Machine learning uncovers aerosol size information from chemistry and meteorology to quantify potential cloud-forming particles

Arshad Nair (aanair@albany.edu)
State University of New York at Albany https://orcid.org/0000-0003-2530-7757
Fangqun Yu
University at Albany, State University of New York https://orcid.org/0000-0001-8862-4835
Pedro Campuzano Jost
University of Colorado, Boulder https://orcid.org/0000-0003-3930-010X
Paul DeMott
Colorado State University https://orcid.org/0000-0002-3719-1889
Ezra Levin
Handix Scientific
Jose Jimenez
University of Colorado-Boulder https://orcid.org/0000-0001-6203-1847
Jeff Peischl
NOAA https://orcid.org/0000-0002-9320-7101
Ilana Pollack
Colorado State University
Carley Fredrickson
University of Washington
Andreas Beyersdorf
California State University
Benjamin Nault
Aerodyne Researcy Inc https://orcid.org/0000-0001-9464-4787
Minsu Park
Yonsei University
Seong Soo Yum
Yonsei University
Brett Palm
University of Colorado Boulder
Lu Xu
California Institute of Technology
Ilann Bourgeois

University of Colorado-Boulder
Bruce Anderson
NASA Langley Research Center
Athanasios Nenes
EPFL
Luke Ziemba
NASA Langley Research Center
Richard H Moore
NASA Langley Research Center https://orcid.org/0000-0003-2911-4469
Taehyoung Lee
Hankuk University of Foreign Studies
Taehyun Park
Hankuk University of Foreign Studies
Chelsea Thompson
University of Colorado Boulder
Frank Flocke
National Center for Atmospheric Research
Lewis Huey
Georgia Institute of Technology
Michelle Kim
Everactive, Inc.
Qiaoyun Peng
University of Washington

Article

Keywords: cloud condensation nuclei, climate change prediction, aerosol composition, machine learning/artificial intelligence

Posted Date: February 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-244416/v1

Version of Record: A version of this preprint was published at Geophysical Research Letters on November 9th, 2021. See the published version at https://doi.org/10.1029/2021GL094133.

Machine learning uncovers aerosol size information from chemistry and meteorology to quantify potential cloud-forming particles

Arshad Arjunan Nair*, Fangqun Yu*, Pedro Campuzano-Jost, Paul J. DeMott, Ezra J. T. Levin, Jose L. Jimenez, Jeff Peischl, Ilana B. Pollack, Carley D. Fredrickson, Andreas J. Beyersdorf, Benjamin A. Nault, Minsu Park, Seong Soo Yum, Brett B. Palm, Lu Xu, Ilann Bourgeois, Bruce E. Anderson, Athanasios Nenes, Luke D. Ziemba, Richard H. Moore, Taehyoung Lee, Taehyun Park, Chelsea R. Thompson, Frank Flocke, Lewis Gregory Huey, Michelle J. Kim & Qiaoyun Peng

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/artificial intelligence model that quantifies CCN from variables of aerosol composition, atmospheric trace gases, and meteorology. Comprehensive multi-campaign airborne measurements, covering varied physicochemical regimes in the troposphere, confirm the validity of and help probe the inner workings of this machine learning model: revealing for the first time that different ranges of atmospheric aerosol composition and mass correspond to distinct aerosol number size distributions. Machine learning extracts this information, important for accurate quantification of CCN, additionally from both chemistry and meteorology. This can provide a physicochemically explainable, computationally efficient, robust machine learning 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.

Introduction

Atmospheric aerosol effects, particularly on cloud radiative forcing, remain the largest source of uncertainty (or model diversity) in climate change prediction[1]. 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 that of precipitation)[2–11]. These resultant effects consequently impact Earth's energy budget and influence climate and weather.

Obtaining agreement of CCN predictions with observations is crucial towards 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)[12,13]. While the debate continues[14–17] 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[18–20]. This is due to current computational constraints, which limit the incorporation into GCMs of size-resolved microphysics models with 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 (the humble and ubiquitous linear regression model) for the last several decades[21,22], in recent years, novel techniques and rapid advances in this field have led to its emergent applications in the atmospheric sciences[23–33], especially in grappling with ordinal, non-linear, 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[33] for deriving CCN number concentrations.

Here, we present an ML model to derive CCN number concentrations ([CCN]) without explicit PNSD information. Comprehensive multi-campaign airborne measurements over varied physicochemical regimes across the tropospheric extent demonstrate its validity. We show that for accurately quantifying [CCN], both size and chemistry of the aerosol matter. More importantly, we demonstrate for the first time that aerosol speciation (and other commonly available atmospheric variables) contain PNSD information. This information, important for

accurately deriving [CCN], is successfully extracted by the ML model. This study provides a robust and explainable AI pathway, without compromising on computational efficiency, for GCMs to incorporate more realistic PNSD information in their simulations of cloud-forming aerosols; this can have potential implications towards reducing uncertainties in the effective radiative forcing of aerosol–cloud interactions.

Results and discussion

Comprehensive (global scope, tropospheric vertical extent, varied seasons, and high temporal resolution) airborne measurements (detailed in the Methods and in Supplementary Fig. 1 and Supplementary Tables 1 and 2) of atmospheric state and composition variables provide an unparalleled opportunity to evaluate the machine learning derivation of number concentrations of CCN at 0.4% supersaturation ([CCN0.4]). We present three approaches: (1) 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, (2) 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 (3) RFRM: a 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). These models are detailed in the Methods.

Machine learning successfully derives CCN number concentrations. Illustrated in Fig. 1 is the comparison for each of these methods with aggregated airborne campaign measurements (for individual campaign comparisons see Supplementary Figs. 5–7) demonstrating improved machine learning skill from LinReg (Fig. 1a) → RFRM-PM (Fig. 1b) → RFRM (Fig. 1c). Fig. 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 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 Fractional Bias ($|FB|$) > 1) do occur, they are rare (5.9%) and in a regime where their effect on cloud

