Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2020-509-AC1, 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

Arshad Arjunan Nair and Fangqun Yu

aanair@albany.edu

Received and published: 23 August 2020

Response to Anonymous Referee #1

We are grateful to Anonymous Referee #1 for constructive comments and suggestions, which have helped us revise and improve the clarity of this manuscript. Our replies addressing the referee comments follow; corresponding revisions (boxed) are in the revised manuscript and marked in a separate tracked-changes version.

General Comments

Referee Point P 1.1 — 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 microphysics 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.

Reply: We are grateful to Anonymous Referee #1 for reviewing this manuscript and affirmation of the value of this work and robustness of the method. Their following questions and suggestions were helpful in the revision of this manuscript.

Specific Comments

Referee Point P 1.2 — 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).

Reply: True. We have now made it clear in the manuscript as follows:

> **Lines 200–201**
>
> For observation–model simultaneity, all data (atmospheric state, composition, and [CCN0.4]) are integrated to the hourly resolution with their geometric mean.

Referee Point P 1.3 — 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 RFRMShortVars and filling of the [CCN0.4] measurement gaps?

Reply: We have not found such application in the atmospheric sciences; this is indeed a useful machine learning approach for experimentalists dealing with missing data and also in data quality checks. Breiman (2003) made the first suggestion of Random Forest for dealing with missing values and there have been subsequent applications in other fields (Stekhoven et al., 2011; Shah et al., 2014; Tang et al., 2017; Kokla et al., 2019). The additional advantage here over these ‘traditional’ RF imputation methods, is that the ‘predictors’ are the same physical parameter ([CCN]) except at different supersaturations, which compensates for the smaller training data size, and results in the exceedingly good agreements seen in Table 1. There are a wide variety of statistical methods of imputation from simple averaging to stochastic imputation, as well as their ensemble approaches; there could be better ways to achieve data filling, but it is dependent on the nature of the data.

We could find only minimal contribution of filling the measurement gaps to the observed decrease in RFRM performance (from RF-AllVars → RF-ShortVars). Fig. 1 (Page C9) shows the MFB distribution à la Figure 13(a) in the manuscript. Omitting the filled-in [CCN0.4] for RF-ShortVars’ [CCN0.4] derivation, Kendall’s τ correlation increased from $0.363 \rightarrow 0.415$, percentage in the good-agreement range ($|\mathrm{MFB}| < 0.6$) from $67.02\% \rightarrow 69.34\%$, and sample size n from $39,811 \rightarrow 29,047$. We have clarified this in the revised manuscript:

> **Lines 334–339**
>
> We note that filling the measurement gaps (per Section 2.4.2) could contribute to this observed decrease in RFRM performance (from RF-AllVars → RF-ShortVars). However, this contribution is minimal: when comparing the RF-ShortVars-derived [CCN0.4] with measurements excluding the filled-in [CCN0.4], Kendall’s τ correlation increased from $0.36 \rightarrow 0.42$, and percentage within the good-agreement range from $67.02\% \rightarrow 69.34\%$, with the sample size n reducing from $39,811 \rightarrow 29,047$. The deteriorated performance is mainly due to the reduction of necessary predictors to the available ones; the uncertainties associated with the measurements themselves may compound this.

Referee Point P 1.4 — 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?

Reply: Lines 208–209 were not very clear and have been modified as below. These 47 sites were chosen based on the availability of aerosol measurements. However, over the US, the ARM SGP site was the only one with ‘good-enough’ (long-term, with less gaps, and available co-located predictor measurements) publicly available data for application and evaluation of the RFRM.

> **Lines 207–208**
>
> The RFRM is trained on a subset of this data. First, the ARM SGP site is ignored; this is to establish a completely independent analysis with available observational data in Section 3.2.2.

Referee Point P 1.5 — Overall, it is suggested to consider publish this paper after some minor revisions, and some specific comments are listed as follows:

Technical Corrections

Referee Point P 1.6 — 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.

