Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2020-509-RC2, 2020

Atmospheric Chemistry and Physics
Discussions
Open Access

Interactive comment on "Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements" *by* Arshad Arjunan Nair and Fangqun Yu

Anonymous Referee #2

Received and published: 3 August 2020

This study demonstrates the use of a random forest regressor (RFR) to predict cloud condensation nuclei (CCN) concentrations at 0.4% supersaturation from features derived from chemical-transport model output (GEOS-Chem-APM) and measurements from a surface monitor. The authors train a RFR using GEOS-Chem-APM output and compare these results to observations at the SGP monitoring site. They then re-train models based on measurements available at this site. While I found the approach promising and the subject matter (using a combination of modeling techniques to predict CCN) fits in well with the subject matter at ACP, I feel that the scientific goal of the

manuscript in its current form is not sufficient to recommend publication.

General comments:

My main concern with this manuscript is in framing the results as a "proof of concept". Given the current state of literature for CCN and machine learning techniques, I do not think a proof of concept is sufficient for publication (for instance, there are a number of previous studies cited here that make use of machine learning techniques for atmospheric science). Further, as a proof of concept, the comparison to observations is limited to one surface site. Could the comparison to observations be expanded to include more sites? This paper could be strengthened by applying this technique further than what the authors did in the present study. The suggestions the authors made in Section 4.2 sounded promising. I especially thought the suggestion at Line 396 (using this approach in models with only bulk aerosol schemes) was a good idea. For instance this paper could be strengthened by: training the model with GEOS-Chem-APM, predicting the CCN based on the standard (bulk microphysics) GEOS-Chem, and then comparing to a larger suite of measurements (perhaps including ATom or sites with an SMPS in Europe). This way the authors could demonstrate if the random forest model provided an improved estimate in CCN over the approach used with the standard GEOS-Chem model. (Note, I am not suggesting the authors follow this suggestion exactly but I am trying to demonstrate how I feel the argument of this paper could be strengthened). Overall, the authors did a detailed job in training the RFR but the use of it is limited to comparing to only one surface site. How robust is the model if we only see it compared to one site?

Similarly, the discussion in Section 3.2.2 (Lines 316 and onwards) about training new models with different features feels somewhat unclear as to what the goal is. This is related to my comment above about the framing of the study. Is the goal to use actual observed pollutant and meteorological measurements to predict CCN or to use the same values used in GEOS-Chem and provide an alternate GEOS-Chem estimate? It seems problematic to have one model trained on global GEOS-Chem-APM output and a second based on observations at one site (unless the goal here is to compare why the two predictions are different?).

Specific comments:

Section 3.2.2: I think it would be useful to compare the performance of the RFR model relative to the observations at SGP and GEOS-Chem-APM at SGP. Does the RFR perform as well at predicting CCN as GEOS-Chem-APM? Or is the decrease in accuracy reasonable given the improvement in computation time of RFR over a microphysical model?

Hyperparameter tuning (Line 215): Why were these three hyperparameters chosen to be tuned as opposed to the other possible hyperparameters (such as max tree depth)? This section is often framed as balancing accuracy with computational cost, but isn't overfitting also a large concern here?

The writing in Section 2.2 can be improved. First, the paragraph on the random forest technique should be better cited. In addition, while a random forest can improve overfitting issues over a single decision tree, it certainly does not completely correct overfitting. Finally, I do not think the paragraph at Line 100 (listing different techniques and supervised/unsupervised learning) is necessary.

The first two paragraphs of the introduction can be cleaned up. For instance:

- Line 2 reads "these particles or aerosols, or rather CCN", equates aerosols to CCN which is not technically correct.

- The word "particles" should be used instead of "particulates".

- Line 25: I suggest changing "directly get into the atmosphere" with "emitted".

- Line 34: "Aerosol-cloud interactions are through CCN" is vague ("depend on"?). This also neglects to mention the intermediate step of cloud droplet number concentration (the authors point is still valid).

Technical comments:

Line ∼135: The paragraph citation of Nair et al. (2019) does not seem needed here, especially since this paragraph cites two other papers.

Line 209 "The remaining data..": Does the word 'data' here refer to GC model output or observed data?

Line 327: A tau of 0.36 is pretty low, correct? This seems like a notable decrease in model accuracy.

Line 353: Minor typo (subscript 2.5)

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2020-509, 2020.