Geophysical Research Letters

RESEARCH LETTER
10.1029/2021GL094133

Key Points:

- Machine learning (ML) derived cloud condensation nuclei numbers in strong agreement with comprehensive multi-campaign aircraft observations
- First demonstration that aerosol size information is contained in aerosol mass speciation, chemistry and meteorology, and is extractable by ML
- A physicochemically explainable (xAI) and robust ML avenue to mitigate aerosol-cloud interaction uncertainties in climate models is realized

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. A. Nair,
aanair@albany.edu

Citation:

Nair, A. A., Yu, F., Campuzano-Jost, P., DeMott, P. J., Levin, E. J. T., Jimenez, J. L., et al. (2021). Machine learning uncovers aerosol size information from chemistry and meteorology to quantify potential cloud-forming particles. *Geophysical Research Letters*, *48*, e2021GL094133. https://doi.org/10.1029/2021GL094133

Received 29 APR 2021
Accepted 8 OCT 2021

Machine Learning Uncovers Aerosol Size Information From Chemistry and Meteorology to Quantify Potential Cloud-Forming Particles

Arshad Arjunan Nair[1], Fangqun Yu[1], Pedro Campuzano-Jost[2,3], Paul J. DeMott[4], Ezra J. T. Levin[4,5], Jose L. Jimenez[2,3], Jeff Peischl[2,6], Ilana B. Pollack[4], Carley D. Fredrickson[7], Andreas J. Beyersdorf[8,9], Benjamin A. Nault[2,3,10], Minsu Park[11], Seong Soo Yum[11], Brett B. Palm[7], Lu Xu[12,13], Ilann Bourgeois[2,6], Bruce E. Anderson[8], Athanasios Nenes[14,15,16], Luke D. Ziemba[8], Richard H. Moore[8], Taehyoung Lee[17], Taehyun Park[17], Chelsea R. Thompson[2,6], Frank Flocke[18], Lewis Gregory Huey[19], Michelle J. Kim[12], and Qiaoyun Peng[7]

[1]Atmospheric Sciences Research Center, State University of New York, Albany, NY, USA, [2]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO, USA, [3]Department of Chemistry, University of Colorado Boulder, Boulder, CO, USA, [4]Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA, [5]Now at Handix Scientific, Boulder, CO, USA, [6]NOAA Chemical Science Laboratory, Boulder, CO, USA, [7]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA, [8]NASA Langley Research Center, Hampton, VA, USA, [9]California State University, San Bernardino, CA, USA, [10]Now at Center for Aerosols and Cloud Chemistry, Aerodyne Research, Inc., Billerica, MA, USA, [11]Department of Atmospheric Sciences, Yonsei University, Seoul, South Korea, [12]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA, [13]Now at NOAA Chemical Science Laboratory, Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO, USA, [14]Institute of Chemical Engineering Sciences, Foundation for Research & Technology-Hellas, Patras, Greece, [15]School of Architecture, Civil and Environmental Engineering, Swiss Federal Institute of Technology Lausanne, Lausanne, Switzerland, [16]School of Earth and Atmospheric Sciences and School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA, USA, [17]Department of Environmental Science, Hankuk University of Foreign Studies, Yongin, South Korea, [18]Atmospheric Chemistry Observations and Modeling, National Center for Atmospheric Research, Boulder, CO, USA, [19]School of Earth and Atmospheric Sciences, Georgia Tech, Atlanta, GA, USA

Abstract Cloud condensation nuclei (CCN) are mediators of aerosol-cloud interactions, which contribute to the largest uncertainty in climate change prediction. Here, we present a machine learning (ML)/artificial intelligence (AI) model that quantifies CCN from model-simulated aerosol composition, atmospheric trace gas, and meteorological variables. Comprehensive multi-campaign airborne measurements, covering varied physicochemical regimes in the troposphere, confirm the validity of and help probe the inner workings of this ML model: revealing for the first time that different ranges of atmospheric aerosol composition and mass correspond to distinct aerosol number size distributions. ML extracts this information, important for accurate quantification of CCN, additionally from both chemistry and meteorology. This can provide a physicochemically explainable, computationally efficient, robust ML pathway in global climate models that only resolve aerosol composition; potentially mitigating the uncertainty of effective radiative forcing due to aerosol-cloud interactions (ERF_{aci}) and improving confidence in assessment of anthropogenic contributions and climate change projections.

Plain Language Summary The largest uncertainties in climate change modeling are linked with cloud condensation nuclei (CCN). These tiny atmospheric particles modulate cloud formation and thus affect the Earth's energy budget. A machine learning/artificial intelligence model that accurately quantifies CCN can potentially reduce these uncertainties. Comprehensive multi-campaign aircraft measurements over varied atmospheric environments validate this model. Importantly, the inner workings of this model are teased out to reveal that its decisions are rooted in physical and chemical principles.

1. Introduction

Atmospheric aerosol effects, particularly on cloud radiative forcing, remain the largest source of uncertainty (or model diversity) in climate change prediction (IPCC, 2013). Those aerosols capable of condensing water droplets and forming clouds—cloud condensation nuclei (CCN)—contribute to this uncertainty. CCN interactions with water vapor thus impact cloud micro- and macrophysics, and consequently modulate cloud formation, its properties (size, number, and optical), and dynamics and precipitation (Ackerman et al., 2000; Albrecht, 1989; Ferek et al., 2000; Hansen et al., 1997; Liou & Ou, 1989; Pincus & Baker, 1994; Rosenfeld, 2000; Twomey, 1974, 1977; Twomey et al., 1984). These resultant effects consequently impact Earth's energy budget and influence climate and weather.

Obtaining agreement of CCN predictions with observations is crucial toward mitigating the uncertainty associated with aerosol-cloud interactions. Two factors play the largest role in determining CCN (at a given water supersaturation): aerosol particle number size distributions (PNSD) and aerosol chemical composition (speciation; Fitzgerald, 1973; Junge & McLaren, 1971). While the debate continues (Crosbie et al., 2015; Dusek et al., 2006; Hudson, 2007; Twohy & Anderson, 2008) as to which factor plays a larger role, the more predominant effect is arguably that of PNSD due to the third order dependence on size for the solute effect that permits water vapor condensation as well as the greater variability of PNSD than that of speciation, except in polluted regions. However, most global climate models (GCMs) use simplified prescriptions to estimate aerosol numbers or CCN from speciation while assuming a fixed PNSD (Boucher & Lohmann, 1995; Menon et al., 2002; Menon & Rotstayn, 2006). This is due to current computational constraints, which limit the incorporation into GCMs of size-resolved microphysics models with a detailed treatment of processes pertinent to a more accurate representation of PNSD and hence CCN number concentrations.

Machine learning (ML) is a subset of artificial intelligence (AI) where computers are trained on a large number of scenarios to acquire knowledge by statistical learning and without explicit instructions. While ML has been in use for the last several decades (Dramsch, 2020; Reichstein et al., 2019), in recent years, novel techniques and rapid advances in ML have led to its emergent applications in the atmospheric sciences (e.g., Grange et al., 2018; Jin et al., 2019; Nair & Yu, 2020; Su et al., 2020), especially in grappling with ordinal, nonlinear, complex, and massive amounts of data. It is key, however, that these increasingly black-box ML/AI techniques remain grounded in reality for trustworthiness and generalizability.

We, therefore, set out to probe the inner workings of our recently proposed ML model (Nair & Yu, 2020) trained on a chemical transport model with detailed size-resolved microphysics for deriving CCN number concentrations, that is, why CCN can be predicted from aerosol speciation (and other commonly available atmospheric variables) without size information. Comprehensive multi-campaign airborne measurements over varied physicochemical regimes across the tropospheric extent are used to explore the key parameters determining [CCN].

