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

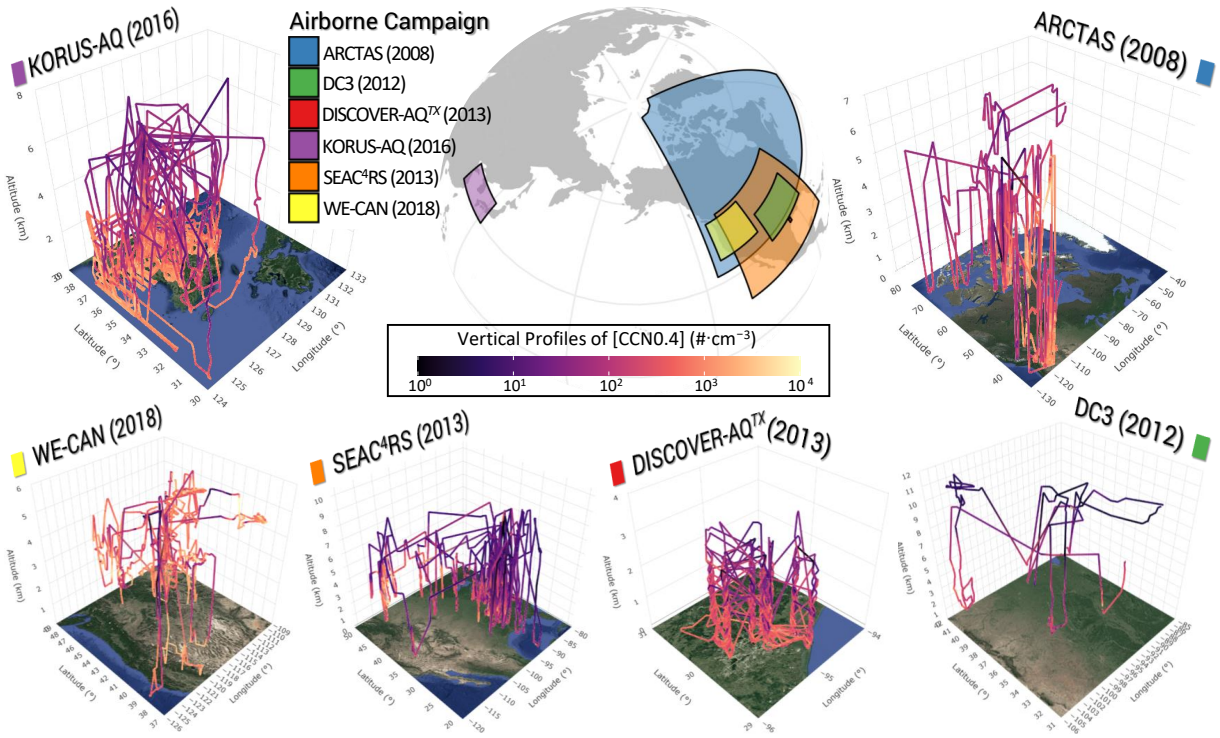
Arshad Arjunan Nair*, Fangqun Yu*, Pedro Campuzano-Jost, Paul J. DeMott, Ezra J. T. Levin, Jose L. Jimenez, Jeff Peischl, Ilana B. Pollack, Carley D. Fredrickson, Andreas J. Beyersdorf, Benjamin A. Nault, Minsu Park, Seong Soo Yum, Brett B. Palm, Lu Xu, Ilann Bourgeois, Bruce E. Anderson, Athanasios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

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

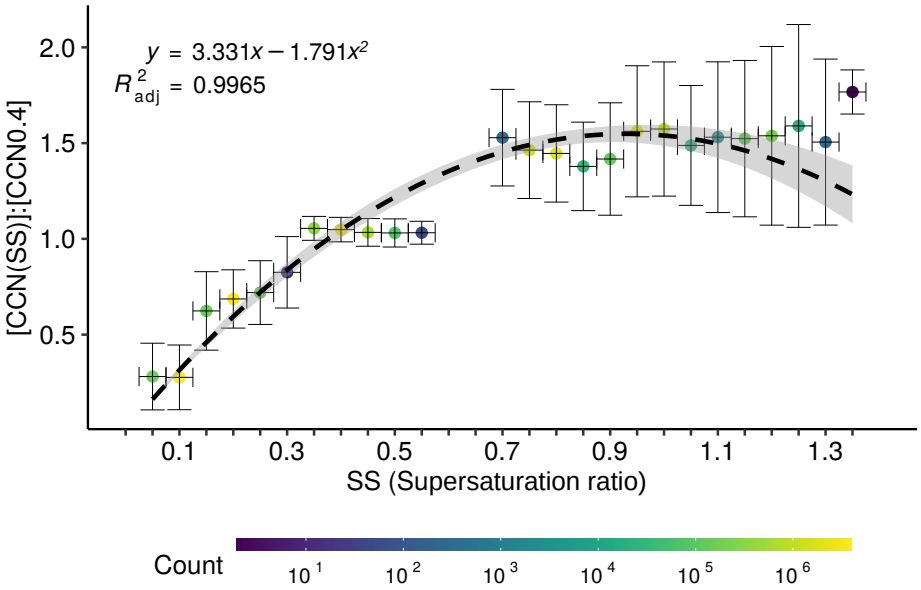
Contents

- Supplementary Figures 1–13
- Supplementary Tables 1 & 2
- Supplementary References