Reply: Fixed as follows:

> **Figure 1 caption, Page 2**
>
> **Figure 1.** The ARM SGP site in Lamont, Oklahoma, U.S.A with marked locations of the instruments. *Legend*— Meteorology: temperature and relative humidity measurements from the ARM Surface Meteorology Systems (MET) (Holdridge and Kyrouac, 1993; Chen and Xie, 1994). ACSM: Aerodyne Aerosol Chemical Speciation Monitor (Ng et al., 2011). [SO$_2$]: concentrations of SO2 measured by the ARM Aerosol Observing System (AOS; Hageman et al., 1996). CCNc: Cloud Condensation Nuclei Particle Counter (CCNc) (Shi and Flynn, 2007; Smith et al., 2011a, b; Hageman et al., 2017). This image is adapted from satellite imagery © 2020 Maxar Technologies, USDA Farm Service Agency obtained through the Google Maps Static API.

Referee Point P 1.7 — Page 8 in Line 188: I guess rate of increase is about 57%.

Reply: The values $(42.4923\%, 65.5809\%, \text{and } 54.336\%)$ are rounded in the paper.

> **Lines 185–186**
>
> Using this approach, we fill in the missing observations and improve data completeness from $\approx 42\% \rightarrow \approx 66\%$, an increase of $\approx 54\%$, for [CCN0.4] during this period (see Fig. 2).

Referee Point P 1.8 — Figure 3 in page 10: Should the line color in the Figure be changed to dark purple to avoid misunderstanding?

Reply: Figure 3 justifies the random sampling of a subset of GEOS-Chem-APM output data for training the RFRM. The overlap cannot be resolved unless zoomed in, which is good for the point we are making here. However, we have made the change as suggested in Fig. 2 (Page C10; Figure 3 in the revised manuscript).

References

Breiman, L.: Manual for Setting Up, Using, and Understanding Random Forest V4.0, Available online at https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf, 2003.

Chen, X. and Xie, S.: ARM: ARMBE: Atmospheric measurements, https://doi.org/10.5439/1095313, 1994.

Hageman, D., Behrens, B., Smith, S., Uin, J., Salwen, C., Koontz, A., Jefferson, A., Watson, T., Sedlacek, A., Kuang, C., Dubey, M., Springston, S., and Senum, G.: ARM: Aerosol Observing System (AOS): aerosol data, 1-min, mentor-QC applied, https://doi.org/10.5439/1025259, 1996.

Hageman, D., Behrens, B., Smith, S., Uin, J., Salwen, C., Koontz, A., Jefferson, A., Watson, T., Sedlacek, A., Kuang, C., Dubey, M., Springston, S., and Senum, G.: ARM: Aerosol Observing System (AOS): cloud condensation nuclei data, https://doi.org/10.5439/1150249, 2017.

Holdridge, D. and Kyrouac, J.: ARM: ARM-standard Meteorological Instrumentation at Surface, https://doi.org/10.5439/1025220, 1993.

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., and Hanhineva, K.: Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study, BMC Bioinformatics, 20, https://doi.org/10.1186/s12859-019-3110-0, 2019.

Ng, N. L., Herndon, S. C., Trimborn, A., Canagaratna, M. R., Croteau, P. L., Onasch, T. B., Sueper, D., Worsnop, D. R., Zhang, Q., Sun, Y. L., and Jayne, J. T.: An Aerosol Chemical

Speciation Monitor (ACSM) for Routine Monitoring of the Composition and Mass Concentrations of Ambient Aerosol, Aerosol Science and Technology, 45, 780–794, https://doi.org/10.1080/02786826.2011.560211, 2011.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H.: Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study, American Journal of Epidemiology, 179, 764–774, https://doi.org/10.1093/aje/kwt312, 2014.

Shi, Y. and Flynn, C.: ARM: Aerosol Observing System (AOS): cloud condensation nuclei data, averaged, https://doi.org/10.5439/1095312, 2007.

Smith, S., Salwen, C., Uin, J., Senum, G., Springston, S., and Jefferson, A.: ARM: AOS: Cloud Condensation Nuclei Counter, https://doi.org/10.5439/1256093, 2011a.

Smith, S., Salwen, C., Uin, J., Senum, G., Springston, S., and Jefferson, A.: ARM: AOS: Cloud Condensation Nuclei Counter (Single Column), averaged, https://doi.org/10.5439/1342133, 2011b.

Stekhoven, D. J. and Buhlmann, P.: MissForest–non-parametric missing value imputation for mixed-type data, Bioinformatics, 28, 112–118, https://doi.org/10.1093/bioinformatics/btr597, 2011.

Tang, F. and Ishwaran, H.: Random forest missing data algorithms, Statistical Analysis and Data Mining: The ASA Data Science Journal, 10, 363–377, https://doi.org/10.1002/sam.11348, 2017.