2. Methods

2.1. Machine Learning Model

Random forest (Breiman, 2001) is a ML technique that can be used for regression analysis and understanding the dependence of an outcome on other variables (its predictors). This is an ensemble (to reduce overfitting) of several decision trees (Breiman et al., 1984), each obtained on random subsets (Breiman, 1996) of the training data. For the generalizability of this ML model, it requires to be trained on a large number of scenarios, for which presently available measurements are scant (see Text S2). Here, the RFRM (Random Forest Regression Model) is trained on 30 yr simulations by GEOS-Chem-APM: a state-of-the-science chemical transport model with detailed size-resolved microphysics (Yu & Luo, 2009). The present study uses the RFRM-ShortVars configuration (Nair & Yu, 2020), a fast implementation (Wright & Ziegler, 2017) of random forest models (Breiman, 2003) in the statistical computing language R (R Core Team, 2020). RFRM-ShortVars, which was developed to use $PM_{2.5}$ (mass of Particulate Matter (PM) with particle diameter ≤ 2.5 μm) speciation as predictors for number concentrations of CCN at 0.4% supersaturation ([CCN0.4]) is retrained to use airborne measurements of PM_1 speciation (in lieu of $PM_{2.5}$ speciation measurements). Henceforth referred to as RFRM, this model derives [CCN0.4] from the following 9 commonly measured

variables of atmospheric state and composition as input predictors: (*Meteorology*) temperature (T) and relative humidity (RH), (*Gas-phase chemistry*) SO_2, NO_x, and O_3, and (*Aerosol composition and mass*) NH_4, SO_4, NO_3, and OA (organic aerosol). The present analysis focuses on [CCN0.4] for the purpose of demonstration and in future work will be extensible for the full CCN spectrum.

2.2. Multi-Campaign Airborne Measurements

Comprehensive (global scope, tropospheric vertical extent, varied seasons, and high temporal resolution) airborne measurements of atmospheric state and composition variables provide an unparalleled opportunity to probe the inner workings of the ML derivation of [CCN0.4] in varied atmospheric environments. Seven airborne campaigns were identified (Table S1) with simultaneous measurements of the 9 predictors as well as [CCN0.4] and with their spatial domain shown in Figure S1, instrumentation details in Table S2, and further details of [CCN0.4] measurements in Text S1. PNSD presented here are <1,000 nm, above which aerosol numbers sharply taper off and negligibly contribute to [CCN0.4]. For the ATom1–4 campaign, PNSD is measured using the aerosol microphysical properties (AMP) package (Brock, Williamson, et al., 2019) and for the other campaigns using a scanning mobility particle sizer (SMPS; and nano-SMPS for WE-CAN) and either an Ultra-High Sensitivity Aerosol Spectrometer (UHSAS: ARCTAS, DISCOVER-AQ^{TX}, and WE-CAN) or a Laser Aerosol Spectrometer (LAS: DC3, KORUS-AQ, and SEAC^{4}RS). To increase data coverage, if a measurement was missing and if there were measurements one second prior and/or after, it was imputed with their mean value. For DC3 [SO_2] (0.1 Hz) and WE-CAN HR-ToF-AMS (0.2 Hz), measurements were assumed constant for 10 and 5 s, respectively.

2.3. Statistical Estimators to Quantify RFRM Performance

In the present study, we use the following statistical estimators for model-observation comparison: Kendall rank correlation coefficient (R_K) to quantify correlation and **%-Good** to quantify agreement. The rationale and advantages of using these statistical metrics to evaluate model-observation comparisons are described in detail elsewhere (Nair et al., 2019). These estimators are defined as follows:

$$R_K = \frac{\sum_{i=2}^{n}(\text{sign}(C_i^m - C_{i-1}^m))(\text{sign}(C_i^o - C_{i-1}^o))}{\sqrt{\binom{n}{2} - \frac{1}{2}\sum_{i=1}^{n} t_i^m(t_i^m - 1)}\sqrt{\binom{n}{2} - \frac{1}{2}\sum_{i=1}^{n} t_i^o(t_i^o - 1)}} \tag{1}$$

$$\mathbf{FB} = \frac{(C_i^m - C_i^o)}{\left(\frac{C_i^m + C_i^o}{2}\right)}; \quad \mathbf{MFB} = \frac{1}{n}\sum_{i=1}^{n}\frac{(C_i^m - C_i^o)}{\left(\frac{C_i^m + C_i^o}{2}\right)} \tag{2}$$

where n is the sample size, C is the value, t is the number of tied ranks in the i^{th} group of tied ranks, and superscripts ^{o} and ^{m} denote observed and modeled values, respectively:

$$\%\textbf{-Good} = 100 \times \frac{1}{n}\sum_{i=1}^{n}((|\mathbf{FB}(i)| \le 0.6) \rightarrow 1) \tag{3}$$

%-Good is defined on the basis of the Fractional Bias (**FB**). It is the percentage of RFRM-derived [CCN0.4] with FB in the range [−0.6, +0.6] with respect to measured [CCN0.4]. Correspondingly, **%-Over**: FB > +0.6 and **%-Under**: FB < −0.6.

3. Results

3.1. Machine Learning Successfully Derives CCN Number Concentrations

We compare three approaches: (a) LinReg: linear regression on the airborne measurements of aerosol speciation for [CCN0.4] as an effective representation for current aerosol mass to number prescriptions in GCMs, (b) RFRM-PM: a Random Forest Regression Model, trained on a global model of atmospheric chemical composition with size-resolved microphysics (GEOS-Chem-APM), for [CCN0.4] on aerosol speciation (PM_1 NH_4, SO_4, NO_3, and OA (organic aerosol)) as a possible improvement on LinReg, and (c) RFRM: a specific