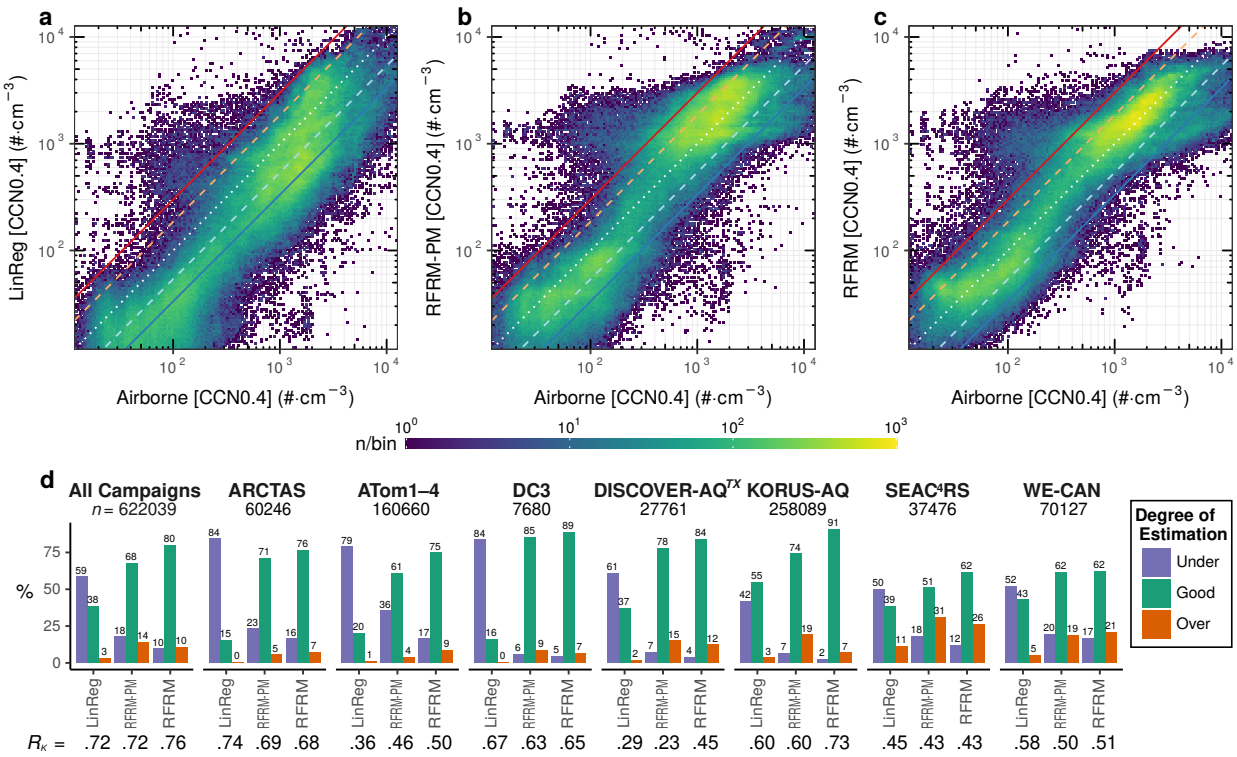

Figure 1: **Comparison of machine learning derived versus airborne measurements of [CCN0.4].** Binned scatter plot for data at the 1Hz resolution from all campaigns. For **a** Linear Regression (LinReg), **b** RFRM-PM, and **c** RFRM. Central 99% range of the airborne-measured [CCN0.4] shown for a zoomed-in 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) on the log-scale. **d** Summary statistics for the degree of model–observation agreement and correlation, as defined in the Methods, for each aircraft campaign.

properties will be smallest[34,35]. For high ($> 3 \times 10^3\,\mathrm{cm}^{-3}$) 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 ($\ll 0.5\%$). 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 be measurement uncertainties; for instance, in Supplementary Fig. 4a, [CCN0.4] measured directly and inferred separately are in large disagreement for SEAC^{4}RS during these instances of apparent RFRM-high-bias.

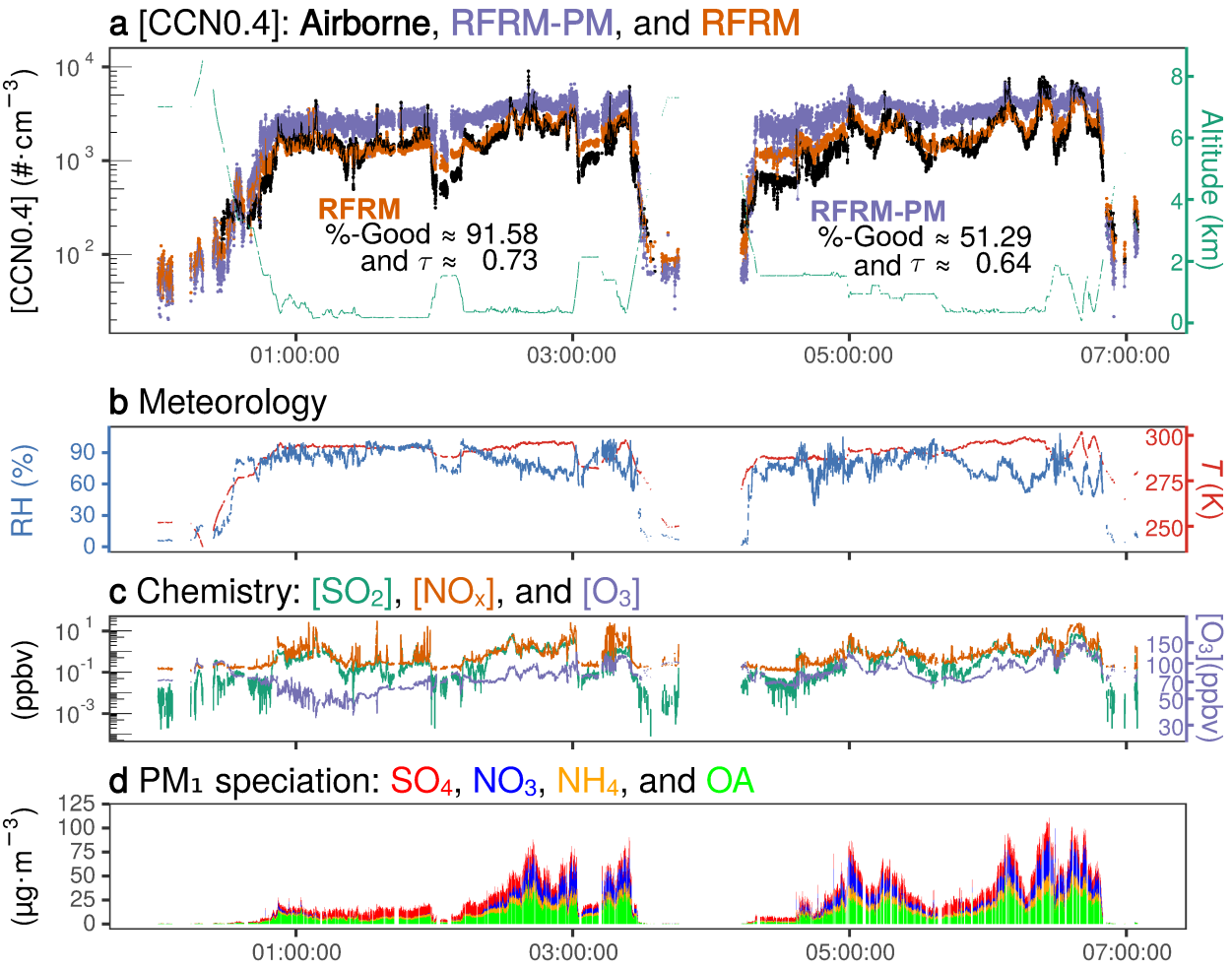

Figure 2: **Time series of [CCN0.4] and variables of atmospheric state and composition shown for a selected campaign day (KORUS-AQ: 10 June 2016). a** [CCN0.4]: (black) Airborne-measurement, (purple) RFRM-PM-derived, and (orange) RFRM-derived; and (green) altitude. **b** Meteorology: (red) temperature (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 the 1 Hz resolution. Solid lines associated with [CCN0.4] are 5 s rolling means.

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 Fig. 2. For illustration, we select a day (10 June 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. Shown in Fig. 2a is the measured [CCN0.4], which varies over a range of roughly 3 orders of magnitude from 10^1–10^4 cm^{-3}. The 9 simultaneously measured predictors (Fig. 2b–d) are used as input predictors for the RFRMs to derive [CCN0.4].

RFRM-PM-derived (purple) and RFRM-derived (orange) [CCN0.4] are shown in Fig. 2a. Even down to the 1 Hz 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 Supplementary Figs. 9&10. 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 (Supplementary Figs. 9&10). 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.