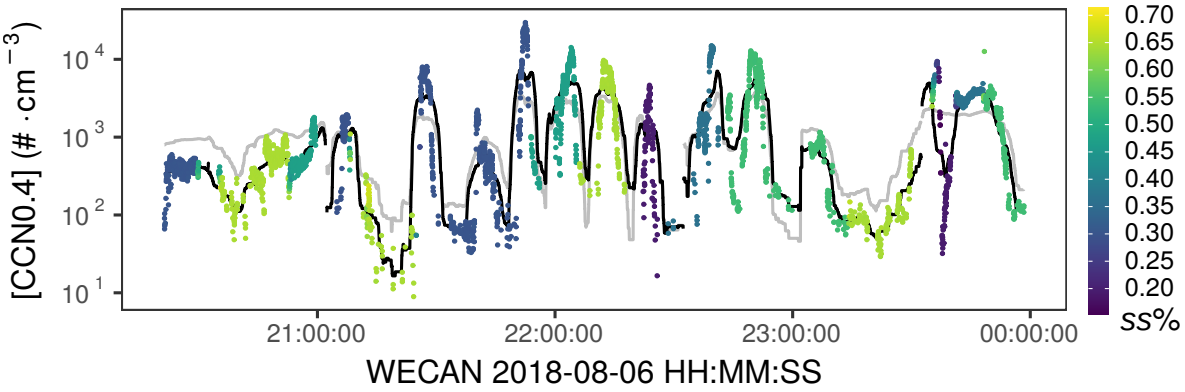
Supplementary Figures



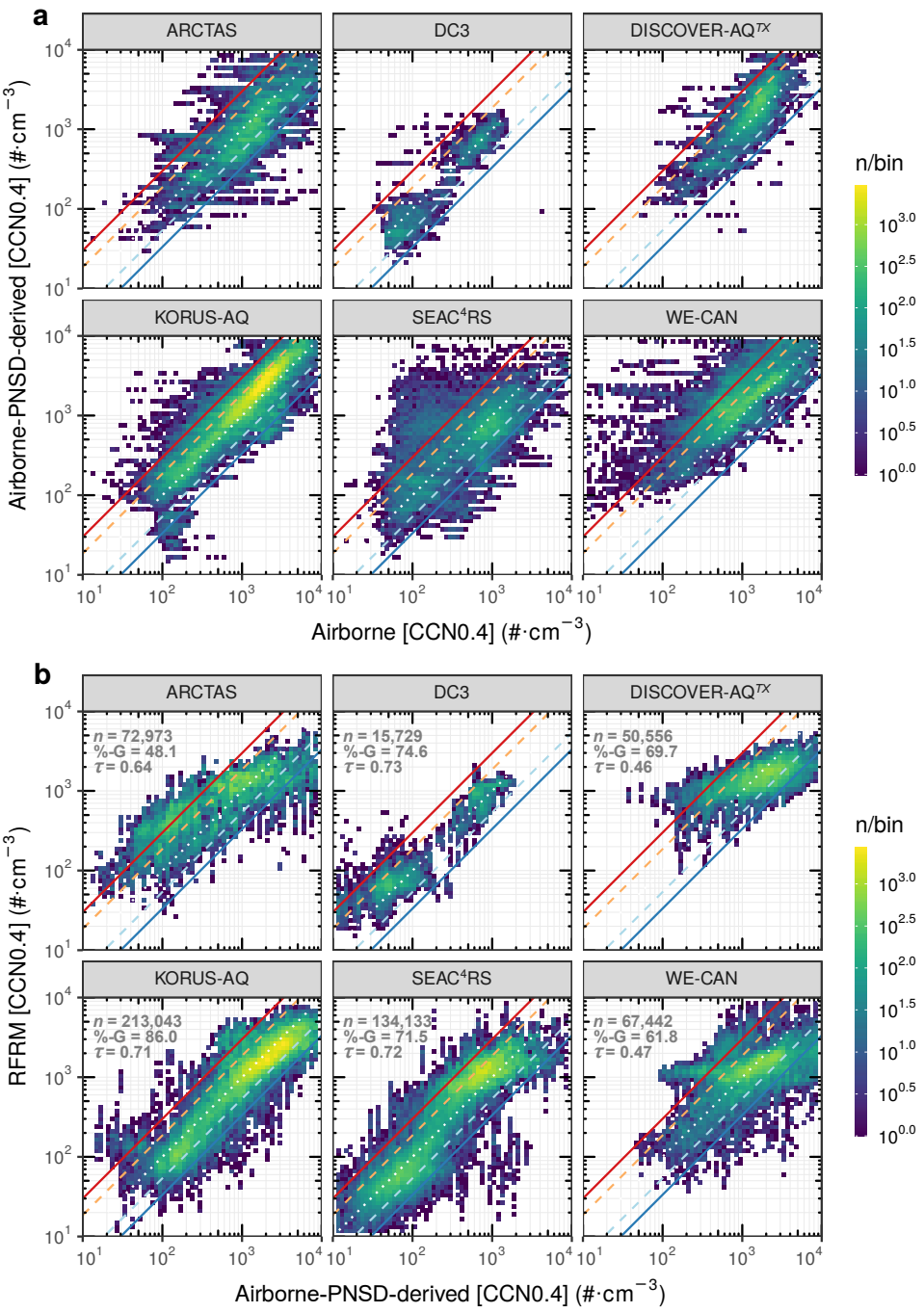
Supplementary Figure 1: **Observational domain and vertical [CCN0.4] profiles.** For six airborne measurement campaigns: KORUS-AQ[1] (2016), ARCTAS[2] (2008), WE-CAN (2018), SEAC^{4}RS[3] (2013), DISCOVER-AQ^{TX} (2013), and DC3[4] (2012). ATom1–4 (2016–18), which has a global domain and indirect measurement of [CCN0.4] is not shown here and can be found elsewhere[5]. Flights' paths shown in color over satellite imagery (© 2020 TerraView obtained through the Google Maps Static API). Colorbar (log-scale) shows the [CCN0.4] number concentration observed along the flight path only when all 9 variables of atmospheric state and composition (RFRM predictors) are also available.