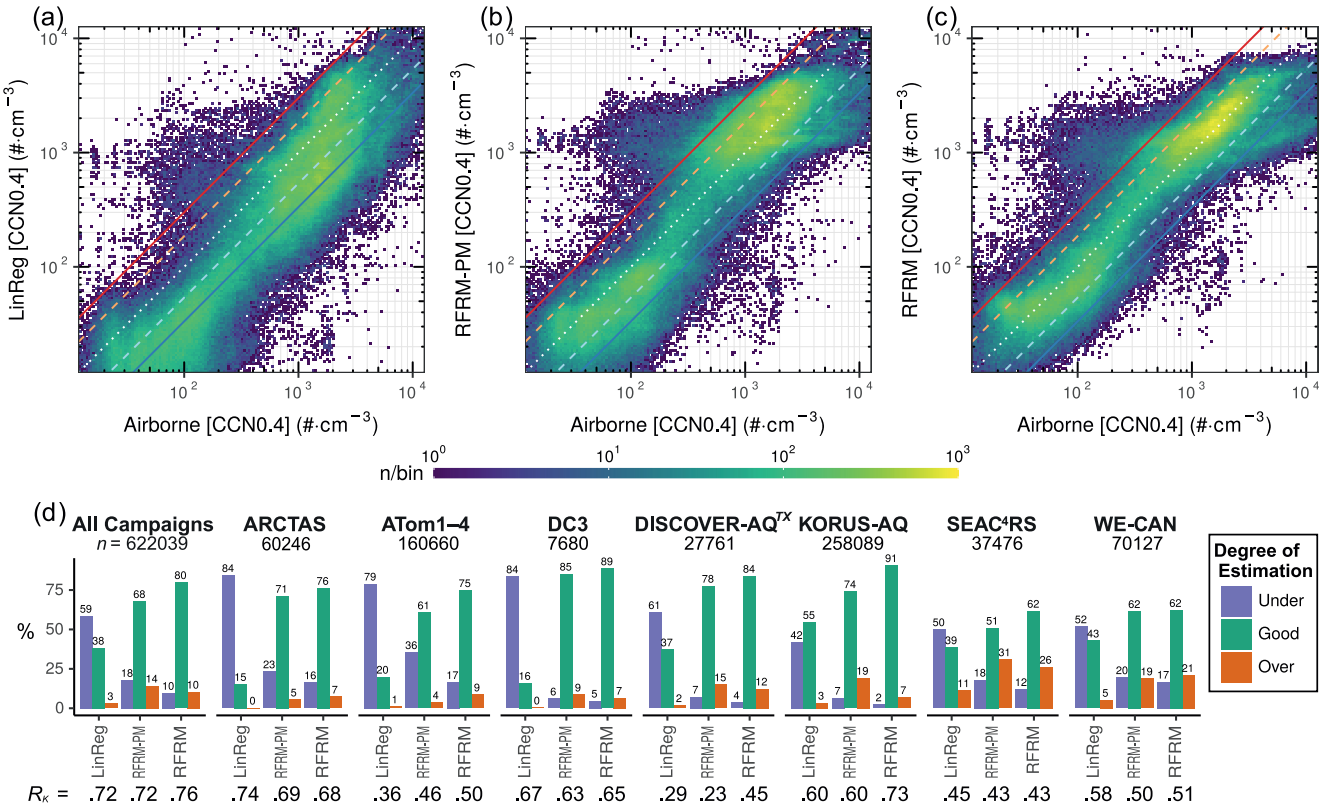

Figure 1. Comparison of machine learning derived versus airborne measurements of [CCN0.4]. Binned scatter plot for data at 1 Hz resolution from all campaigns. For (a) Linear Regression (LinReg), (b) RFRM-PM, and (c) RFRM. Central 99% range of the airborne-measured [CCN0.4] shown for a zoomed-in view. The lines, in the order of decreasing *y*-intercept, indicate fractional bias (FB) of (solid red) + 1, (dashed light red) + 0.6, (dotted white) 0 or 1 : 1 agreement, (dashed light blue) − 0.6, and (solid blue) − 1, respectively. Logscale colorbar shows the count per bin. Bin-width is 0.02 (arbitrary; corresponding to ± 2.3%) on the logscale. (d) Summary statistics for the degree of model-observation agreement and correlation, as defined in the Methods, for each aircraft campaign.

Random Forest Regression Model, trained on GEOS-Chem-APM, for [CCN0.4] on aerosol speciation and additional variables of (*Gas-phase chemistry*) [SO_2], [NO_x], and [O_3], and (*Meteorology*) temperature (*T*) and relative humidity (RH).

Illustrated in Figure 1 is the comparison for each of these methods with aggregated airborne campaign measurements (for individual campaign comparisons see Figures S5–S7) demonstrating improved ML skill from LinReg (Figure 1a) → RFRM-PM (Figure 1b) → RFRM (Figure 1c). Figure 1d provides the summary statistics quantifying model-observation degree of agreement and correlation. In comparison with airborne measurements of [CCN0.4], RFRM-derived values show strong agreement (%-Good, defined in Methods, of ≈ 80%) and high correlation ($R_K \approx 0.76$). The highest density is on or around the dotted white line indicating 1:1 model–observation agreement and the majority (green bars in Figure 1d) of the derived values are within the corridor of good-agreement between the dashed light red and dashed light blue lines. While the RFRM is overall robust, we examine the cases where it deviates from airborne measurements. When these model-observation disagreements (absolute FB (|FB|) > 1) do occur, they are rare (5.9%) and in a regime where their effect on cloud properties will be smallest (Martin et al., 1994; Ramanathan, 2001), that is, the sensitivity of cloud droplet numbers to changes in aerosol numbers is reduced at their high concentrations. For high (> 3×10^3 cm^{-3}) measured [CCN0.4] RFRM low bias (FB < −1) is largely associated with the wildfire plume measurements during the ARCTAS and WE-CAN campaigns. It must be noted here that the low likelihood of the RFRM being exposed to these scenarios of high [CCN0.4] and predictor values in its training (on the GEOS-Chem-APM global simulations) may contribute to this observed low bias. Ultimately, however, this scenario is infrequent: ARCTAS (8.7% of its measurements), WE-CAN (8.3%), SEAC^{4}RS (2.7%), and other campaigns (≪ 0.5%). The high bias (FB > +1) of RFRM-derived [CCN0.4] occurs mainly during SEAC^{4}RS (14%) and WE-CAN (7.1%). While the reason for this remains to be determined, there may

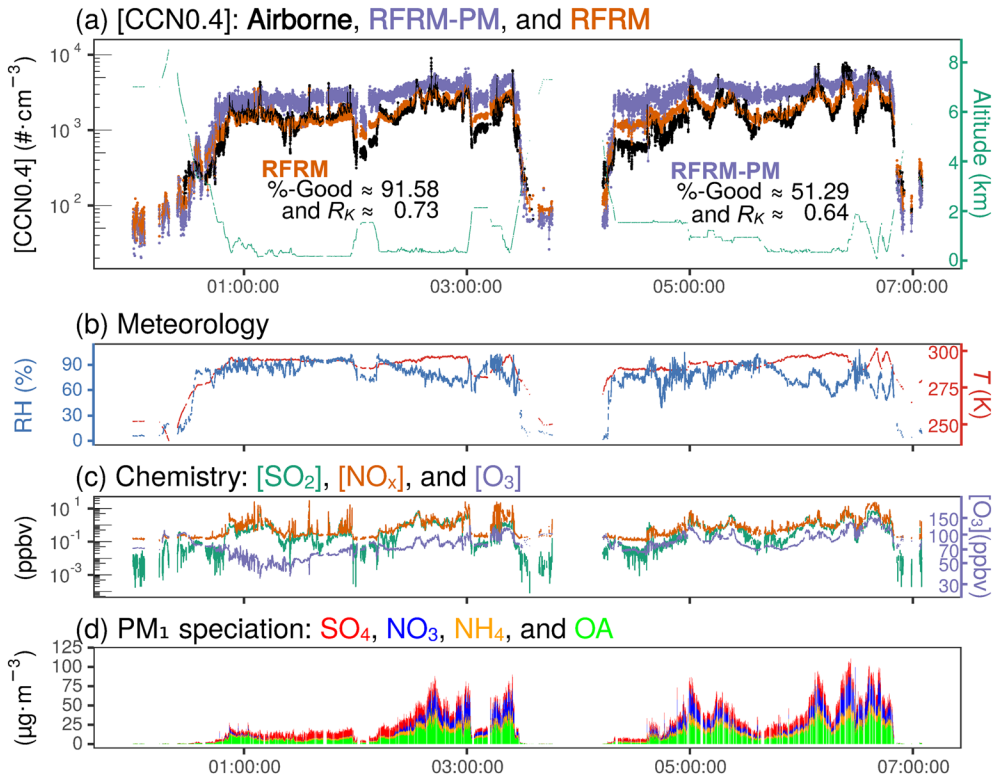

Figure 2. Time series of [CCN0.4] and variables of atmospheric state and composition shown for a selected campaign day (KORUS-AQ: June 10, 2016). (a) [CCN0.4]: (black) Airborne-measurement, (purple) RFRM-PM-derived, and (orange) RFRM-derived; and (green) altitude. (b) Meteorology: (red) temperature (T) and (blue) relative humidity (RH). (c) Chemistry: (green) [SO_2], (orange) [NO_x], and (blue) [O_3]. (d) PM_1 speciated masses of (red) SO_4, (blue) NO_3, (orange) NH_4, and (green) OA. Data is shown at 1 Hz resolution. Solid lines associated with [CCN0.4] are 5 s rolling means.

be measurement uncertainties; for instance, in Figure S4a, [CCN0.4] measured directly and inferred separately are in large disagreement for SEAC^{4}RS during these instances of apparent RFRM-high-bias.

