Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2020-509-RC1, 2020

Atmospheric Chemistry and Physics Discussions

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 #1

Received and published: 28 July 2020

The paper proposed a Random Forest Regression Model (RFRM) based on machine learning to derive [CCN0.4] number concentrations from commonly available measurements (8 fractions of PM2.5, 7 gaseous specie, and 4 meteorological variables) over varying spatial and temporal scales. The CCN number concentrations is an essential task for environment evaluation and can be used for many different applications. The optic of this paper is interesting and it is valuable to investigate. The author explained the detailed data acquisition and processing steps of the proposed model. The suggested RFRM is trained on the long-term simulations in a global size-resolved particle

mi-crophysics model and can be applied to any area of the world. The experimental results demonstrate robustness of the proposed method.

Also, there are still serval problems that should be answered by the authors for a better understanding of the paper.

In the process of geospatial analysis, it is essential to ensure that all the relevant elements are at the same or very approximate temporal stamps. How do you confirm the measure data at the SGP site (Meteorology and Chemical Species) are all in the same temporal scale? No specific detailed explanation was found in section 2.4.1([CCN0.4] measurements) and section 2.4.3 (Atmospheric state and composition measurements),

This work use the Random Forest approach to fill the missing observations with other reported supersaturation ratios (0.2-0.6%). I am wondering whether other work had ever done using this method before. Are there any other different better ways to achieve data filling? Is there any connection between the reduction performance of RFRM-ShortVars and filling of the [CCN0.4] measurement gaps?

In Section 2.4.1(RFRM: training, testing, and optimising), why did the author just ignore the ARM SGP site? Is it possible to choose more sites among the 47 sites in later section?

Overall, it is suggested to consider publish this paper after some minor revisions, and some specific comments are listed as follows:

1. Figure 1 in page 7: It is recommended to describe the figure in more detail, E.g. the description of the Meteorology rectangle is not very clear. 2. Page 8 in Line 188: I guess rate of increase is about 57%. 3. Figure 3 in page 10: Should the line color in the Figure be changed to dark purple to avoid misunderstanding?

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