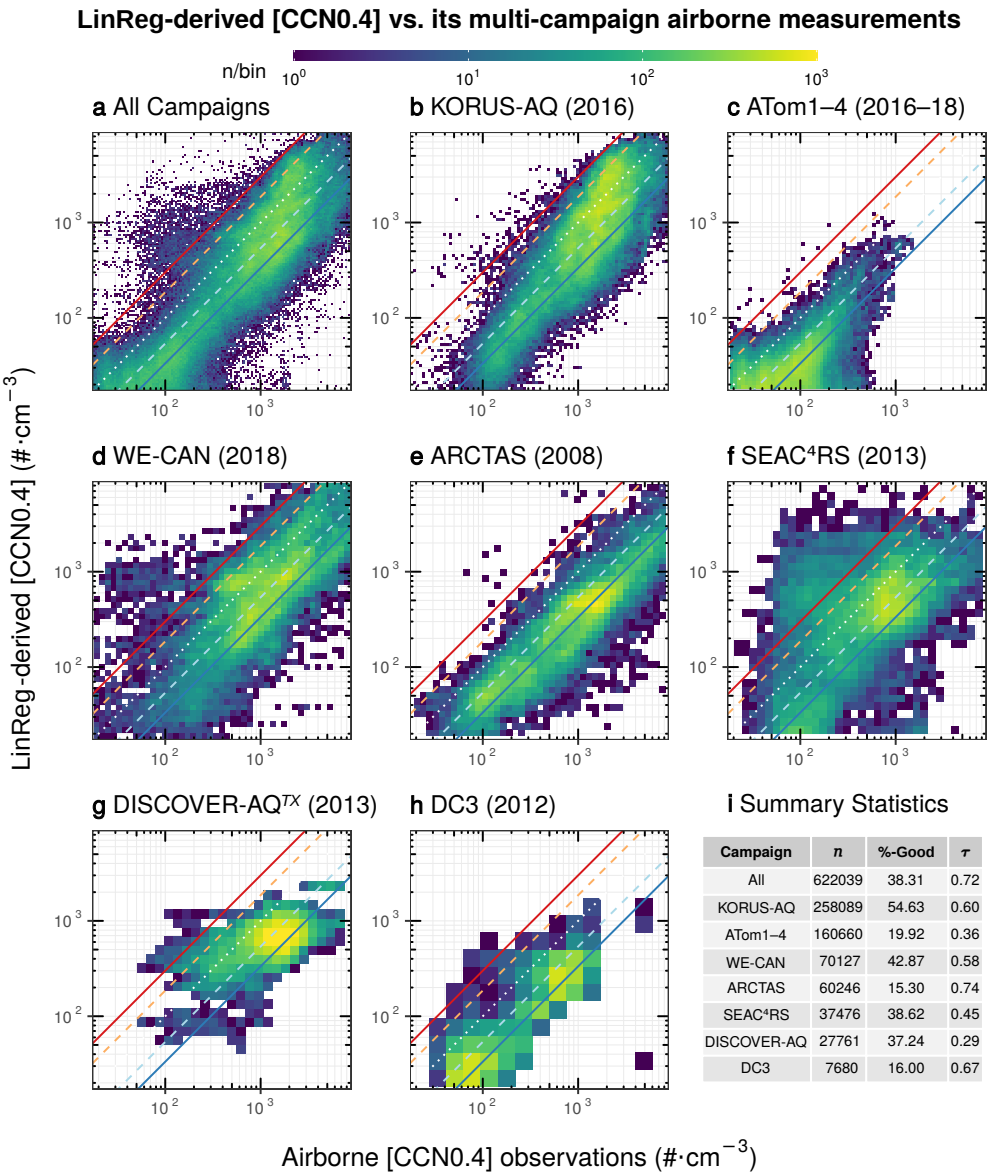
Supplementary Figure 2: **Empirical fit function for approximating [CCN0.4].** Ratio of CCN at various supersaturation ratios to CCN at 0.4% supersaturation versus supersaturation ratio from the SGP dual column CCNc. A polynomial fit is obtained to approximate [CCN0.4] from airborne measurements of [CCN]. Points show the median and vertical error bars show the median absolute deviation. Horizontal error bars show the supersaturation ratio bin range. Logscale color bar shows the number of observations associated with each supersaturation rounded to the nearest 0.05.



Supplementary Figure 3: **Example application of [CCN0.4] approximation when** $ss \neq 0.4$**.** Chosen is a day with largest variations in CCNc instrument supersaturations. Points show [CCN0.4] approximated per the polynomial fit in Supplementary Figure 2 with its colors corresponding to the instrument supersaturation. The 5 minute rolling mean is shown by the lines: (black) [CCN0.4] approximated from measurements of [CCNss] and (grey) RFRM-derived [CCN0.4] from the 9 predictors of atmospheric state and composition.

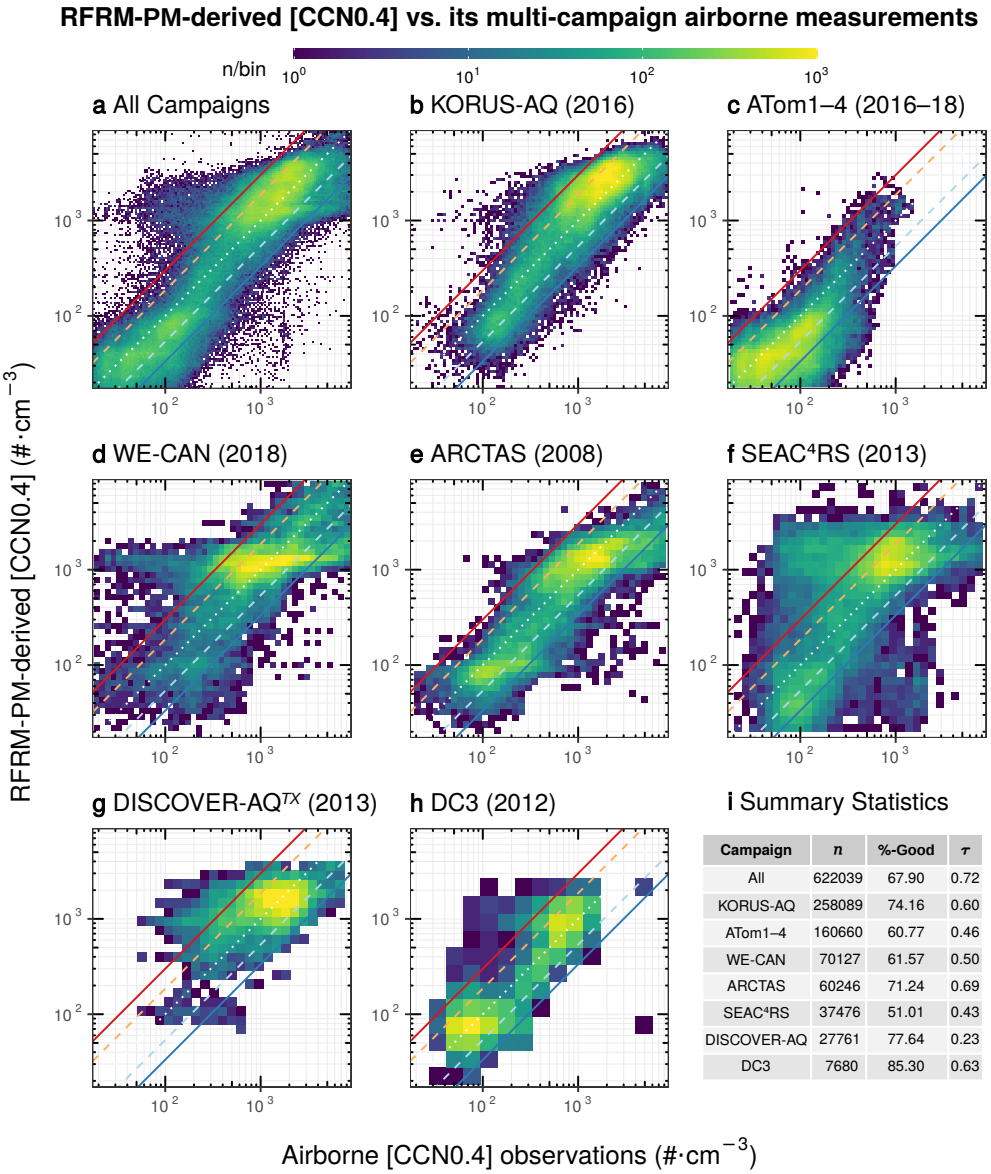


Supplementary Figure 4: **Approximating [CCN0.4] from PNSD when CCN is not directly measured.** Application to other campaigns of the same approach— count of particles larger than a cut-off diameter of 60 nm —used to determine ATom [CCN0.4] from PNSD data. **a** [CCN0.4] derived from airborne PNSD measurements versus that measured using a CCNc. **b** [CCN0.4] derived by the RFRM versus that derived from PNSD. Inset are the statistical metrics (%-G: %-Good and τ: correlation) corresponding to Figure 1**d**.