While the Random Forest Regression Models demonstrate a high degree of predictive performance overall, we examine their performance in higher detail, leveraging the high temporal resolution of airborne measurements, in Figure 2. For illustration, we select a day (June 10, 2016 from the KORUS-AQ campaign) with large variability in altitude (surface–8.5 km) as well as the 9 predictors. Shown is the time series of the measurements of these variables during this day: measured [CCN0.4] in black in Figure 2a and the 9 simultaneously measured predictors (Figures 2b–2d) used as input predictors for the RFRMs to derive [CCN0.4]. RFRM-PM-derived (purple) and RFRM-derived (orange) [CCN0.4] are shown in Figure 2a. Even down to 1 Hz of resolution, RFRM is able to capture [CCN0.4] variations with high skill (%-Good≈ 92% and $R_K \approx 0.73$). During periods (1st, 3rd, and 6th hours) of aircraft ascent and descent and the corresponding large change in magnitude of [CCN0.4], the RFRM demonstrates its robustness in varying physicochemical environments. The consistency of the RFRM performance across the vertical extent of the troposphere is illustrated for each campaign in Figures S9 and S10. For WE-CAN (4–6 km) and ARCTAS (1–3 km), the earlier noted tendency of the RFRM to underpredict [CCN0.4] is seen in the splitting and skewing left of the violin distribution (Figures S9 and S10). Examining this in further detail, for observations with $PM_1OA > 40\ \mu g \cdot m^{-3}$, mean fractional bias (MFB) for ARCTAS(WE-CAN) is − 1.3(−0.6) as compared to − 0.03(+0.2) when otherwise ($PM_1OA \leq 40\ \mu g \cdot m^{-3}$). This suggests that the RFRM-underestimation is due mostly to the high organic mass (likely in biomass burning plumes) not experienced by the RFRM during its training or the underestimation of the potential contribution of organic aerosol to CCN numbers in current models or a combination of these factors.

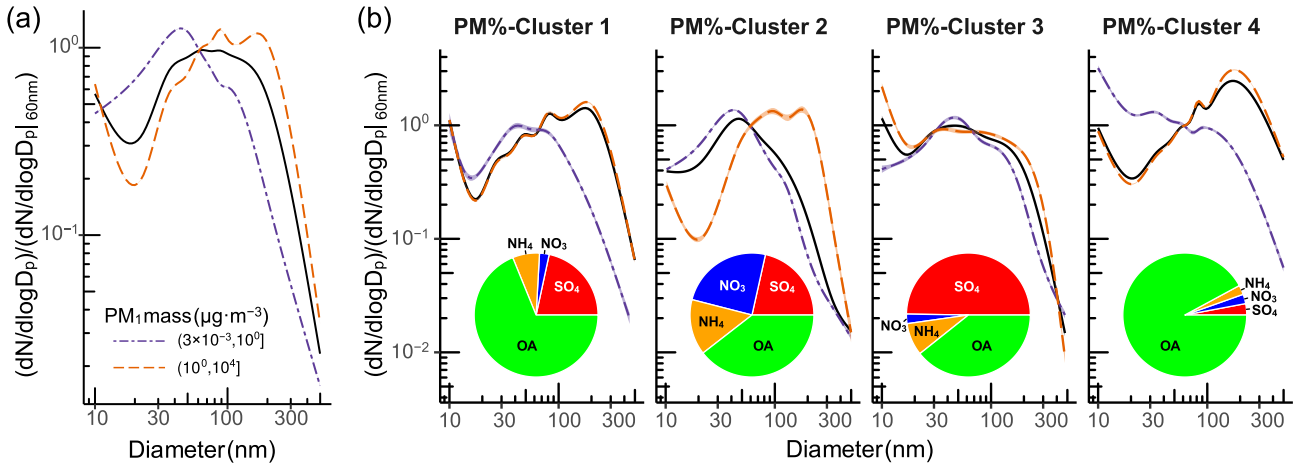

Figure 3. Aerosol mass and composition carry its number size distribution information. Average (generalized additive model) airborne measured aerosol number size distributions (PNSD) normalized to ≈60 nm. For (purple, dotted-dashed) $PM_{tot} \leq 1\ \mu g \cdot m^{-3}$, and (orange, dashed) $PM_{tot} > 1\ \mu g \cdot m^{-3}$. Solid black curve in (a) is for all data. (b) For each cluster: Cluster 1 (SO_4: 19%–24%, OA: 66%–71%, NO_3: 0%–5%, and NH_4: 4.5%–9.5%), Cluster 2 (SO_4: 19%–24%, OA: 37%–42%, NO_3: 22%–27%, and NH_4: 12%–17%), Cluster 3 (SO_4: 47.5%–52.5%, OA: 37%–42%, NO_3: 0%–5%, and NH_4: 6%–11%), Cluster 4 (SO_4: 0.5%–5.5%, OA: 91%–96%, NO_3: 0%–5%, and NH_4: 0%–5%), and (black) respective cluster-wise average. Typical aerosol composition for each cluster is illustrated by the inset pie charts.

3.2. Aerosol Mass Speciation Contains Size Distribution Information as Revealed by Machine Learning

In GCMs that do not resolve particle size distributions, proxies for aerosol numbers or cloud droplet numbers are obtained from aerosol mass speciation alone, assuming a fixed aerosol number size distribution. In this study, LinReg is an effective representation of the aerosol mass-to-number prescription in GCMs. This is due to linearly regressing for measured [CCN0.4] on all the measured aerosol speciation variables. Therefore, by virtue of overfitting, there can be no better aerosol mass-to-number prescription for the airborne measurements used in this study. Despite this, LinReg is demonstrated to be inadequate (Figure 1a). A potential improvement—RFRM-PM—employs one of the most accurate ML approaches for regression and appreciably (%-Good: 38 → 68%) improves the degree of agreement with CCN measurements. The importance of considering 19 predictor variables of atmospheric state and composition (not limited to aerosol mass speciation) for accurate RFRM-derivation of [CCN0.4] has been demonstrated (Nair & Yu, 2020). Considering observational limitations, reduction to nine important predictors including T, RH, $[SO_2]$, $[NO_x]$, and $[O_3]$ is possible without significant deterioration of model performance. RFRM, which considers these variables in addition to only aerosol speciated mass, is in agreement with measured [CCN0.4] to a much greater degree (%-Good: 38 → 68 → 80%; Figures 1 and 2, and Figures S9 and S10). With the significant amount of measurement data that these airborne campaigns provide, we search for the reasons for why consideration of predictors beyond PM_1 speciation helps improve the machine-learning model derivation of [CCN0.4].