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. We have demonstrated that LinReg, directly obtained from the comprehensive airborne measurements and by virtue of this overfitting is an effective representation of current aerosol mass to number prescriptions in GCMs, can be inadequate. A potential improvement, RFRM-PM, which employs one of the most accurate machine learning approaches for regression appreciably (%-Good: $38 \rightarrow 68\%$) 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[33]. Considering observational limitations, a maximum reduction to 9 most important predictors including T, RH, [SO_2], [NOx], 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: $38 \rightarrow 68 \rightarrow 80\%$; Figs. 1, 2, and Supplementary Figs. 9&10). With the significant amount of measurement data that these airborne campaigns provide, we probe into the reasons for why consideration of predictors beyond PM_1 speciation helps improve the machine-learning model derivation of [CCN0.4].

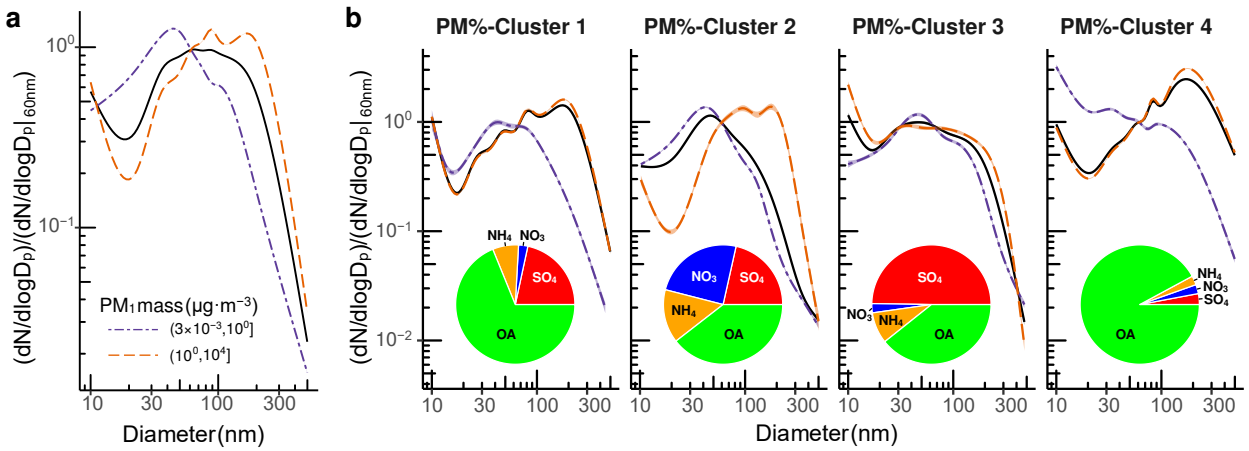

Figure 3: **Aerosol mass and composition carry its number size distribution information.** Average (generalized additive model) airborne measured aerosol number size distributions (PNSD) normalized to ≈60 nm. For (purple, dot-dashed) $PM_{tot} \leq 1$ μg cm^{-3}, and (orange, dashed) $PM_{tot} > 1$ μg cm^{-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.

The RFRM-PM performs better than LinReg for deriving [CCN0.4] when only the PM_1 speciated masses are used as input (Fig. 1). To examine the reason for this, Fig. 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[36]: the rough cut-off size for CCN0.4) airborne measured PNSD is shown in Fig. 3. Fig. 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, Fig. 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 machine learning technique (k-means clustering), described in the Methods and illustrated in Supplementary Figs. 12&13. Thus 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.

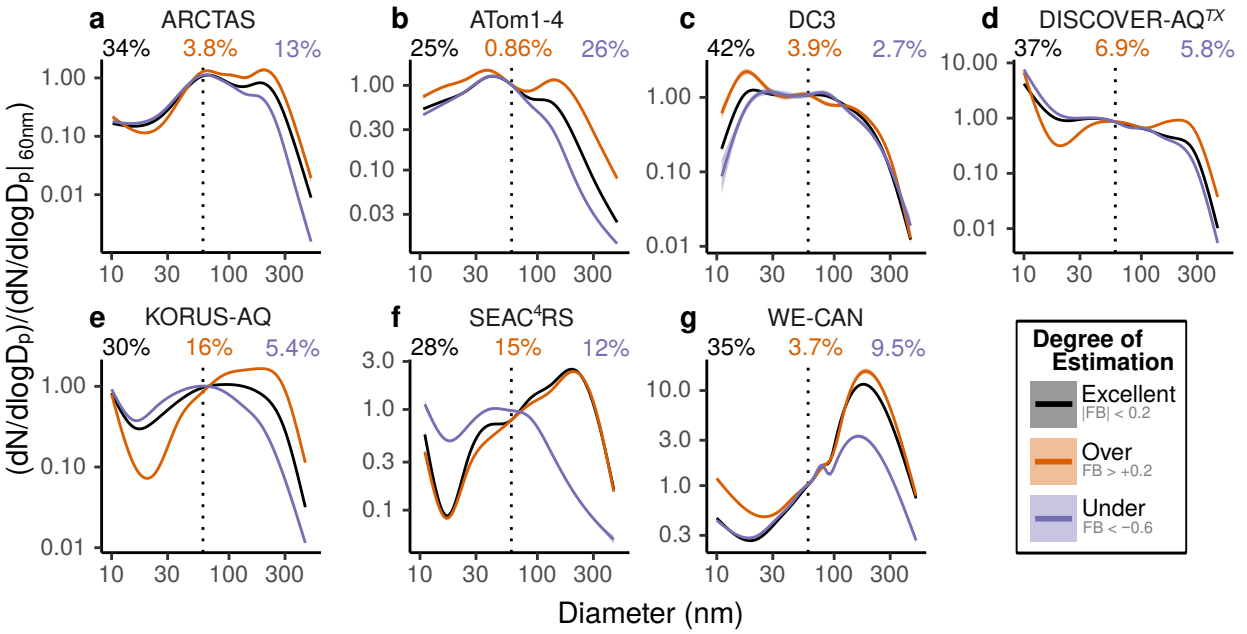

Figure 4: **Machine learning can extract aerosol number size information from chemistry and meteorology.** Average (generalized additive model) aerosol number size distributions (PNSD) normalized to ≈60 nm for each campaign: **a** ARCTAS, **b** ATom1–4, **c** DC3, **d** DISCOVER-AQ^{TX}, **e** KORUS-AQ, **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.

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 Fig. 4, RFRM-PM-derived [CCN0.4] is classified into excellent-agreement ($|\text{FB}| < 0.2$; roughly 22% deviation from airborne measurement of [CCN0.4]; black), overestimation (orange), and underestimation (purple). The percentage 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 Supplementary Fig. 11. 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 towards 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 towards 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 towards improved reflection of the ambient PNSD, and delineation of the physicochemical processes that determine CCN (spectrum) number concentrations.

This work demonstrates, using comprehensive airborne multi-campaign measurements encompassing the varied physicochemical conditions across the troposphere, the overall success of machine learning in deriving CCN number concentrations. Importantly, machine learning 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 machine learning model can provide a computationally efficient pathway for a more accurate representation of CCN in global climate models and potentially reduce the uncertainties associated with aerosol–cloud interactions in assessing anthropogenic forcing and climate change projection.

Methods

Multi-campaign airborne measurements. Data were collected from seven airborne campaigns with measurements of the 9 predictors as well as [CCN0.4] identified in Supplementary Table 1 and with their spatial domain shown in Supplementary Fig. 1. The 9 predictor variables required by the RFRM for its derivation of [CCN0.4] are measured in these campaigns using instrumentation detailed in Supplementary Table 2. PNSD, here, are limited to ∼1000 nm, by the size of 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[37] 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).