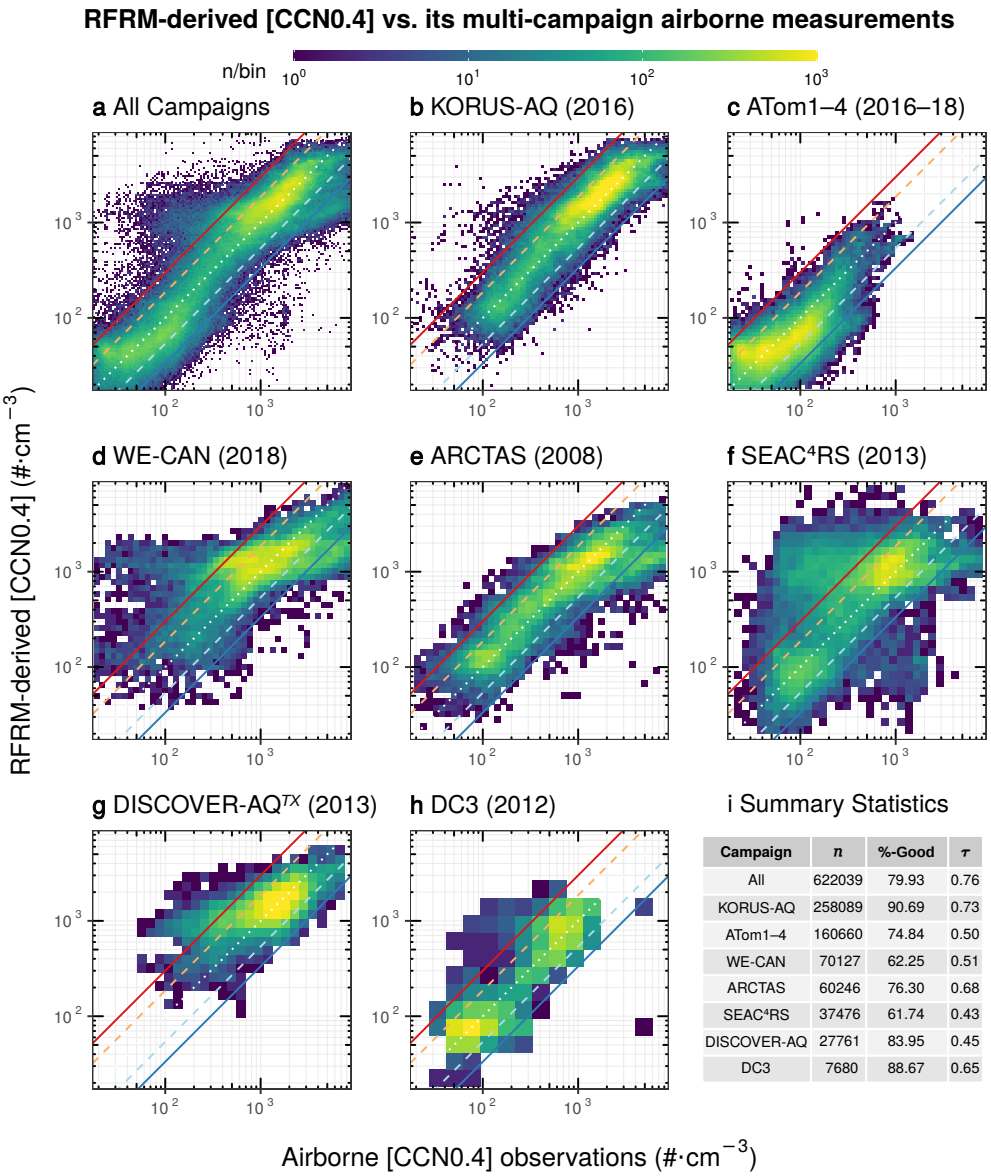
Campaign	n	%-Good	τ
All	622039	38.31	0.72
KORUS-AQ	258089	54.63	0.60
ATom1–4	160660	19.92	0.36
WE-CAN	70127	42.87	0.58
ARCTAS	60246	15.30	0.74
SEAC⁴RS	37476	38.62	0.45
DISCOVER-AQ	27761	37.24	0.29
DC3	7680	16.00	0.67

Supplementary Figure 5: **Comparison of linear regression (LinReg) derived versus airborne measurements of [CCN0.4].** Binned scatter plot for data at the 1Hz resolution for **a** all campaigns and **b**–**h** each campaign. 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 MFB 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 on the log-scale is 0.02 (for comparison with main manuscript Figure 1) times the ratio of the square root of the total number of observations divided by the square root of the number of observations corresponding to each campaign, such that bin area is normalized to the number of observations. **i** Summary statistics for the degree of model–observation agreement and correlation, in decreasing order of the number of observations per campaign.



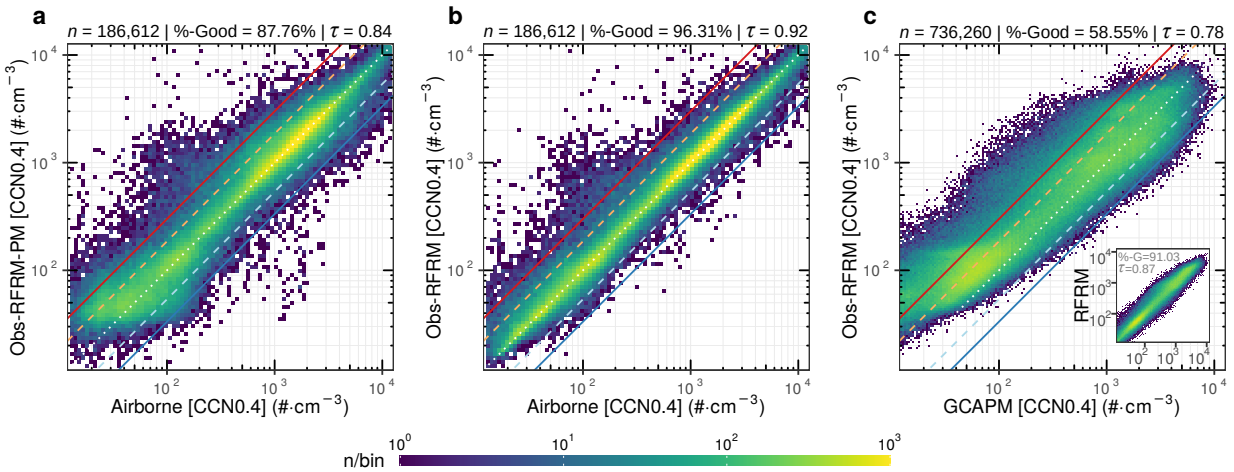
Campaign	n	%-Good	τ
All	622039	67.90	0.72
KORUS-AQ	258089	74.16	0.60
ATom1–4	160660	60.77	0.46
WE-CAN	70127	61.57	0.50
ARCTAS	60246	71.24	0.69
SEAC⁴RS	37476	51.01	0.43
DISCOVER-AQ	27761	77.64	0.23
DC3	7680	85.30	0.63