The RFRM-PM performs better than LinReg for deriving [CCN0.4] when only the PM_1 speciated masses are used as input (Figure 1). To examine the reason for this, Figure 3 shows how the PM_1 mass contains information about the aerosol number size distribution (PNSD; P: particle/aerosol) that the random forest approach can leverage. The average normalized (to ≈60 nm (Williamson et al., 2019): the rough cut-off size for CCN0.4) airborne measured PNSD is shown in Figure 3. Figure 3a shows that for two different total PM_1 mass ranges the PNSD profile varies. While the linear regression implicitly assumes a fixed average PNSD (black curve), the RFRM derives [CCN0.4] using decisions in the subspace corresponding to the PM_1 total mass, which defines more representative variations of PNSD. In addition, Figure 3b demonstrates that the aerosol composition (speciated mass fractions of aerosol mass) also carries PNSD information. The four panels correspond to distinct clusters of aerosol composition, and each cluster with speciated composition of the total PM_1 mass within a range of ± 2.5% to ensure in-cluster homogeneity as well as each cluster spanning the entire range of PM_1 total mass. The clusters are determined with the aid of an unsupervised ML technique (k-means clustering), described in the Text S2 and illustrated in Figures S12 and S13. Thus

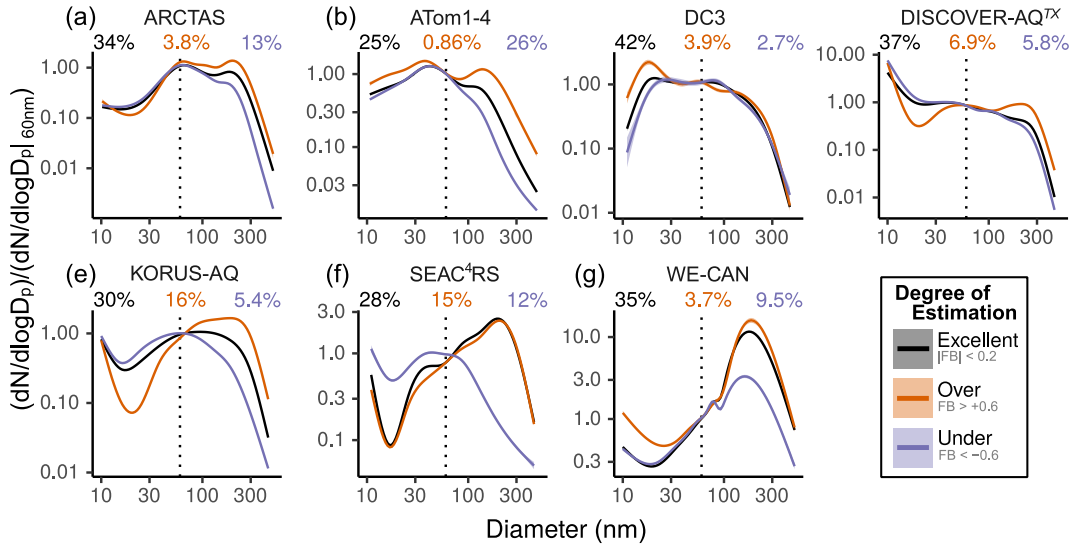

Figure 4. Machine learning can extract aerosol number size information from chemistry and meteorology. Average (generalized additive model) aerosol number size distributions (PNSD) normalized to ≈60 nm for each campaign: (a) ARCTAS, (b) ATom1–4, (c) DC3, (d) DISCOVER-AQ^{TX}, (e) KORUS-AQ, (f) SEAC^{4}RS, and (g) WE-CAN. Data shown for the subset of RFRM in good-agreement and where RFRM-PM (orange) overestimates, (purple) underestimates, or is in (black) excellent agreement with airborne measurements of [CCN0.4]. Percentage of the number of observations in each class of degree of agreement shown with respectively colored text in panel sub-headings.

aerosol mass and composition confer to the RFRM-PM the ability to implicitly consider the PNSDs pertinent to PM_1 mass and speciation in its derivation of [CCN0.4] and enhance its skill compared to linear regression with an assumed mean PNSD.

3.3. Further Size Information Can Be Machine-Learned From Additional Chemistry and Meteorology

To examine why RFRM is more robust than RFRM-PM in its derivation of [CCN0.4], we consider the subset of the data where RFRM-derived [CCN0.4] is in good-agreement with airborne measurements. Counterintuitively, RFRM-PM overestimates (FB > 0.6) mostly (83.6%) when higher [CCN0.4] is measured and underestimates (FB < −0.6) mostly (82.4%) when lower [CCN0.4] is measured. This is indicative that rather than a general bias in the RFRM, it is the non-consideration of the predictors other than PM speciation contributing to the RFRM-PM bias. In Figure 4, RFRM-PM-derived [CCN0.4] is classified into excellent-agreement (|FB| < 0.2; roughly 22% deviation from airborne measurement of [CCN0.4]; black), overestimation (orange), and underestimation (purple). The percentages corresponding to these classes are noted in each campaign's panel. Illustrated are the typical PNSD normalized to the ∼60 nm diameter, corresponding roughly to the cut-off size of CCN0.4. Across all campaigns, differences in these size distributions with respect to the degree of estimation remain consistent. More detailed differences in PNSD across the vertical extent of the troposphere are also illustrated in Figure S11. In the scenario of a more typical PNSD, with high Aitken and low accumulation mode, both RFRM and RFRM-PM are in agreement with measurements. When the accumulation mode is much higher and Aitken mode is much lower than average, RFRM is in agreement but RFRM-PM overestimates. This is because the aerosol mass distribution toward the larger diameters results in less numerous particles than a mean size distribution would suggest. When the Aitken mode is much higher and the accumulation mode much lower than average, the corollary follows. The additional consideration of chemical species of SO_2, NO_x, and O_3 and meteorology (T and RH), which are important for chemistry and gas-to-particle conversion (including new particle formation and growth) and hence PNSD, enables RFRM to contain more discerning subspaces for its decision making than RFRM-PM. With regards to the PNSD, these additional predictors carry rich information about the air mass history, sources of primary aerosols, and occurrence of atmospheric new particle formation and growth and photochemical processing toward the secondary aerosol formation. Future investigations will focus

on comprehensive assessment of individual contributions of each predictor variable, consideration of all variables in the full-RFRM pertinent toward the improved reflection of the ambient PNSD, and delineation of the physicochemical processes that determine CCN (spectrum) number concentrations.

4. Conclusions

This work demonstrates, using comprehensive airborne multi-campaign measurements encompassing the varied physicochemical conditions across the troposphere, the overall success of ML in deriving CCN number concentrations. Importantly, ML can extract aerosol size information from aerosol composition and additionally from atmospheric chemical and meteorological variables; this demonstrates that the statistical learning of ML/AI algorithms is emergent from the underlying physical (and chemical) laws. This physicochemically explainable and robust ML model can provide a computationally efficient pathway for a more accurate representation of CCN in GCMs. This may potentially reduce the uncertainties associated with aerosol-cloud interactions in the assessment of anthropogenic forcing and climate change projection.