[CCN0.4] when $ss \neq 0.4\%$. [CCN] measurements during the ARCTAS, DC3, DISCOVER-AQ^{TX}, SEAC^{4}RS, and WE-CAN campaigns were typically made over a range of supersaturations, either by changing the temperature gradient or the flow rate in the growth chamber[38], and not necessarily for constant (0.4%) supersaturation. For these campaigns, the DMT CCN-100 instrument data is reported for a range of supersaturations from 0.08–0.86%, with $>$ 99% within the range 0.1–0.7%. For the KORUS-AQ campaign, the DMT CCN-100 instrument data is reported for 0.6% supersaturation. For a 1:1 comparison of RFRM-derived [CCN0.4] to measurements, we convert the [CCN] measured at supersaturations other than 0.4% using an empirical fitting function. This is to increase the data coverage during the campaign periods. This function is determined from [CCN] measurements at the US Department of Energy's (DOE) Atmospheric Radiation Measurement (ARM) Southern Great Plains (SGP) site located in Lamont, Oklahoma (36°36′18″ N, 9°29′6″ W, 318 m). ARM SGP [CCN] measurements made using a DMT CCN-200[39–41], which is the two-column version of the instrument deployed during the airborne campaigns, were used. The measurement data[42,43] is made publicly available by ARM (https://adc.arm.gov/discovery). The two columns provide the opportunity to compare coterminous [CCN0.4] and [CCNss], where ss is some supersaturation other than 0.4%, and obtain an empirical fit for [CCN0.4] as a function of [CCNss] and ss. Supersaturations are rounded to the nearest 0.05 and the ratios of [CCNss]:[CCN0.4] (where $ss \in \{n \times 0.05\% : n \in \mathbb{Z}^+$ where $1 \leq n \leq 27\}$) are plotted against ss in Supplementary Fig. 2. A polynomial fit is obtained on the median values corresponding to the supersaturation ratios. Within the range of supersaturations of the measurements here (0.08–0.86%), $\mathcal{O}3$ and higher terms have a negligible contribu-

tion. The function is equivalent to the Taylor expansion to $\mathcal{O}2$ of the functional relationship CCN(ss), with no assumptions of its form (such as Twomey's approach[44]). This empirical fit function is used to approximate [CCN0.4] from airborne measurements of [CCNss] where $ss \neq 0.4$. An example from the WECAN campaign where [CCN] was measured across a range of supersaturations ($\{0.2,0.3,0.4,0.6\}\%$) is shown in Supplementary Fig. 3. We note that although this fit is on the median values in Supplementary Fig. 3 and the median absolute deviations shown with the error bars are not large, there exists the possibility of deviation from the polynomial fit. It must also be noted that the empirical fit is obtained from the ground-based ARM SGP site, which may be in an atmospheric environment drastically different from where the airborne measurements are made. However, compositional effects are ameliorated to an extent by using the ratio of simultaneous [CCNss]:[CCN0.4] rather than [CCNss] alone for this fit. In the absence of sufficient direct measurement of [CCN0.4], this approach, while not ideal, is presently the best possible for evaluating the performance of the machine learning models. While rigorously demonstrating the effectiveness of this approximation is currently impossible due to the airborne observational constraints, we note that there are no significant discontinuities in approximated [CCN0.4] values occurring over the change in instrument supersaturation; additionally, the approximation is consistent with RFRM derived values. For ATom1–4, direct measurement of [CCN] was unavailable and [CCN0.4] is estimated as the number of particles with diameter greater than 60 nm based on Köhler theory[36]. For illustration, Supplementary Fig. 4a shows this estimation applied to the other campaigns' data. Supplementary Fig. 4b illustrates the comparison, in similar vein as Fig. 1c, of RFRM-derived [CCN0.4] to these airborne-PNSD-derived [CCN0.4].

Data imputation. 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 s and 5 s, respectively.

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[45]. These estimators are defined as follows:

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

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

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

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

Machine learning models. A Random Forest Regression Model (RFRM) to derive the number concentrations of cloud condensation nuclei at 0.4% supersaturation ([CCN0.4]) from atmospheric state and composition variables has been developed and is described in detail elsewhere[33]. The present analysis focuses on [CCN0.4] for the purpose of demonstration and in future work will be extensible for the full CCN spectrum. A Random Forest[46] is a machine learning 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[47], each obtained on random subsets[48] of the training data. Here, the RFRM is trained on 30-year simulations by GEOS-Chem-APM: a state-of-the-science chemical transport model with detailed size-resolved microphysics. The present study uses the RFRM-ShortVars configuration[33], a fast implementation[49] of Random Forest models[50] in the statistical computing language R[51], retrained using PM_1 speciation variables as predictors in the absence of airborne measurements of $PM_{2.5}$ speciation. 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).

To identify clusters of similar aerosol composition, k-means clustering[52,53] is applied to partition the four dimensional PM_1(SO_4, NO_3, NH_4, OA)% space into optimally separate Voronoi cells by minimization of variance within each cluster and maximization of silhouetting[54]. To ensure more homogeneity of composition within each cluster, a tolerance of $\pm 2.5\%$ around the highest density is applied. This clustering method is schematically illustrated in Supplementary Figs. 12&13. In obtaining these clusters, no cognitive bias is introduced, yet the statistical basis of clustering is in line with a chemical basis, reflecting varied atmospheric environments (biogenic/urban/biomass-burning/remote).

A Linear Regression model with minimization of least squares using the fast column-

pivoted QR decomposition method[55] is also developed on the airborne dataset. This intentional overfitting is to obtain an effective representation of linear regression-like current GCMs' prescriptions of aerosol mass to number.

While, typically, machine learning models require large amounts of data for their training, we explore the possibility of using the comprehensive airborne measurements to develop an observation-based RFRM (Obs-RFRM). The airborne measurements are split 7 : 3 for training : testing. Illustrated in Supplementary Fig. 8a&b is the performance of the trained Obs-RFRM-PM & Obs-RFRM, respectively, using the airborne measurement data set aside for testing. Obs-RFRM using airborne measurements of the 9 predictors as input derives [CCN0.4] in excellent agreement ($R_K \approx 0.92$ and %-Good $\approx 96\%$) with its airborne measurements. This is however only an academic exercise for the sake of completeness, i.e., for equivalence of panels b&c to the overfit panel a in Fig. 1 as well as for demonstration that the most of the variability of [CCN0.4] can be explained by the 9 predictors considered here. Supplementary Fig. 8c shows this Obs-RFRM applied to a testing dataset for GEOS-Chem-APM (which RFRM and RFRM-PM have not been exposed to during their training, as well). Inset is the RFRM performance with respect to GEOS-Chem-APM. This demonstrates that empirical observations are, at present, still limited in number to develop a practicable and generalizable machine learning model. However, a synergistic approach of measurements–modeling–machine-learning has the potential to provide revealing insights.

Data availability

All datasets used in this paper are publicly available as detailed below.

[CCN0.18–0.86], PNSD, PM_1 composition and mass, $[SO_2]$, $[NO_x]$, $[O_3]$, T, and RH measured during the ARCTAS[56] campaign are publicly available[57] (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[37] campaigns are publicly available[58–61] (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[62] campaign are publicly available[63] (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 are publicly available[64] (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[65] campaign are publicly available[66] (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[67] campaign are publicly available[68] (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 are publicly available[69] (https://www-air.larc.nasa.gov/cgi-bin/ArcView/firexaq).

Dual column CCNc measurement[42,43] 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/), last accessed on October 4 2020.

Data to reproduce the figures and analysis in this paper will be made publicly available on Figshare upon publication.

Code availability

R scripts to reproduce the figures and analysis in this paper will be made publicly available on Figshare upon publication.