Supplementary Figure 6: **Comparison of RFRM-PM (with only PM_1 speciation predictors) derived versus airborne measurements of [CCN0.4].** Binned scatter plot for data at the 1Hz resolution for **a** all campaigns and **b**–**h** each campaign. 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 MFB 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 on the log-scale is 0.02 (for comparison with main manuscript Figure 1) times the ratio of the square root of the total number of observations divided by the square root of the number of observations corresponding to each campaign, such that bin area is normalized to the number of observations. **i** Summary statistics for the degree of model–observation agreement and correlation, in decreasing order of the number of observations per campaign.

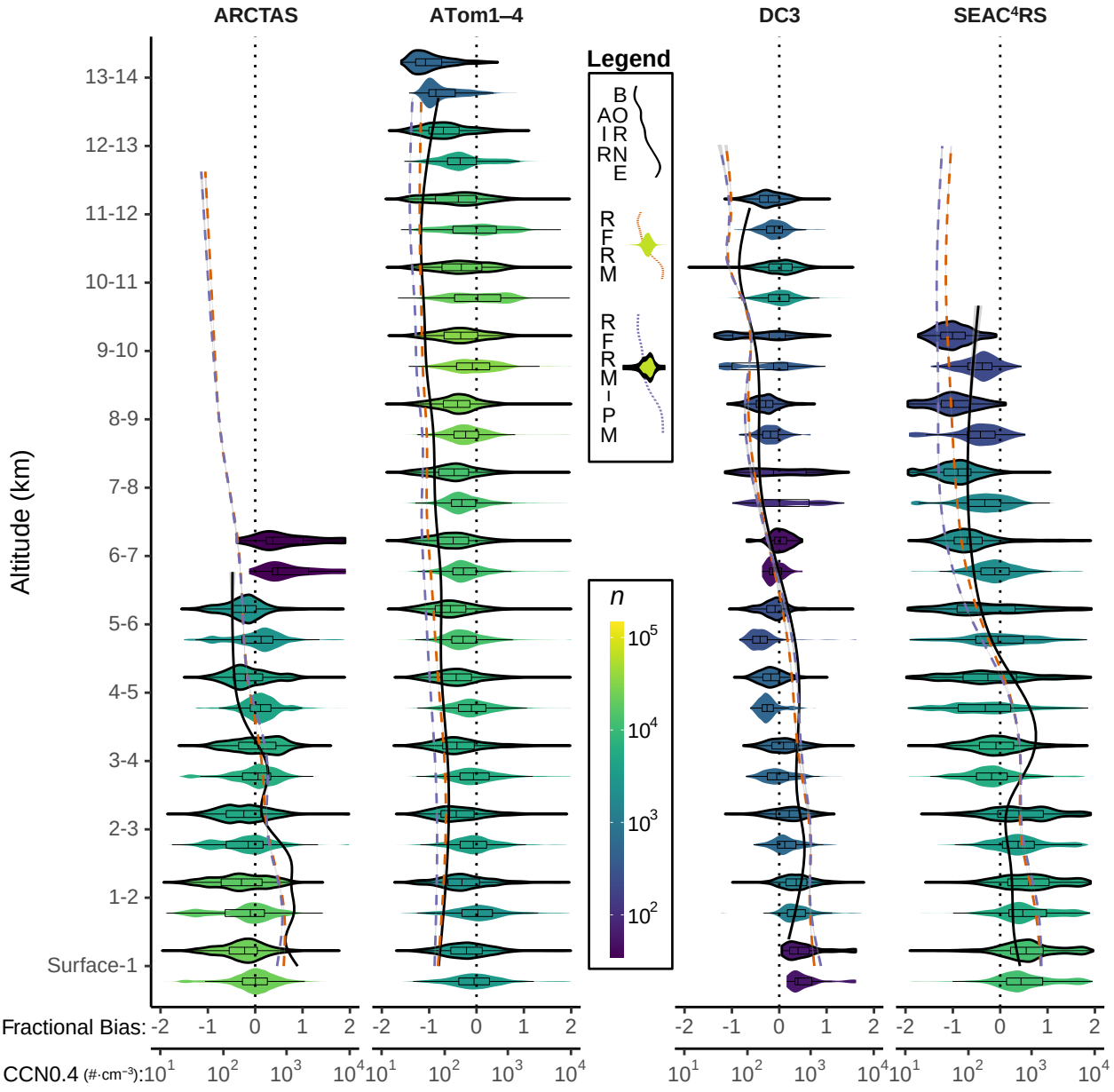


Campaign	n	%-Good	τ
All	622039	79.93	0.76
KORUS-AQ	258089	90.69	0.73
ATom1–4	160660	74.84	0.50
WE-CAN	70127	62.25	0.51
ARCTAS	60246	76.30	0.68
SEAC⁴RS	37476	61.74	0.43
DISCOVER-AQ	27761	83.95	0.45
DC3	7680	88.67	0.65