Data Availability Statement

Data from the following aircraft campaigns were used in this study—ARCTAS (Jacob et al., 2010): ARCTAS Team (2020), ATom1–4 (Brock, Williamson, et al., 2019): Allen et al. (2019), Brock, Kupc, et al. (2019), Jimenez et al. (2019), and Ryerson et al. (2019), DC3 (Barth et al., 2015): DC3 Team (2013), DISCOVER-AQ^{TX}: DISCOVER-AQ Team (2014), KORUS-AQ (Jordan et al., 2020): KORUS-AQ Team (2018), SEAC^{4}RS (Toon et al., 2016): SEAC4RS Team (2014), WE-CAN: WE-CAN Team (2019). Additional dual column CCNc measurement (Uin et al., 2017a, 2017b) data were obtained from the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science User Facility managed by the Biological and Environmental Research program. All data sets used in this study are publicly available and individually detailed as follows:

- [CCN0.18–0.86], PNSD, PM_1 composition and mass, $[SO_2]$, $[NO_x]$, $[O_3]$, T, and RH measured during the ARCTAS (Jacob et al., 2010) campaign: ARCTAS Team (2020, https://www-air.larc.nasa.gov/cgi-bin/ArcView/arctas).
- PNSD, PM_1 composition and mass, $[SO_2]$, $[NO_x]$, $[O_3]$, T, and RH measured during the ATom1–4 (Brock, Williamson, et al., 2019) campaigns: Allen, Crounse, Kim, Teng, and Wennberg (2019); Ryerson, Thompson, Peischl, and Bourgeois (2019); Jimenez et al. (2019); Brock, Kupc, et al. (2019, https://espo.nasa.gov/atom/archive/browse/atom/DC8).
- [CCN0.13–0.68], PNSD, PM_1 composition and mass, $[SO_2]$, $[NO_x]$, $[O_3]$, T, and RH measured during the DC3 (Barth et al., 2015) campaign: DC3 Team (2013, https://www-air.larc.nasa.gov/missions/dc3-seac4rs/).
- [CCN0.14–0.60], PNSD, PM composition and mass, $[SO_2]$, $[NO_x]$, $[O_3]$, T, and RH measured during the DISCOVER-AQ^{TX} campaign: DISCOVER-AQ Team (2014, https://www-air.larc.nasa.gov/cgi-bin/ArcView/discover-aq.tx-2013).
- [CCN0.6], PNSD, PM_1 composition and mass, $[SO_2]$, $[NO_x]$, $[O_3]$, T, and RH measured during the KORUS-AQ (Jordan et al., 2020) campaign: KORUS-AQ Team (2018, https://www-air.larc.nasa.gov/missions/korus-aq/).
- [CCN0.09–0.56], PNSD, PM_1 composition and mass, $[SO_2]$, $[NO_x]$, $[O_3]$, T, and RH measured during the SEAC^{4}RS (Toon et al., 2016) campaign: SEAC4RS Team (2014, https://www-air.larc.nasa.gov/cgi-bin/ArcView/seac4rs).
- [CCN0.079–0.73], PNSD, PM_1 composition and mass, $[SO_2]$, $[NO_x]$, $[O_3]$, T, and RH measured during the WE-CAN campaign: WE-CAN Team (2019, https://www-air.larc.nasa.gov/cgi-bin/ArcView/firexaq).
- Dual column CCNc measurement (Uin et al., 2017a, 2017b) data were obtained from the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science User Facility managed by the Biological and Environmental Research program, which is publicly available at the ARM Discovery Data Portal (https://www.archive.arm.gov/discovery/).

Acknowledgments
This research has been supported by the National Aeronautics and Space Administration (NASA grant no. NNX17AG35G) and the National Science Foundation (NSF grant no. AGS-1550816). BAN, PCJ & JLJ were supported by NASA (grant nos. NNX15AJ23G, NNX15AH33A, 80NSSC19K0124, and 80NSSC18K0630). PJD & EJTL acknowledge support from the NSF (grant no. AGS-1650786). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. MP & SSY acknowledge the support from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT; grant no. NRF-2018R1A2B2006965). MJK was supported by an NSF Atmospheric and Geospace Sciences Postdoctoral Research Fellowship (AGS-PRF; grant no. 1524860). CDF, BBP & QP acknowledge support from the NSF (grant no. AGS-1652688) and the National Oceanic and Atmospheric Administration (NOAA grant no. NA17OAR4310012).
The authors are grateful to Chuck Brock (NOAA) for the in situ measurements of aerosol microphysical properties during the ATom1–4 campaigns, Andrew J. Weinheimer, Denise D. Montzka & David J. Knapp (ARCTAS, DISCOVER-AQ^{TX}, KORUS-AQ, and WE-CAN) and Thomas B. Ryerson (ATom1–4, DC3, and SEAC^{4}RS) for (NO_x) and (O_3) measurements, Paul O. Wennberg, John D. Crounse & Hannah M. Allen for (SO_2) measurements during the ARCTAS and ATom1–4 campaigns, Kevin R. Barry (funded by NSF grant no. AGS-1660486; WE-CAN: CCNc and SMPS), Sonia M. Kreidenweis (WE-CAN: CCNc and HR-ToF-AMS), Kathryn A. Moore (funded by the NSF Graduate Research Fellowship grant no. 006784; WE-CAN: CCNc), Darin W. Toohey & Michael Reeves (WE-CAN: UHSAS), Lauren A. Garofalo & Delphine K. Farmer (funded by the NOAA grant no. NA17OAR4310010; WE-CAN: HR-ToF-AMS), and Joel A. Thornton (WE-CAN: CIMS (SO_2) measurements). The authors are thankful to Michael Shook & Gao Chen at the NASA Langley Research Center Airborne Science Data for Atmospheric Composition (https://www-air.larc.nasa.gov/index.html) for data curation. The authors also thank the DOE ARM SGP Research Facility teams for the operation and maintenance of instruments, quality checks, and making their measurement data publicly available.

References

Ackerman, A. S., Toon, O. B., Stevens, D. E., Heymsfield, A. J., Ramanathan, V., & Welton, E. J. (2000). Reduction of tropical cloudiness by soot. *Science*, *288*(5468), 1042–1047. https://doi.org/10.1126/science.288.5468.1042

Albrecht, B. A. (1989). Aerosols, cloud microphysics, and fractional cloudiness. *Science*, *245*(4923), 1227–1230. https://doi.org/10.1126/science.245.4923.1227

Allen, H. M., Crounse, J. D., Kim, M. J., Teng, A. P., & Wennberg, P. O. (2019). *ATom: L2 In Situ Data from Caltech Chemical Ionization Mass Spectrometer (CIT-CIMS)*. ORNL Distributed Active Archive Center. Retrieved from https://espo.nasa.gov/atom/archive/browse/atom/DC8/CIT-SO2

ARCTAS Team. (2020). *ARCTAS Field Campaign Data*. NASA Langley Atmospheric Science Data Center DAAC. https://doi.org/10.5067/SUBORBITAL/ARCTAS2008/DATA001

Barth, M. C., Cantrell, C. A., Brune, W. H., Rutledge, S. A., Crawford, J. H., Huntrieser, H., et al. (2015). The Deep Convective Clouds and Chemistry (DC3) field campaign. *Bulletin of the American Meteorological Society*, *96*(8), 1281–1309. https://doi.org/10.1175/bams-d-13-00290.1

Boucher, O., & Lohmann, U. (1995). The sulfate-CCN-cloud albedo effect. *Tellus B: Chemical and Physical Meteorology*, *47*(3), 281–300. https://doi.org/10.3402/tellusb.v47i3.16048

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2003). Manual for Setting Up, Using, and Understanding Random Forest V4.0 [Computer software manual]. Retrieved from https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J.897 (1984). *Classification and regression trees*. Routledge. https://doi.org/10.1201/9781315139470

Brock, C. A., Kupc, A., Williamson, C. J., Froyd, K., Erdesz, F., Murphy, D. M., & Wilson, J. C. (2019). *ATom: L2 In Situ Measurements of Aerosol Microphysical Properties (AMP)*. ORNL Distributed Active Archive Center. Retrieved from https://espo.nasa.gov/atom/archive/browse/atom/DC8/SDAerosol https://doi.org/10.3334/ORNLDAAC/1671