References

1. IPCC. *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 University Press, Cambridge, United Kingdom and New York, NY, USA, 2013).

2. Twomey, S. A. Pollution and the planetary albedo. *Atmospheric Environment (1967)* **8**, 1251–1256 (1974).

3. Twomey, S. A. The influence of pollution on the shortwave albedo of clouds. *Journal of the Atmospheric Sciences* **34**, 1149–1152 (1977).

4. Twomey, S. A., Piepgrass, M. & Wolfe, T. An assessment of the impact of pollution on global cloud albedo. *Tellus B* **36B**, 356–366 (1984).

5. Albrecht, B. A. Aerosols, Cloud Microphysics, and Fractional Cloudiness. *Science* **245**, 1227–1230 (1989).

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

7. Pincus, R. & Baker, M. B. Effect of precipitation on the albedo susceptibility of clouds in the marine boundary layer. *Nature* **372**, 250–252 (1994).

8. Hansen, J., Sato, M. & Ruedy, R. Radiative forcing and climate response. *Journal of Geophysical Research: Atmospheres* **102**, 6831–6864 (1997).

9. Ackerman, A. S. *et al.* Reduction of Tropical Cloudiness by Soot. *Science* **288**, 1042–1047 (2000).

10. Ferek, R. J. *et al.* Drizzle suppression in ship tracks. *Journal of the Atmospheric Sciences* **57**, 2707–2728 (2000).

11. Rosenfeld, D. Suppression of rain and snow by urban and industrial air pollution. *Science* **287**, 1793–1796 (2000).

12. Junge, C. & McLaren, E. Relationship of Cloud Nuclei Spectra to Aerosol Size Distribution and Composition. *Journal of the Atmospheric Sciences* **28**, 382–390 (1971).

13. Fitzgerald, J. W. Dependence of the supersaturation spectrum of CCN on aerosol size distribution and composition. *Journal of the Atmospheric Sciences* **30**, 628–634 (1973).

14. Dusek, U. *et al.* Size matters more than chemistry for cloud-nucleating ability of aerosol particles. *Science* **312**, 1375–1378 (2006).

15. Hudson, J. G. Variability of the relationship between particle size and cloud-nucleating ability. *Geophysical Research Letters* **34** (2007).

16. Twohy, C. H. & Anderson, J. R. Droplet nuclei in non-precipitating clouds: composition and size matter. *Environmental Research Letters* **3**, 045002 (2008).

17. Crosbie, E. *et al.* 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**, 6943–6958 (2015).

18. Boucher, O. & Lohmann, U. The sulfate-CCN-cloud albedo effect. *Tellus B: Chemical and Physical Meteorology* **47**, 281–300 (1995).

19. Menon, S., Genio, A. D. D., Koch, D. & Tselioudis, G. GCM simulations of the aerosol indirect effect: Sensitivity to cloud parameterization and aerosol burden. *Journal of the Atmospheric Sciences* **59**, 692–713 (2002).

20. Menon, S. & Rotstayn, L. The radiative influence of aerosol effects on liquid-phase cumulus and stratiform clouds based on sensitivity studies with two climate models. *Climate Dynamics* **27**, 345–356 (2006).

21. Reichstein, M. *et al.* Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195–204 (2019).

22. Dramsch, J. S. 70 years of machine learning in geoscience in review. In *Machine Learning in Geosciences*, 1–55 (Elsevier, 2020).

23. Joutsensaari, J. *et al.* Identification of new particle formation events with deep learning. *Atmospheric Chemistry and Physics* **18**, 9597–9615 (2018).

24. Zaidan, M. A. *et al.* Predicting atmospheric particle formation days by Bayesian classification of the time series features. *Tellus B: Chemical and Physical Meteorology* **70**, 1–10 (2018).

25. Christopoulos, C. D., Garimella, S., Zawadowicz, M. A., Möhler, O. & Cziczo, D. J. A machine learning approach to aerosol classification for single-particle mass spectrometry. *Atmospheric Measurement Techniques* **11**, 5687–5699 (2018).

26. Hughes, M., Kodros, J., Pierce, J., West, M. & Riemer, N. Machine learning to predict the global distribution of aerosol mixing state metrics. *Atmosphere* **9**, 15 (2018).

27. Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E. & Hueglin, C. Random forest meteorological normalisation models for swiss PM_{10} trend analysis. *Atmospheric Chemistry and Physics* **18**, 6223–6239 (2018).

28. Fuchs, J., Cermak, J. & Andersen, H. Building a cloud in the southeast Atlantic: understanding low-cloud controls based on satellite observations with machine learning. *Atmospheric Chemistry and Physics* **18**, 16537–16552 (2018).

29. Mauceri, S., Kindel, B., Massie, S. & Pilewskie, P. Neural network for aerosol retrieval from hyperspectral imagery. *Atmospheric Measurement Techniques* **12**, 6017–6036 (2019).

30. Okamura, R., Iwabuchi, H. & Schmidt, K. S. Feasibility study of multi-pixel retrieval of optical thickness and droplet effective radius of inhomogeneous clouds using deep learning. *Atmospheric Measurement Techniques* **10**, 4747–4759 (2017).

31. Dou, X. & Yang, Y. Comprehensive evaluation of machine learning techniques for estimating the responses of carbon fluxes to climatic forces in different terrestrial ecosystems. *Atmosphere* **9**, 83 (2018).

32. Jin, J., Lin, H. X., Segers, A., Xie, Y. & Heemink, A. Machine learning for observation bias correction with application to dust storm data assimilation. *Atmospheric Chemistry and Physics* **19**, 10009–10026 (2019).

33. Nair, A. A. & Yu, F. Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements. *Atmospheric Chemistry and Physics* **20**, 12853–12869 (2020).

34. Martin, G. M., Johnson, D. W. & Spice, A. The measurement and parameterization of effective radius of droplets in warm stratocumulus clouds. *Journal of the Atmospheric Sciences* **51**, 1823–1842 (1994).

35. Ramanathan, V. Aerosols, climate, and the hydrological cycle. *Science* **294**, 2119–2124 (2001).

36. Williamson, C. J. *et al.* A large source of cloud condensation nuclei from new particle formation in the tropics. *Nature* **574**, 399–403 (2019).

37. Brock, C. A. *et al.* Aerosol size distributions during the Atmospheric Tomography Mission (ATom): methods, uncertainties, and data products. *Atmospheric Measurement Techniques* **12**, 3081–3099 (2019).

38. Moore, R. H. & Nenes, A. Scanning Flow CCN Analysis—A Method for Fast Measurements of CCN Spectra. *Aerosol Science and Technology* **43**, 1192–1207 (2009).

39. Roberts, G. C. & Nenes, A. A Continuous-Flow Streamwise Thermal-Gradient CCN Chamber for Atmospheric Measurements. *Aerosol Science and Technology* **39**, 206–221 (2005).

40. Uin, J. Cloud Condensation Nuclei Particle Counter (CCN) Instrument Handbook. Tech. Rep., DOE Office of Science Atmospheric Radiation Measurement (ARM) Program (2016).

41. Uin, J. & Smith, S. Southern Great Plains (SGP) Aerosol Observing System (AOS) Instrument Handbook (2021).

42. Uin, J., Salwen, C. & Senum, G. 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) (2017). URL https://adc.arm.gov/discovery/#/results/datastream::sgpaosccn2colaE13.b1.

43. Uin, J., Salwen, C. & Senum, G. 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) (2017). URL https://adc.arm.gov/discovery/#/results/datastream::sgpaosccn2colbE13.b1.

