models against Our Criteria" as we set out two specific criteria in the introduction that we would use to judge our results (resolution independence and agreement between models). The focus of the study is two folds, therefore the results have two focuses: the science focus of how aerosol may have increased cloudiness and the ML focus on how to create a set of accurate models agree with each other.

Figure 6: That is .4% not 40%. The zoomed in portion shows values below 1%, hence all percentages are a fractional percentage.

From lines 301 to 310: We have rephrased the paragraph. The core of the paragraph now simply reads:

"The explained variance scores not only lend credence to the simple ML models, but to our methodology, as a core assumption of our methodology was that the environment in "clean" scenes could explain a majority of the variations in cloudiness. "

Figure 11: We have updated the caption to read: "The change in cloud fraction (y-axis) for each region sorted by their similarity (x-axis) at 96 (top) and 12 km (bottom) resolutions . The change in cloud fraction is weighted by the regional occurrence of warm clouds. The similarity is found by sorting the median change in cloud fraction from the ML models from smallest to largest values. This allows us a unique viewpoint on how the ML models compare to the MVLR."

References

Klein, Stephen A., et al. "Low-cloud feedbacks from cloud-controlling factors: A review." Shallow clouds, water vapor, circulation, and climate sensitivity (2017): 135-157.

Wood, Robert, and Christopher S. Bretherton. "On the relationship between stratiform low cloud cover and lower-tropospheric stability." Journal of climate 19.24 (2006): 6425-6432.