Brock, C. A., Williamson, C., Kupc, A., Froyd, K. D., Erdesz, F., Wagner, N., et al. (2019). Aerosol size distributions during the Atmospheric Tomography Mission (ATom): Methods, uncertainties, and data products. *Atmospheric Measurement Techniques*, *12*(6), 3081–3099. https://doi.org/10.5194/amt-12-3081-2019

Crosbie, E., Youn, J.-S., Balch, B., Wonaschütz, A., Shingler, T., Wang, Z., et al. (2015). On the competition among aerosol number, size and composition in predicting CCN variability: A multi-annual field study in an urbanized desert. *Atmospheric Chemistry and Physics*, *15*(12), 6943–6958. https://doi.org/10.5194/acp-15-6943-2015

DC3 Team. (2013). *DC3 Field Campaign Data*. NASA Langley Atmospheric Science Data Center DAAC. https://doi.org/10.5067/AIRCRAFT/DC3/DC8/AEROSOL-TRACEGAS

DISCOVER-AQ Team (2014). DISCOVER-AQ Field Campaign Data. NASA Langley Atmospheric Science Data Center DAAC. https://doi.org/10.5067/AIRCRAFT/DISCOVER-AQ/AEROSOL-TRACEGAS

Dramsch, J. S. (2020). 70 yr of machine learning in geoscience in review. In *Machine learning in geosciences* (pp. 1–55). Elsevier. https://doi.org/10.1016/bs.agph.2020.08.002

Dusek, U., Frank, G. P., Hildebrandt, L., Curtius, J., Schneider, J., Walter, S., et al. (2006). Size matters more than chemistry for cloud-nucleating ability of aerosol particles. *Science*, *312*(5778), 1375–1378. https://doi.org/10.1126/science.1125261

Ferek, R. J., Garrett, T., Hobbs, P. V., Strader, S., Johnson, D., Taylor, J. P., et al. (2000). Drizzle suppression in ship tracks. *Journal of the Atmospheric Sciences*, *57*(16), 2707–2728. https://doi.org/10.1175/1520-0469(2000)057<2707:dsist>2.0.co;2

Fitzgerald, J. W. (1973). Dependence of the supersaturation spectrum of CCN on aerosol size distribution and composition. *Journal of the Atmospheric Sciences*, *30*(4), 628–634. https://doi.org/10.1175/1520-0469(1973)030<0628:dotsso>2.0.co;2

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., & Hueglin, C. (2018). Random forest meteorological normalization models for Swiss PM_{10} trend analysis. *Atmospheric Chemistry and Physics*, *18*(9), 6223–6239. https://doi.org/10.5194/acp-18-6223-2018

Hansen, J., Sato, M., & Ruedy, R. (1997). Radiative forcing and climate response. *Journal of Geophysical Research: Atmospheres*, *102*(D6), 6831–6864. https://doi.org/10.1029/96jd03436

Hudson, J. G. (2007). Variability of the relationship between particle size and cloud-nucleating ability. *Geophysical Research Letters*, *34*(8). https://doi.org/10.1029/2006gl028850

IPCC. (2013). In T. F. Stocker et al. (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107415324

Jacob, D. J., Crawford, J. H., Maring, H., Clarke, A. D., Dibb, J. E., Emmons, L. K., et al. (2010). The Arctic Research of the Composition of the Troposphere from Aircraft and Satellites (ARCTAS) mission: Design, execution, and first results. *Atmospheric Chemistry and Physics*, *10*(11), 5191–5212. https://doi.org/10.5194/acp-10-5191-2010

Jimenez, J. L., Campuzano-Jost, P., Day, D. A., Nault, B. A., Price, D. J., & Schroder, J. C. (2019). *ATom: L2 Measurements from CU High-Resolution Aerosol Mass Spectrometer (HR-AMS)*. ORNL Distributed Active Archive Center. Retrieved from https://espo.nasa.gov/atom/archive/browse/atom/DC8/AMS https://doi.org/10.3334/ORNLDAAC/1716

Jin, J., Lin, H. X., Segers, A., Xie, Y., & Heemink, A. (2019). Machine learning for observation bias correction with application to dust storm data assimilation. *Atmospheric Chemistry and Physics*, *19*(15), 10009–10026. https://doi.org/10.5194/acp-19-10009-2019

Jordan, C. E., Crawford, J. H., Beyersdorf, A. J., Eck, T. F., Halliday, H. S., Nault, B. A., et al. (2020). Investigation of factors controlling $PM_{2.5}$ variability across the South Korean Peninsula during KORUS-AQ. *Elementa: Science of the Anthropocene*, *8*. https://doi.org/10.1525/elementa.424

Junge, C., & McLaren, E. (1971). Relationship of cloud nuclei spectra to aerosol size distribution and composition. *Journal of the Atmospheric Sciences*, *28*(3), 382–390. https://doi.org/10.1175/1520-0469(1971)028<0382:rocnst>2.0.co;2

KORUS-AQ Team. (2018). *KORUS-AQ Field Campaign Data*. NASA Langley Atmospheric Science Data Center DAAC. Retrieved from https://www-air.larc.nasa.gov/missions/korus-aq/

Liou, K.-N., & Ou, S.-C. (1989). The role of cloud microphysical processes in climate: An assessment from a one-dimensional perspective. *Journal of Geophysical Research*, *94*(D6), 8599–8607. https://doi.org/10.1029/jd094id06p08599

Martin, G. M., Johnson, D. W., & Spice, A. (1994). The measurement and parameterization of effective radius of droplets in warm stratocumulus clouds. *Journal of the Atmospheric Sciences*, *51*(13), 1823–1842. https://doi.org/10.1175/1520-0469(1994)051<1823:tmapoe>2.0.co;2

Menon, S., Genio, A. D. D., Koch, D., & Tselioudis, G. (2002). GCM simulations of the aerosol indirect effect: Sensitivity to cloud parameterization and aerosol burden. *Journal of the Atmospheric Sciences*, *59*(3), 692–713. https://doi.org/10.1175/1520-0469(2002)059<0692:gsotai>2.0.co;2

Menon, S., & Rotstayn, L. (2006). The radiative influence of aerosol effects on liquid-phase cumulus and stratiform clouds based on sensitivity studies with two climate models. *Climate Dynamics*, *27*(4), 345–356. https://doi.org/10.1007/s00382-006-0139-3

Nair, A. A., & Yu, F. (2020). Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements. *Atmospheric Chemistry and Physics*, *20*(21), 12853–12869. https://doi.org/10.5194/acp-20-12853-2020

Nair, A. A., Yu, F., & Luo, G. (2019). Spatioseasonal variations of atmospheric ammonia concentrations over the United States: Comprehensive model-observation comparison. *Journal of Geophysical Research: Atmospheres*, *124*(12), 6571–6582. https://doi.org/10.1029/2018JD030057

Pincus, R., & Baker, M. B. (1994). Effect of precipitation on the albedo susceptibility of clouds in the marine boundary layer. *Nature*, *372*(6503), 250–252. https://doi.org/10.1038/372250a0

Ramanathan, V., Crutzen, P. J., Kiehl, J. T., & Rosenfeld, D. (2001). Aerosols, climate, and the hydrological cycle. *Science*, *294*(5549), 2119–2124. https://doi.org/10.1126/science.1064034