44. Twomey, S. 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**, 243–249 (1959).

45. Nair, A. A., Yu, F. & Luo, G. Spatioseasonal Variations of Atmospheric Ammonia Concentrations Over the United States: Comprehensive Model-Observation Comparison. *Journal of Geophysical Research: Atmospheres* **124**, 6571–6582 (2019).

46. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).

47. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification And Regression Trees* (Routledge, 1984).

48. Breiman, L. Bagging predictors. *Machine Learning* **24**, 123–140 (1996).

49. Wright, M. N. & Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software* **77** (2017).

50. Breiman, L. *Manual for Setting Up, Using, and Understanding Random Forest V4.0* (2003). https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.

51. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2020). URL https://www.R-project.org/.

52. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**, 129–137 (1982).

53. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics* **28**, 100 (1979).

54. Kaufman, L. & Rousseeuw, P. J. (eds.) *Finding Groups in Data* (John Wiley & Sons, Inc., 1990).

55. Bates, D. & Eddelbuettel, D. Fast and Elegant Numerical Linear Algebra Using the RcppEigenPackage. *Journal of Statistical Software* **52** (2013).

56. Jacob, D. J. *et al.* 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**, 5191–5212 (2010).

57. ARCTAS Team. ARCTAS Field Campaign Data (2020). URL https://www-air.larc.nasa.gov/cgi-bin/ArcView/arctas.

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

59. Ryerson, T. B., Thompson, C. R., Peischl, J. & Bourgeois, I. ATom: L2 In Situ Measurements from NOAA Nitrogen Oxides and Ozone (NOyO3) Instrument (2019). URL https://espo.nasa.gov/atom/archive/browse/atom/DC8/.

60. Jimenez, J. L. *et al.* ATom: L2 Measurements from CU High-Resolution Aerosol Mass Spectrometer (HR-AMS) (2019). URL https://espo.nasa.gov/atom/archive/browse/atom/DC8/AMS.

61. Brock, C. A. *et al.* ATom: L2 In Situ Measurements of Aerosol Microphysical Properties (AMP) (2019). URL https://espo.nasa.gov/atom/archive/browse/atom/DC8/SDAerosol.

62. Barth, M. C. *et al.* The deep convective clouds and chemistry (DC3) field campaign. *Bulletin of the American Meteorological Society* **96**, 1281–1309 (2015).

63. DC3 Team. DC3 Field Campaign Data (2013). URL https://www-air.larc.nasa.gov/missions/dc3-seac4rs/.

64. DISCOVER-AQ Team. DISCOVER-AQ Field Campaign Data (2014). URL https://www-air.larc.nasa.gov/cgi-bin/ArcView/discover-aq.tx-2013.

65. Jordan, C. E. *et al.* Investigation of factors controlling PM2.5 variability across the south korean peninsula during KORUS-AQ. *Elementa: Science of the Anthropocene* **8** (2020).

66. KORUS-AQ Team. KORUS-AQ Field Campaign Data (2018). URL https://www-air.larc.nasa.gov/missions/korus-aq/.

67. Toon, O. B. *et al.* 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**, 4967–5009 (2016).

68. SEAC4RS Team. SEAC4RS Field Campaign Data (2014). URL https://www-air.larc.nasa.gov/cgi-bin/ArcView/seac4rs.

69. WE-CAN Team. WE-CAN Field Campaign Data (2019). URL https://www-air.larc.nasa.gov/cgi-bin/ArcView/firexaq?MERGE=1.

Acknowledgments

We 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). We 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. We also thank the DOE ARM SGP Research Facility data teams for the operation and maintenance of instruments, quality checks, and making their measurement data publicly available.

Funding

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). B.A.N., P.C.J. & J.L.J. were supported by NASA (grant nos. NNX15AJ23G, NNX15AH33A, 80NSSC19K0124, and 80NSSC18K0630). P.J.D. & E.J.T.L. 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. M.P. & S.S.Y. acknowledge the support from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (grant no. NRF-2018R1A2B2006965). M.J.K. was supported by an NSF Atmospheric and Geospace Sciences Postdoctoral Research Fellowship (AGS-PRF; grant no. 1524860). C.D.F., B.B.P. & Q.P. acknowledge support from the NSF (grant no. AGS-1652688) and the National Oceanic and Atmospheric Administration (NOAA grant no. NA17OAR4310012).

Author information

Benjamin A. Nault

Present address: Center for Aerosols and Cloud Chemistry, Aerodyne Research, Inc., Billerica, Massachusetts 01821, USA

Ezra J. T. Levin

Present address: Handix Scientific, Boulder, Colorado 80301, USA

Affiliations

Atmospheric Sciences Research Center, State University of New York, Albany, New York 12203, USA

Arshad Arjunan Nair & Fangqun Yu

Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, Colorado 80309, USA

Pedro Campuzano-Jost, Jose L. Jimenez, Jeff Peischl, Benjamin A. Nault, Chelsea R. Thompson & Ilann Bourgeois

Department of Chemistry, University of Colorado Boulder, Boulder, Colorado 80309, USA

Pedro Campuzano-Jost, Jose L. Jimenez & Benjamin A. Nault

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado 80523, USA
Paul J. DeMott, Ezra J. T. Levin & Ilana B. Pollack

NOAA Chemical Science Laboratory, Boulder, Colorado 80305, USA
Jeff Peischl, Ilann Bourgeois, Chelsea R. Thompson, Thomas B. Ryerson

Department of Atmospheric Sciences, University of Washington, Seattle, Washington 98195, USA
Carley D. Fredrickson, Brett B. Palm & Qiaoyun Peng

NASA Langley Research Center, Hampton, Virginia 23666, USA
Andreas J. Beyersdorf, Bruce E. Anderson, Luke D. Ziemba & Richard H. Moore

California State University, San Bernardino, California 92407, USA
Andreas J. Beyersdorf

Department of Atmospheric Sciences, Yonsei University, Seoul 03722, South Korea
Minsu Park & Seong Soo Yum

Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA
Lu Xu, & Michelle J. Kim

Institute of Chemical Engineering Sciences, Foundation for Research & Technology-Hellas, Patras 26504, Greece
Athanasios Nenes

School of Architecture, Civil and Environmental Engineering, Swiss Federal Institute of Technology Lausanne, Lausanne 1015, Switzerland
Athanasios Nenes

School of Earth and Atmospheric Sciences and School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA
Athanasios Nenes

Department of Environmental Science, Hankuk University of Foreign Studies, Yongin 17035, South Korea
Taehyoung Lee & Taehyun Park

Atmospheric Chemistry Observations and Modeling, National Center for Atmospheric Research, Boulder, Colorado 80307, USA
Frank Flocke

School of Earth and Atmospheric Sciences, Georgia Tech, Atlanta, Georgia 30332, USA
Lewis Gregory Huey

Contributions

A.A.N. and F.Y. conceptualized this study. B.E.A., A.N., A.J.B., L.D.Z., R.H.M., M.P., S.S.Y., E.J.T.L., P.J.D., P.C.J., B.A.N., B.B.P., J.L.J., T.L., T.P., C.R.T., F.F., I.B.P., I.B., J.P., L.X., M.J.K., L.G.H., C.D.F., and Q.P. carried out, analyzed, and archived airborne measurements. A.A.N. led the analysis and writing with major input from F.Y. and further input from P.C.J., P.J.D., E.J.T.L., J.L.J., J.P., I.B.P., C.D.F., A.J.B., B.A.N., M.P., S.S.Y., B.B.P., L.X., and I.B..