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

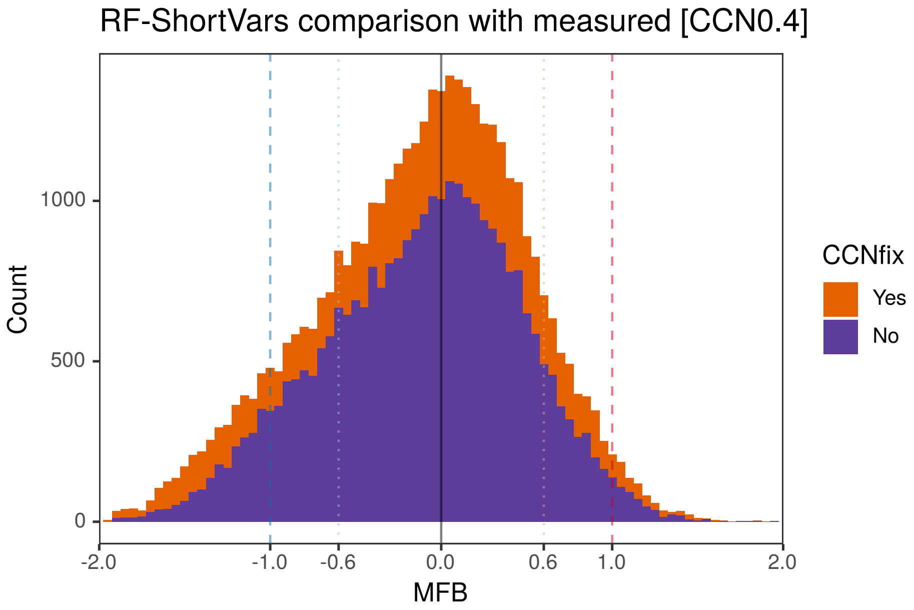

Fig. 1. Mean fractional bias (MFB) of the RF-derived compared to measured [CCN0.4]. Stacked histogram shows the pairwise counts by MFB, binwidth = 0.05.

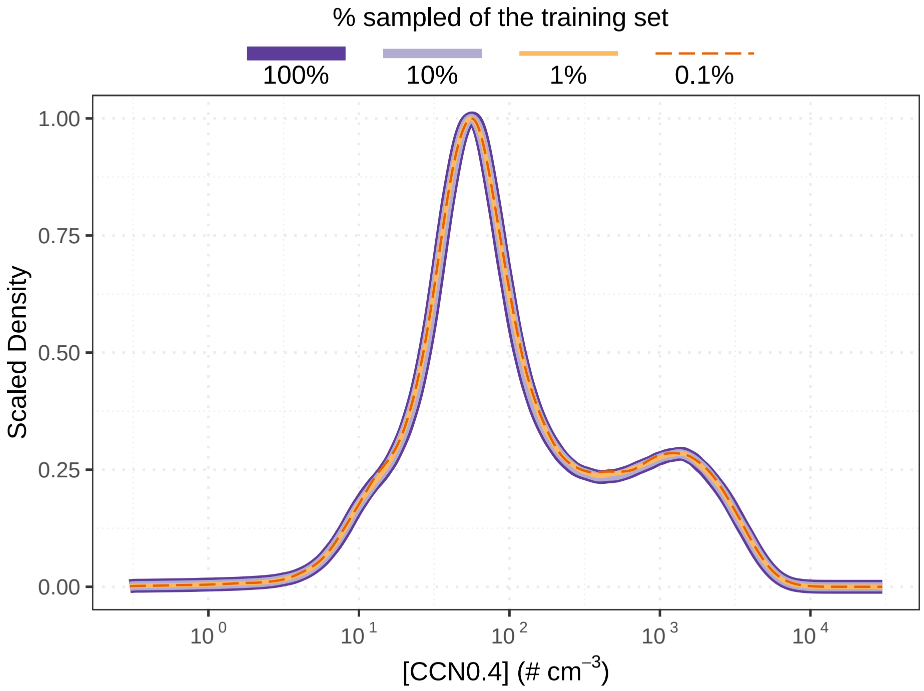

Fig. 2. Scaled Gaussian kernel density estimate for [CCN0.4] for the training set (dark purple) and each of its subsets (10%: light purple; 1%: light orange; 0.1%: dark orange).