R Core Team. (2020). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rosenfeld, D. (2000). Suppression of rain and snow by urban and industrial air pollution. *Science*, *287*(5459), 1793–1796. https://doi.org/10.1126/science.287.5459.1793

Ryerson, T. B., Thompson, C. R., Peischl, J., & Bourgeois, I. (2019). *ATom: L2 In Situ Measurements from NOAA Nitrogen Oxides and Ozone (NOyO3) Instrument*. ORNL Distributed Active Archive Center. https://doi.org/10.3334/ORNLDAAC/1734

SEAC4RS Team. (2014). *SEAC4RS Field Campaign Data*. NASA Langley Atmospheric Science Data Center DAAC. https://doi.org/10.5067/AIRCRAFT/SEAC4RS/AEROSOL-TRACEGAS-CLOUD

Su, H., Wu, L., Jiang, J. H., Pai, R., Liu, A., Zhai, A. J., et al. (2020). Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophysical Research Letters*, *47*(17). https://doi.org/10.1029/2020gl089102

Toon, O. B., Maring, H., Dibb, J., Ferrare, R., Jacob, D. J., Jensen, E. J., et al. (2016). Planning, implementation, and scientific goals of the Studies of Emissions and Atmospheric Composition, Clouds and Climate Coupling by Regional Surveys (SEAC4RS) field mission. *Journal of Geophysical Research: Atmospheres*, *121*(9), 4967–5009. https://doi.org/10.1002/2015JD024297

Twohy, C. H., & Anderson, J. R. (2008). Droplet nuclei in non-precipitating clouds: Composition and size matter. *Environmental Research Letters*, *3*(4), 045002. https://doi.org/10.1088/1748-9326/3/4/045002

Twomey, S. A. (1974). Pollution and the planetary albedo. *Atmospheric Environment*, *8*(12), 1251–1256. https://doi.org/10.1016/j.atmosenv.2007.10.062

Twomey, S. A. (1977). The influence of pollution on the shortwave albedo of Clouds. *Journal of the Atmospheric Sciences*, *34*(7), 1149–1152. https://doi.org/10.1175/1520-0469(1977)034<1149:tiopot>2.0.co;2

Twomey, S. A., Piepgrass, M., & Wolfe, T. L. (1984). An assessment of the impact of pollution on global cloud albedo. *Tellus B: Chemical and Physical Meteorology*, *36B*(5), 356–366. https://doi.org/10.1111/j.1600-0889.1984.tb00254.x

Uin, J., Salwen, C. & Senum, G. (2017a). Cloud Condensation Nuclei Particle Counter (AOSCCN2COLA). 2017-04-12 to 2020-08-11, Southern Great Plains (SGP) Lamont, OK (Extended and Co-located with C1) (E13). Atmospheric Radiation Measurement (ARM) Archive, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (US). Retrieved from https://adc.arm.gov/discovery/#/results/datastream::sgpaosccn2colaE13.b1

Uin, J., Salwen, C., & Senum, G. (2017b). *Cloud Condensation Nuclei Particle Counter (AOSCCN2COLB). 2017-04-12 to 2020-08-11, Southern Great Plains (SGP) Lamont, OK (Extended and Co-located with C1) (E13).* Atmospheric Radiation Measurement (ARM) Archive, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (US). Retrieved from https://adc.arm.gov/discovery/#/results/datastream::sgpaosccn2colbE13.b1

WE-CAN Team. (2019). *WE-CAN Field Campaign Data*. NASA Langley Atmospheric Science Data Center DAAC. Retrieved from https://www-air.larc.nasa.gov/cgi-bin/ArcView/firexaq?MERGE=1

Williamson, C. J., Kupc, A., Axisa, D., Bilsback, K. R., Bui, T., Campuzano-Jost, P., et al. (2019). A large source of cloud condensation nuclei from new particle formation in the tropics. *Nature*, *574*(7778), 399–403. https://doi.org/10.1038/s41586-019-1638-9

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

Yu, F., & Luo, G. (2009). Simulation of particle size distribution with a global aerosol model: Contribution of nucleation to aerosol and CCN number concentrations. *Atmospheric Chemistry and Physics*, *9*(20), 7691–7710. https://doi.org/10.5194/acp-9-7691-2009

References From the Supporting Information

Bahreini, R., Ervens, B., Middlebrook, A. M., Warneke, C., de Gouw, J. A., DeCarlo, P. F., et al. (2009). Organic aerosol formation in urban and industrial plumes near Houston and Dallas, Texas. *Journal of Geophysical Research*, *114*. https://doi.org/10.1029/2008jd011493

Bates, D., & Eddelbuettel, D. (2013). Fast and elegant numerical linear algebra using the RcppEigenPackage. *Journal of Statistical Software*, *52*(5). https://doi.org/10.18637/jss.v052.i05

DeCarlo, P. F., Kimmel, J. R., Trimborn, A., Northway, M. J., Jayne, J. T., Aiken, A. C., et al. (2006). Field-deployable, high-resolution, time-of-flight aerosol mass spectrometer. *Analytical Chemistry*, *78*(24), 8281–8289. https://doi.org/10.1021/ac061249n

Guo, H., Campuzano-Jost, P., Nault, B. A., Day, D. A., Schroder, J. C., Kim, D., et al. (2021). The importance of size ranges in aerosol instrument intercomparisons: A case study for the atmospheric tomography mission. *Atmospheric Measurement Techniques*, *14*(5), 3631–3655. https://doi.org/10.5194/amt-14-3631-2021

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Hu, W., Campuzano-Jost, P., Day, D. A., Croteau, P., Canagaratna, M. R., Jayne, J. T., et al. (2017). Evaluation of the new capture vaporizer for aerosol mass spectrometers (AMS) through laboratory studies of inorganic species. *Atmospheric Measurement Techniques*, *10*(8), 2897–2921. https://doi.org/10.5194/amt-10-2897-2017-supplement

Kaufman, L., & Rousseeuw, P. J. (Eds.). (1990). *Finding groups in data*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316801

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/tit.1982.1056489

Moore, R. H., & Nenes, A. (2009). Scanning flow CCN analysis—A method for fast measurements of CCN spectra. *Aerosol Science and Technology*, *43*(12), 1192–1207. https://doi.org/10.1080/02786820903289780

Nault, B. A., Campuzano-Jost, P., Day, D. A., Schroder, J. C., Anderson, B., Beyersdorf, A. J., et al. (2018). Secondary organic aerosol production from local emissions dominates the organic aerosol budget over Seoul, South Korea, during KORUS-AQ. *Atmospheric Chemistry and Physics*, *18*(24), 17769–17800. https://doi.org/10.5194/acp-18-17769-2018-supplement

Roberts, G. C., & Nenes, A. (2005). A continuous-flow streamwise thermal-gradient CCN chamber for atmospheric measurements. *Aerosol Science and Technology*, *39*(3), 206–221. https://doi.org/10.1080/027868290913988

Twomey, S. A. (1959). The nuclei of natural cloud formation part II: The supersaturation in natural clouds and the variation of cloud droplet concentration. *Geofisica Pura e Applicata*, *43*(1), 243–249. https://doi.org/10.1007/bf01993560

Uin, J. (2016). *Cloud condensation nuclei particle counter (CCN) instrument handbook (Tech. Rep.)*. DOE Office of Science Atmospheric Radiation Measurement (ARM) Program. https://doi.org/10.2172/1251411

Uin, J., & Smith, S. (2021). *Southern Great Plains (SGP) Aerosol Observing System (AOS) instrument handbook*. Office of Scientific and Technical Information (OSTI). https://doi.org/10.2172/1756406