Corresponding authors

Correspondence to Arshad Arjunan Nair (aanair@albany.edu) and Fangqun Yu (fyu@albany.edu).

Ethics declarations

Competing interests

The authors declare no competing interests.

Figures

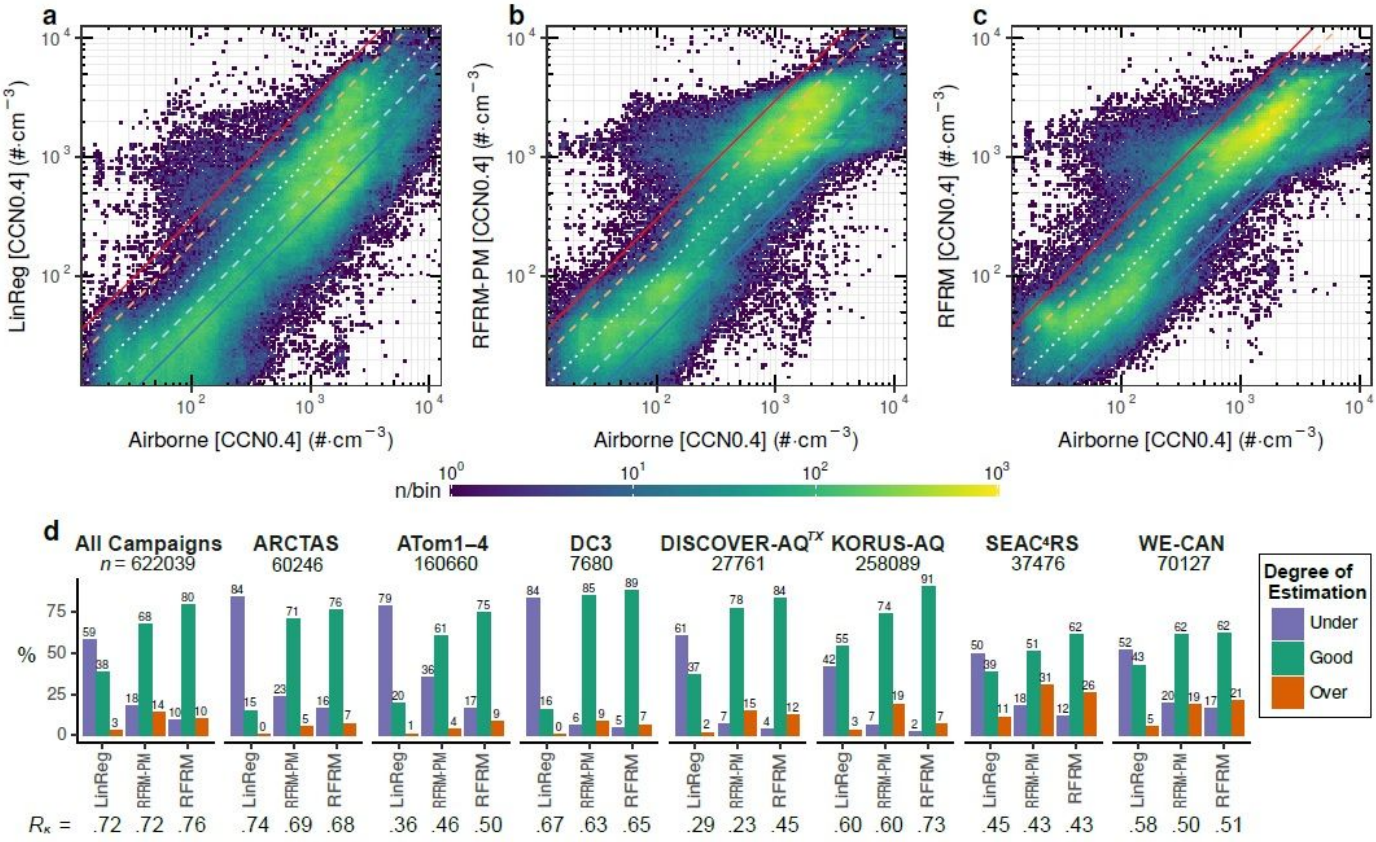

Figure 1

Comparison of machine learning derived versus airborne measurements of [CCN0.4]. Binned scatter plot for data at the 1Hz resolution from all campaigns. For a Linear Regression (LinReg), b RFRM-PM, and c RFRM. Central 99% range of the airbornemeasured [CCN0.4] shown for a zoomed-in view. The lines, in the order of decreasing yintercept, 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. d Summary statistics for the degree of model{observation agreement and correlation, as defined in the Methods, for each aircraft campaign.