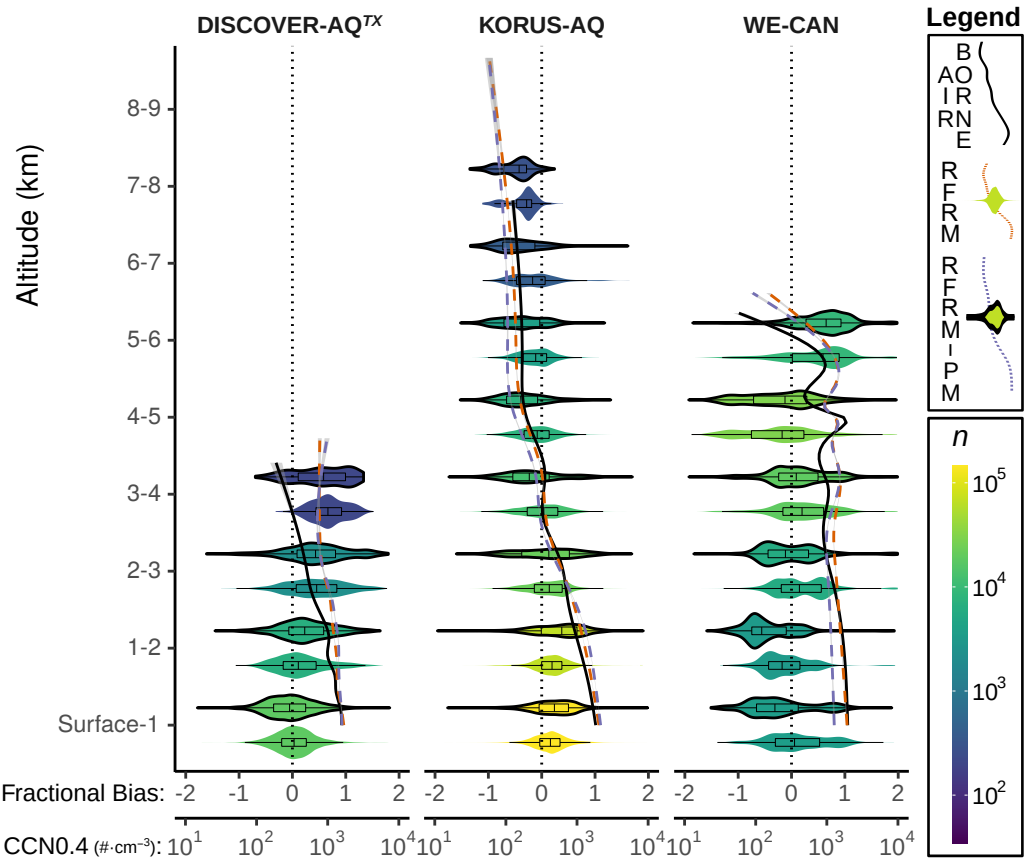
Supplementary Figure 7: **Comparison of RFRM (with 9 predictors) derived versus airborne measurements of [CCN0.4].** Binned scatter plot for data at the 1Hz resolution for **a** all campaigns and **b**–**h** each campaign. Central 99% range of the airborne-measured [CCN0.4] shown for a zoomed-in view. The lines, in the order of decreasing y-intercept, indicate MFB 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 on the log-scale is 0.02 (for comparison with main manuscript Figure 1) times the ratio of the square root of the total number of observations divided by the square root of the number of observations corresponding to each campaign, such that bin area is normalized to the number of observations. **i** Summary statistics for the degree of model–observation agreement and correlation, in decreasing order of the number of observations per campaign.



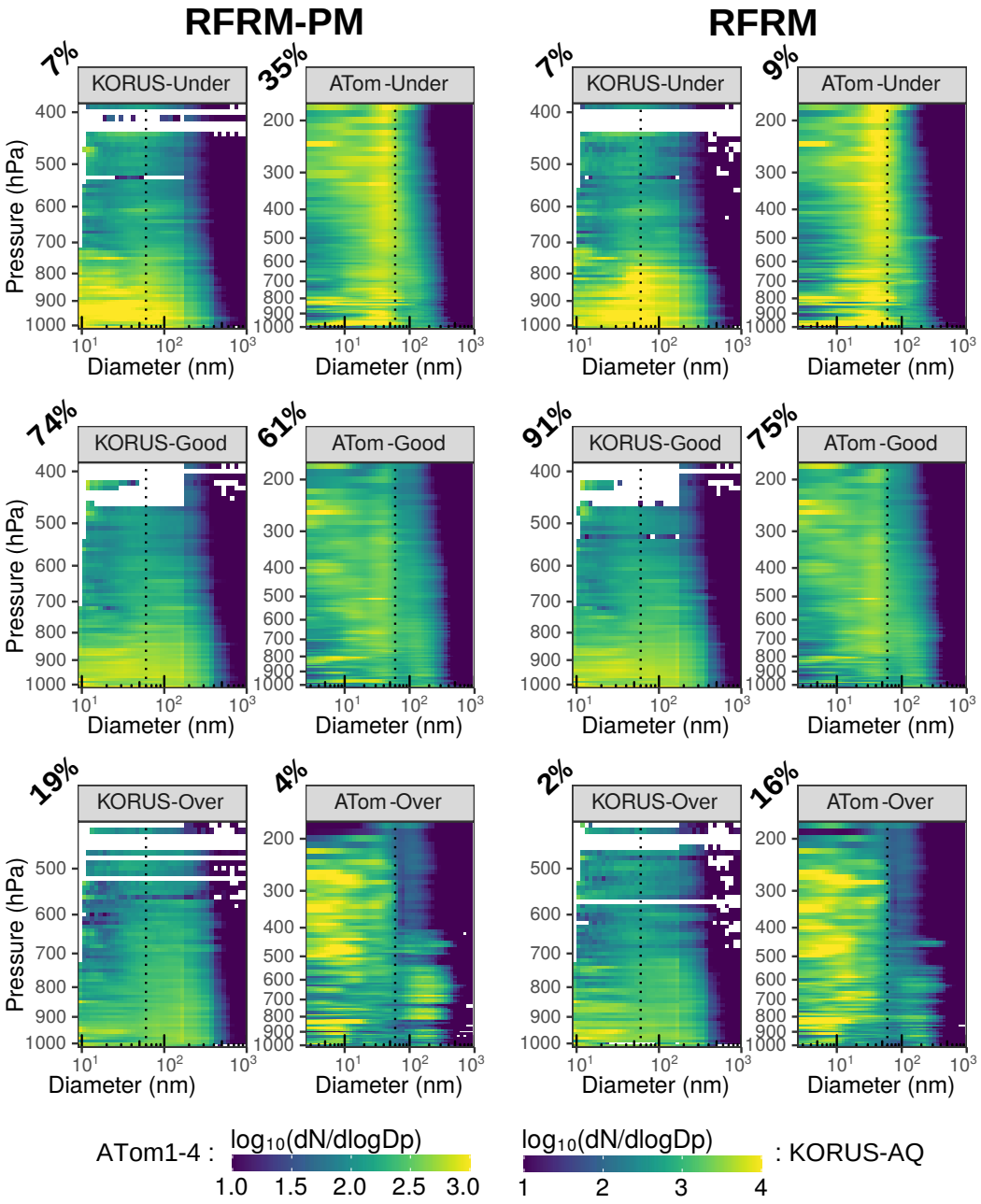
Supplementary Figure 8: **Comparison of machine learning derived versus expected values of [CCN0.4].** Binned scatter plot for [CCN0.4] derived using **a** RFRM trained on airborne data using only PM_1 speciation as predictors (Obs-RFRM-PM), **b** RFRM trained on airborne data using all 9 predictors (Obs-RFRM) versus airborne measurements of [CCN0.4].**c** Obs-RFRM derived [CCN0.4] using GEOS-Chem-APM (GCAPM) input predictors versus GCAPM [CCN0.4]. Inset in **c** is RFRM-derived versus GCAPM [CCN0.4]. x-axis limited to central 99% of the data 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 that of Figure 1 scaled to the number of data points (n).



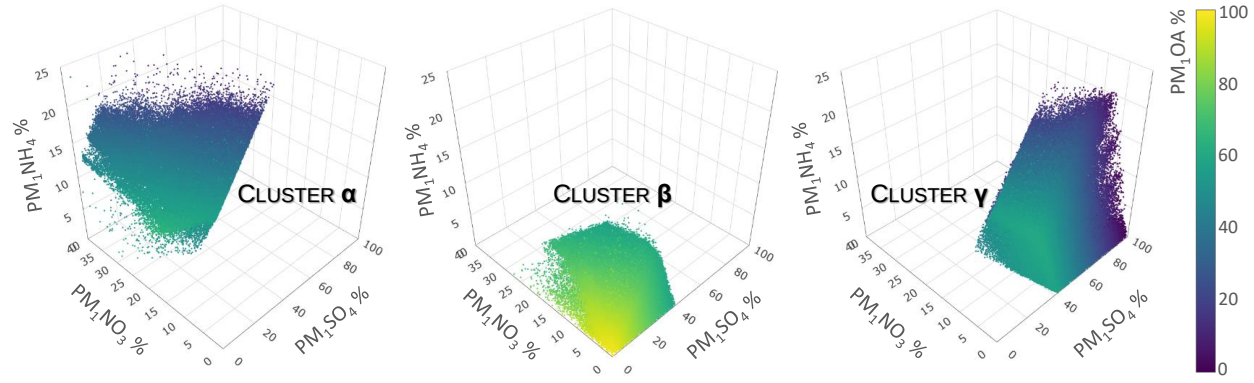
Supplementary Figure 9: **Model performance along the tropospheric vertical extent.** RFRM-derived (dashed orange curves; borderless violins) and RFRM-PM-derived (dashed purple; bordered violins) compared to airborne-measured (black) [CCN0.4] across the vertical tropospheric extent. Vertical curves are the generalized additive model fits with confidence interval shaded in grey. Violin plots show the distribution of fractional bias at each 1 km layer. Color bar (log scale) shows the number of observations in each violin. Also shown are the simple box plots, without the kernel density. Although average (generalized additive model fits) for RFRM-derived and RFRM-PM are both in good agreement with airborne-measurements and the fractional bias distributions are somewhat similarly centered for both RFRM and RFRM-PM, the skew is significantly lower for RFRM in both directions, revealing its higher performance.



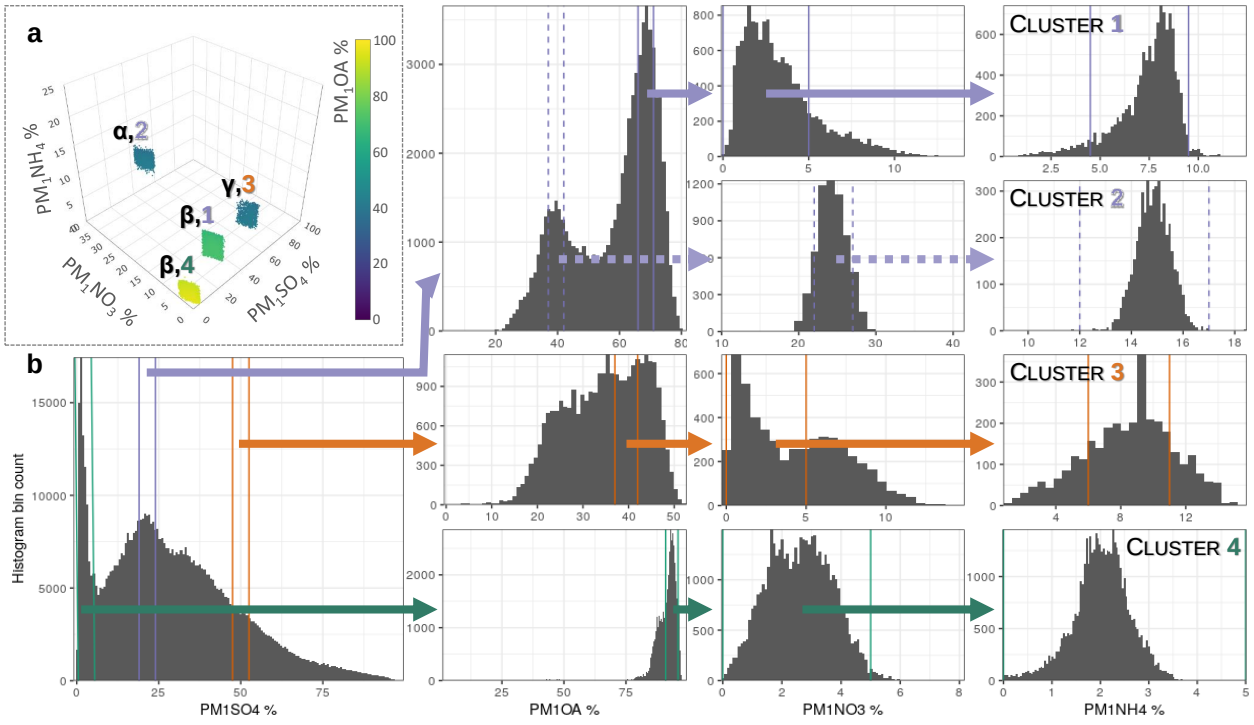
Supplementary Figure 10: **Model performance along the tropospheric vertical extent.** RFRM-derived (dashed orange curves; borderless violins) and RFRM-PM-derived (dashed purple; bordered violins) compared to airborne-measured (black) [CCN0.4] across the vertical tropospheric extent. Vertical curves are the generalized additive model fits with confidence interval shaded in grey. Violin plots show the distribution of fractional bias at each 1 km layer. Color bar (log scale) shows the number of observations in each violin. Also shown are the simple box plots, without the kernel density. Although average (generalized additive model fits) for RFRM-derived and RFRM-PM are both in good agreement with airborne-measurements and the fractional bias distributions are somewhat similarly centered for both RFRM and RFRM-PM, the skew is significantly lower for RFRM in both directions, revealing its higher performance.



Supplementary Figure 11: **Mean PNSD w.r.t. tropospheric height and by model performance.** Aerosol number size distribution across the vertical corresponding to degree of estimation of RFRM and RFRM-PM. Percentage within each class (Underestimation, Good-Agreement, or Overestimation) are noted to the left of each panel heading. Note the x and y axes and colorscale have different ranges for KORUS-AQ and ATom1–4 to capture features.



Supplementary Figure 12: ***k*-means clustering of aerosol composition.** Four-dimensional plot of the composition of PM_1 measured during all airborne campaigns. Shown on the axes are the percentage mass of each inorganic species and for organic aerosol on the colorscale. Axes for NO_3 and NH_4 are truncated to ≈99th percentile for a zoomed-in view. The plot is separated into the three optimum clusters determined by the k-means clustering algorithm. The clustering is also sound from the chemistry point-of-view— α: ↑ NH_4NO_3, β: ↑ OA, and γ: ↑ SO_4.



Supplementary Figure 13: **Selected aerosol composition clusters. a** The chosen clusters of maximum density and within a tolerance of ±2.5% for similarity of composition of each PM_1 component. **b** Schematic of how this choice of clusters was made by selecting these 5% ranges of individual PM_1 composition.

Supplementary Tables

Table 1: Details of airborne campaigns identified with RFRM's 9 predictors' measurements.

Campaign	Agency	Year	Months	Area	Data
ARCTAS	NASA	2008	Apr-Jul	Arctic	6
ATom-1	NASA	2016	Jul-Aug	Global	7–10
ATom-2	NASA	2017	Jan-Feb	Global	7–10
ATom-3	NASA	2017	Sep-Oct	Global	7–10
ATom-4	NASA	2018	Apr-May	Global	7–10
DC3	NSF/NCAR	2012	May-Jun	Mid-west and South US	11
DISCOVER-AQ^{TX}	NASA	2013	Sep	Texas	12
KORUS-AQ	NASA, NIER	2016	Apr-Jun	Korean Peninsula	13
SEAC^{4}RS	NASA	2013	Aug-Sep	Southeast US, Gulf of Mexico, and California	14
WE-CAN	NSF/NCAR	2018	Jul-Sep	Western US	15

Abbreviations— **ARCTAS**: Arctic Research of the Composition of the Troposphere from airborne and Satellites, **ATom**: Atmospheric Tomography Mission, **DC3**: Deep Convection Clouds & Chemistry experiment, **DISCOVER-AQ**: Deriving Information on Surface Conditions from Column and Vertically Resolved Observations Relevant to Air Quality, **KORUS-AQ**: Korea-United States Air Quality Study, **SEAC4RS**: Studies of Emissions and Atmospheric Composition, Clouds and Climate Coupling by Regional Surveys, **WE-CAN**: Western wildfire Experiment for Cloud chemistry, Aerosol absorption and Nitrogen. **NASA**: US National Aeronautics and Space Administration, **NSF**: US National Science Foundation, **NCAR**: US National Center for Atmospheric Research, **NIER**: South Korea's National Institute of Environmental Research.

Table 2: Details of instrumentation for the 1 Hz measurements of the 9 predictor variables of atmospheric state and composition and of [CCN0.4].

	Instrument	Accuracy (±)	Detection Limit	Precision
$\boldsymbol{T}$	Rosemount Inc. 102 E4AL	1.05 K	174–333 K	0.1 K
RH	Edgetech 3-Stage Hygrometer	2%	0–100%	0.10%
	Lyman-alpha^{w}	5%	0–100%	0.20%
$\mathbf{SO_2}$	CIMS	10–15%	10 pptv	NA
$\mathbf{NO}_x$	NOAA NOyO3*	5–7%	10 pptv	6–20 pptv
	NCAR NOxyO3†	4%	10 pptv	30 pptv
	NCAR NO-NO2^{w}	4%	10 pptv	30 pptv
$\mathbf{O_3}$	NOAA NOyO3*	2%	NA	15 pptv
	NCAR NOxyO3†	2–3%	NA	15–40 pptv
	NCAR NO-NO2^{w}	5%	0–300 ppbv	500 pptv
$\mathbf{SO_4}$	HR-ToF-AMS^{d}	36%	40.3 ng·m^{-3}	10–36%
$\mathbf{NH_4}$		34%	294 ng·m^{-3}	10–34%
$\mathbf{NO_3}$		34%	22.5 ng·m^{-3}	10–34%
OA		38%	170 ng·m^{-3}	10–38%
CCN	DMT CCN-100^{a}	10–20%	60000‡ cm^{-3}	10 cm^{-3}
CCNss		0.04	NA	0.04

Abbreviations— **CIMS**: Chemical Ionization Mass Spectrometer, **DMT CCN-100**: Droplet Measurement Technologies CCN counter (scanning flow analysis mode), **HR-ToF-AMS**: High Resolution Time-of-Flight Aerosol Mass Spectrometer; instrument quantification characteristics from 16,17 and accuracy ranges are $\pm 2\sigma$, **Lyman-alpha**: Lyman-alpha absorption hygrometer, **NCAR NO-NO2**: NCAR 2-Channel Chemiluminescence Instrument, **NCAR NOxyO3**: NCAR 4-Channel Chemiluminescence Instrument, **NOAA NOyO3**: National Oceanic and Atmospheric Administration 4-Channel Chemiluminescence Instrument, **CCNss**: instrument supersaturation during CCN measurement.
Superscripts— w for WE-CAN (2018). * for DC3, SEAC^{4}RS, and ATom1–4. † for ARCTAS, DISCOVER-AQ^{TX}, KORUS-AQ, and WE-CAN. d for DISCOVER-AQ^{TX}, aerosol speciation measurements were using a BMI Particle-Into-Liquid Sampler (PILS) with Ion Chromatograph (IC) for inorganics and with a Sievert Total Organic Carbon (TOC) Analyzer for organics. a for ATom1–4 CCN was not directly measured but inferred from the particle number size distributions measured by the Aerosol Microphysical Properties (AMP) package[5]. ‡ for upper limit of detection for the lowest sampling flow-rate of $0.\overline{3}$ cm^{3}s^{-1}.

Supplementary References

1. 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).

2. 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).

3. 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).

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

5. 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).

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

7. 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.

8. 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/.

9. 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.

10. 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.

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

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

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

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

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

16. Bahreini, R. *et al.* Organic aerosol formation in urban and industrial plumes near houston and dallas, texas. *Journal of Geophysical Research* **114** (2009).

17. DeCarlo, P. F. *et al.* Field-deployable, high-resolution, time-of-flight aerosol mass spectrometer. *Analytical Chemistry* **78**, 8281–8289 (2006).