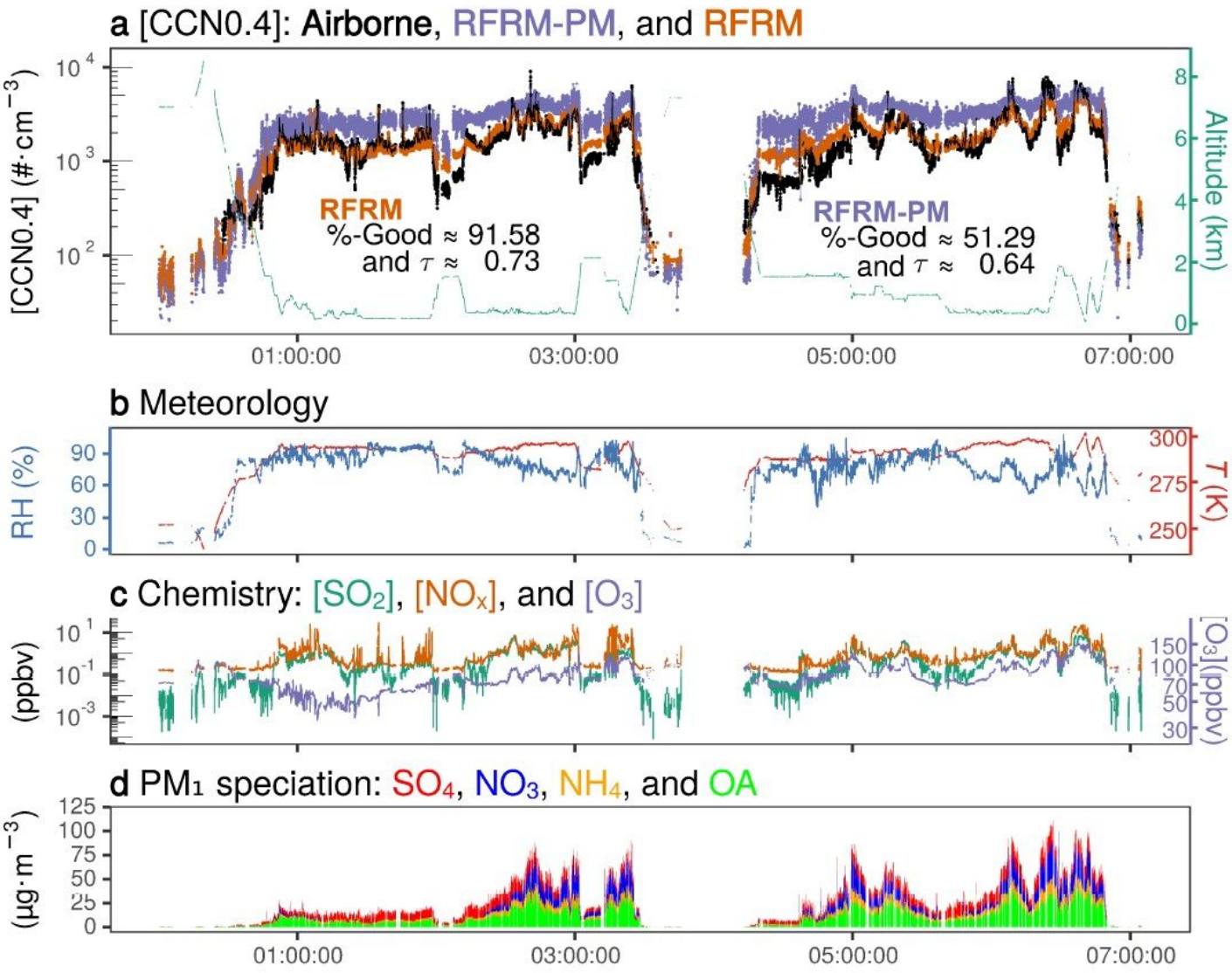

Figure 2

Time series of [CCN0.4] and variables of atmospheric state and composition shown for a selected campaign day (KORUS-AQ: 10 June 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) [SO2], (orange) [NOx], and (blue) [O3]. d PM1 speciated masses of (red) SO4, (blue) NO3, (orange) NH4, and (green) OA. Data is shown at the 1 Hz resolution. Solid lines associated with [CCN0.4] are 5 s rolling means.

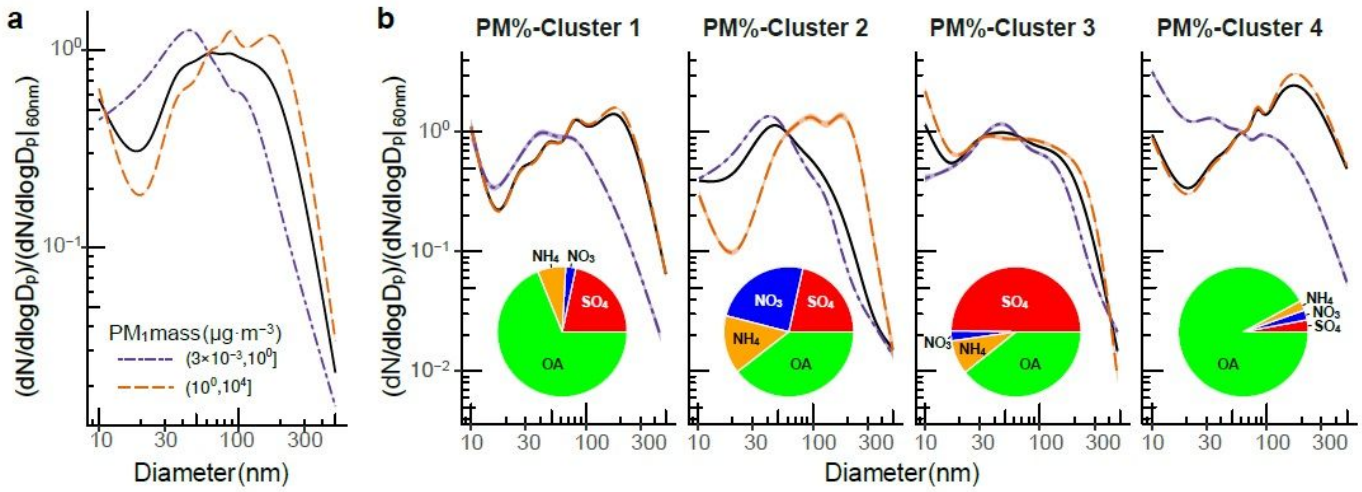

Figure 3

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, dot-dashed) PMtot < 1 μg cm-3, and (orange, dashed) PMtot > 1 μg cm-3. Solid black curve in a is for all data. b For each cluster: Cluster 1 (SO4: 19{24%, OA: 66{71%, NO3: 0{5%, and NH4: 4.5{9.5%), Cluster 2 (SO4: 19{24%, OA: 37{42%, NO3: 22{27%, and NH4: 12{17%), Cluster 3 (SO4: 47.5{ 52.5%, OA: 37{42%, NO3: 0{5%, and NH4: 6{11%), Cluster 4 (SO4: 0.5{5.5%, OA: 91{96%, NO3: 0{5%, and NH4: 0{5%), and (black) respective cluster-wise average. Typical aerosol composition for each cluster is illustrated by the inset pie charts.

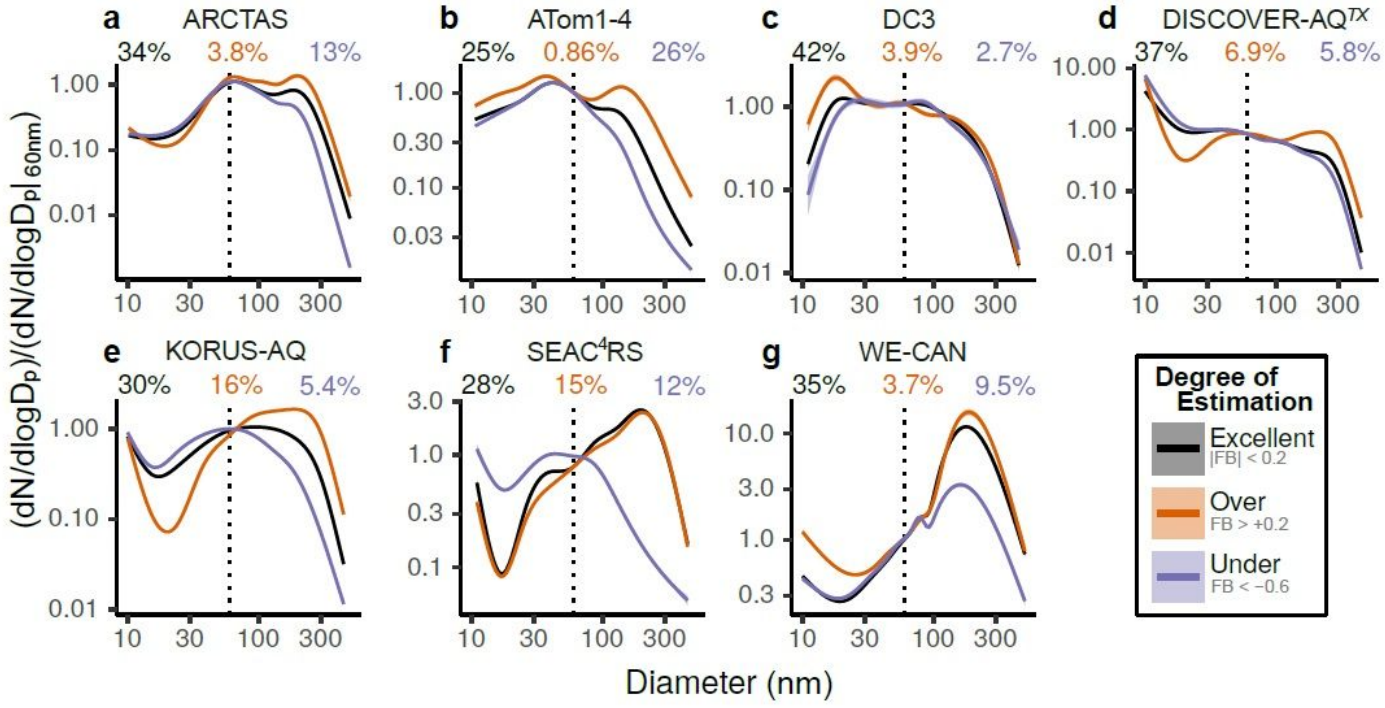

Figure 4

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-AQTX, e KORUS-AQ, f SEAC4RS, 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf