Supplementary information for: "Tracing biological, human, and inorganic sources of coarse aerosols via single-particle fluorescence and optical morphology"

Aiden Jönsson[1,2], Jinglan Fu[1,2,3], Gabriel Pereira Freitas[1,2], Ian Crawford[4], Pavla Dagsson-Waldhauserová[5,6], Radovan Krejci[1,2], Yutaka Tobo[7,8], Karl Espen Yttri[9], and Paul Zieger[1,2]

[1]Department of Environmental Science, Stockholm University, Stockholm, Sweden
[2]Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden
[3]Now at: Centre for Isotope Research, Faculty of Science and Engineering, University of Groningen, Groningen, the Netherlands
[4]Department of Earth and Environmental Science, University of Manchester, Manchester, United Kingdom
[5]Agricultural University of Iceland, Hvanneyri, Iceland
[6]Czech University of Life Sciences, Prague, Czech Republic
[7]National Institute of Polar Research, Tackikawa, Tokyo, Japan
[8]Graduate Institute for Advanced Studies SOKENDAI, Tachikawa, Tokyo, Japan
[9]The Climate and Environmental Research Institute NILU, Kjeller, Norway

Correspondence: A. Jönsson (aiden.jonsson@aces.su.se) and P. Zieger (paul.zieger@aces.su.se)

Contents

List of Tables

List of Figures

S1 Domain adaptation

Following the initial untuned model assessment, we perform the domain adaptation steps for the LRMs by defining functions expressing the relative influence between two tracer types; for the pollution model, this is between biological and pollution/-combustion influence, and for the dust model, this is between sea spray aerosol (SSA) and dust influence. These are calculated for each tracer sampling period and are then appended as "soft labels" I to all particles observed with the MBS during that time period. These soft label metrics are encoded into "hard labels" $L(I)$, or best-guess flags used as the training labels for the tuning step, using the rounding function:

$$L(I) = \begin{cases} 1 & \text{if } I \geq 0.5, \\ 0 & \text{else.} \end{cases} \tag{1}$$

The tuning block samples are then used to further train the LRM using these hard labels and a weight proportional to the confidence in the hard label based on I_{pol}. We estimate the confidence in the hard label using the Shannon information entropy H for N number of components given by the equation:

$$H = -\sum_{i}^{N} p_i \log_2 p_i, \tag{2}$$

where p_i is the probability of influence by individual components. Because Shannon information entropy is the inverse of certainty in that label, we estimate the confidence in the hard label C with $1 - H$. For our binary classification problems where the probability of influence by a given component is estimated with I, this yields the formula:

$$C(I) = 1 - \Big(I \log_2(I) - (1 - I) \log_2(1 - I) \Big). \tag{3}$$

This confidence metric is then used to derive tuning block trainign weights for each particle. We choose a maximum amount of influence from the tuning block β to be 20% of that of the source training data ($\beta = 0.2$). The individual tuning block sample weights must sum to this proportion to ensure this level of influence ($\sum w_{tune} = \beta$). Since the number of particles in the tuning and training blocks are different, the weights are scaled by the proportion of tuning data points by the relative size of tuning block data points to lab training data points ($\frac{N_{tune}}{N_{lab}}$). The weights are finally scaled by the mean confidence in all particles $\bar{C}$, which is necessary for summing to β. The tuning block weights for each particle appended with I are thus calculated as:

$$w_{tune}(I) = \beta \frac{N_{tune}}{N_{lab}} \frac{C(I)}{\bar{C}}. \tag{4}$$

We calculate the pollution soft label using eBC and arabitol concentrations; because a number of biological tracers correlate with biomass burning tracers (Figure S7), we choose arabitol for having the weakest correlation among them. We define the pollution influence metric I_{pol}, a type of soft label, as the proportion of normalized (by maximum value over the entire period) eBC (eBC$_{norm}$) to the sum of eBC$_{norm}$ and normalized (also by maximum value) biological tracer arabitol concentrations $C_{ara,norm}$:

$$I_{pol} = \frac{\text{eBC}_{norm}}{\text{eBC}_{norm} + C_{ara,norm}}. \tag{5}$$

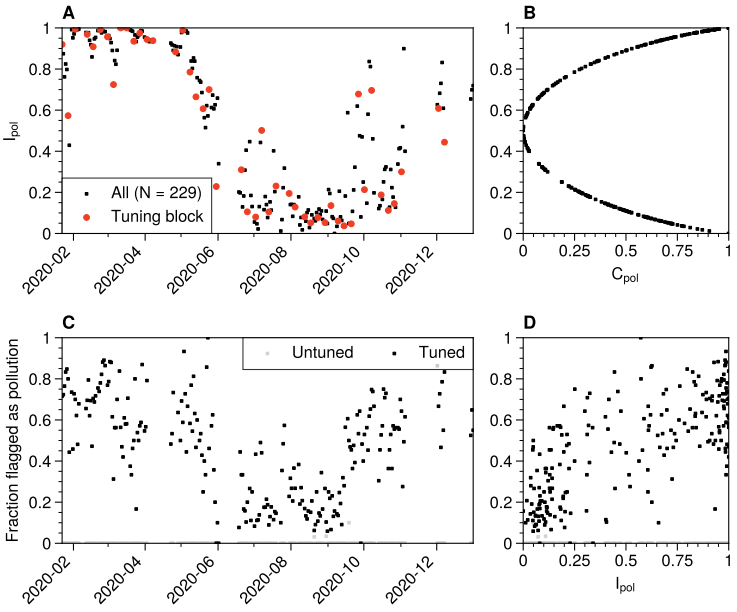

Figure S1. Domain adaptation parameters for the tuning of the pollution logistic regression model (LRM). The time series of the pollution soft label I_{pol} is shown in (a); the distribution of confidence estimates C_{pol} according to I_{pol}, calculated with equation (3), is shown in (b). (c): Time series of the fraction of highly fluorescent particles (HFPs) flagged as pollution by the pollution LRM. (d): Fractions of HFPs flagged as pollution plotted against I_{pol}.

Because arabitol measurements were taken weekly and eBC measurements are daily values, we use the collocated weekly arabitol concentration for all daily eBC concentrations during that week. For the tuning block, we choose every fifth daily sampling period in the time series and select all HFPs during those periods; the resulting tuning block includes 4,012 particles. Figure S2 illustrates the evolution of I_{pol} throughout the measurement period, the distribution of C_{pol}, and the fractions of HFPs identified as pollution using. These time periods cover a distribution of high and very low influences by pollution and a range in between. The resulting metric has the expected behavior; the time series (Fig. S1a) shows strong influence by pollution during the winter ($I_{pol} \sim 1$) and low influence during the summer ($I_{pol} \sim 0$). The tuned model reproduces the annual cycle of I_{pol} (Fig. S1c) and responds roughly linearly to I_{pol} (Fig. S1d).

For the dust soft label, we use mineral dust mass approximated using silicon, aluminum, iron, manganese, titanium, nickel, and chromium concentrations, derived with inductively coupled plasma mass spectroscopy, from Aas (2024). The total mineral dust mass concentration [MD] was approximated with the assumption that all measured elements were present as oxides:

$$[\mathrm{MD}] \approx [\mathrm{SiO_2}] + [\mathrm{Al_2O_3}] + [\mathrm{Fe_2O_3}] + [\mathrm{MnO}] + [\mathrm{TiO_2}] + [\mathrm{NiO}] + [\mathrm{Cr_2O_4}]. \quad (6)$$

In eq. (6), silicon concentrations were calculated using an assumed Si/Al ratio of 2.05, derived from Si and Al measurements in aerosol filter samples collected at Zeppelin Observatory in 2020:

$$[\mathrm{Si}] = 2.05 \times [\mathrm{Al}]. \tag{7}$$

Because silicon measurements by (Calzolai, 2025) were derived using a different technique, this was done in order to maintain consistency with measurements for the other mineral constituents. For iron oxide, an equal partitioning between Fe_2O_3 and FeO was assumed. For SSA tracers, the sum concentration of the three major constituents sodium, chloride, and magnesium, comprising the vast majority of supermicron SSA mass fraction (Gamage et al., 2025), were chosen to approximate SSA mass ([SSA]):

$$[\mathrm{SSA}] \approx [\mathrm{Na}^+] + [\mathrm{Cl}^-] + [\mathrm{Mg}^{2+}]. \tag{8}$$

The dust soft label I_{dust} is then calculated as:

$$I_{dust} = \frac{[\mathrm{MD}]}{[\mathrm{MD}] + [\mathrm{SSA}]}. \tag{9}$$

The resulting dust soft label estimates the influence of dust mass at Zeppelin Observatory and varies considerably throughout the year, including events with high dust influence (Figure S2a). Note that using the proportion of mass estimates necessarily comes with the assumption that SSA and dust have approximately the same mass distributions across particle sizes, although typical dust and SSA size distributions can significantly differ. Sample periods for the tuning block were taken as the extremes in I_{dust} (the 0-5th percentile and 95-100th percentile periods); due to high numbers of data points, random subsets were taken for an equal number of laboratory and field data points. Weights were calculated as in equation (4). Figure S2 shows the time series of I_{dust} and the corresponding C_{dust} curve, along with fractions of particles identified as dust computed with both tuned and untuned models. The tuned model displays greater sensitivity, with the dust fraction increasing linearly with I_{dust}. Dust fractions are overall lower and more variable in the tuned model.

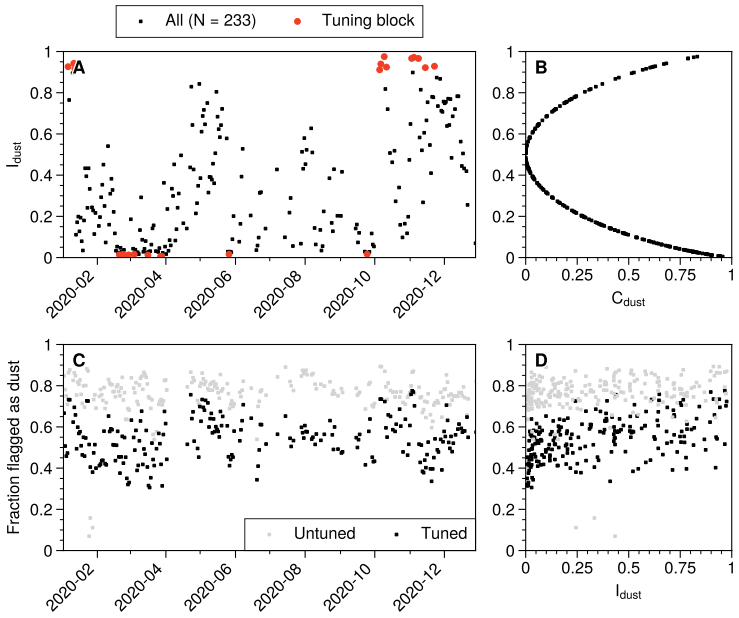

Figure S2. Domain adaptation parameters for the tuning of the dust logistic regression model (LRM). The time series of the dust soft label I_{dust} is shown in (a); the distribution of confidence estimates C_{dust} according to I_{dust}, calculated with equation (3), is shown in (b). (c): Time series of the fraction of particles flagged as dust by the dust LRM. (d): Fractions of particles flagged as dust plotted against I_{dust}.

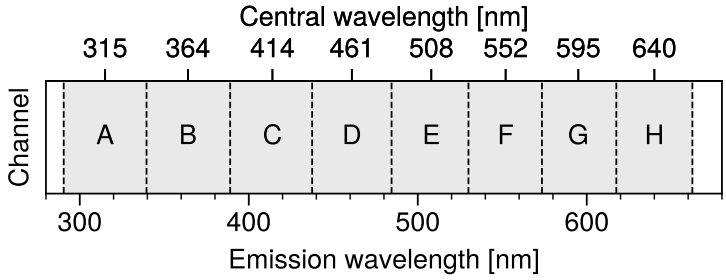

Figure S3. Fluorescence detection channels in the MBS; the detectable region is shaded in gray, band (acceptance interval) edges are shown with dashed lines, and each channel's central wavelengths are given on the top axis.

Pollen type	Collection and processing	Reported size range (microscopy/acetic blue measurement)
Ash	Dry/mechanical	18-30/21-31 μm
Alder	Dry/mechanical	20-30/18-30 μm
Birch	Dry/mechanical	20-30/17-28 μm
Hazel	Dry/mechanical	20-30/22-32 μm
Juniper	Dry/mechanical	20-35/15-33 μm
Pine	Dry/mechanical	50-100/31-77 μm
Willow	Acetone defatting/mechanical	18-25/17-29 μm

Table S1. Pollen samples, their processing, and size ranges reported by Pharmallerga CZ s.r.o., our pollen sample provider.

Parameter	Description [unit, if applicable]	Used as input in:		
		Pollution LRM	UMAP	Dust LRM
Size	Optical diameter [μm]	X	X	–
XE1_1_norm	A channel fluorescence intensity (normalized)	X	X	–
XE1_2_norm	B channel fluorescence intensity (normalized)	X	X	–
XE1_3_norm	C channel fluorescence intensity (normalized)	X	X	–
XE1_4_norm	D channel fluorescence intensity (normalized)	X	X	–
XE1_5_norm	E channel fluorescence intensity (normalized)	X	X	–
XE1_6_norm	F channel fluorescence intensity (normalized)	X	X	–
XE1_7_norm	G channel fluorescence intensity (normalized)	X	X	–
XE1_8_norm	H channel fluorescence intensity (normalized)	X	X	–
FL_ratio	Ratio of the sums of the first two to last six fluorescence channels	X	X	–
FL	Summed fluorescence intensity of all channels [au]	X	X	–
Sat	Flag for saturation in any fluorescence channel (boolean)	X	X	–
AsymLR%	Left-right mirror asymmetry across corresponding pixels [%]	X	X	X
AsymLRinv%	Left-right mirror asymmetry, inverted [%]	X	X	X
PeakMeanL, PeakMeanR	Peak-to-mean-ratio, left and right arrays	X	X	X
PeakWidthL, PeakWidthR	Peak width at half height, left and right arrays	X	X	X
PeakCountL, PeakCountR	Number of peaks over threshold, left and right arrays	X	X	X
KurtosisL, KurtosisR	Kurtosis of signal treated as a distribution, left and right arrays	X	X	X
VarianceL, VarianceR	Variance of signal treated as a distribution, left and right arrays	X	X	X
MeanL, MeanR	Mean signal intensity, left and right arrays	X	X	X
SumL, SumR	Sum of signal intensities, left and right arrays	X	X	X
SkewL, SkewR	Skew of signal treated as a distribution, left and right arrays	X	X	X
MirrorL%, MirrorR%	Mirror symmetry across middle pixel, left and right arrays [%]	X	X	X

Table S2. Output variables from the Multiparameter Bioaerosol Spectrometer (MBS). Normalized fluorescence channel intensities are divided by the maximum intensity detected across the spectrum for each measurement. The right three columns denote (with an "X") where the variable is used as input between the pollution logistic regression model (LRM), uniform manifold approximation and projection (UMAP) transformer, and dust LRM in our classification algorithm.

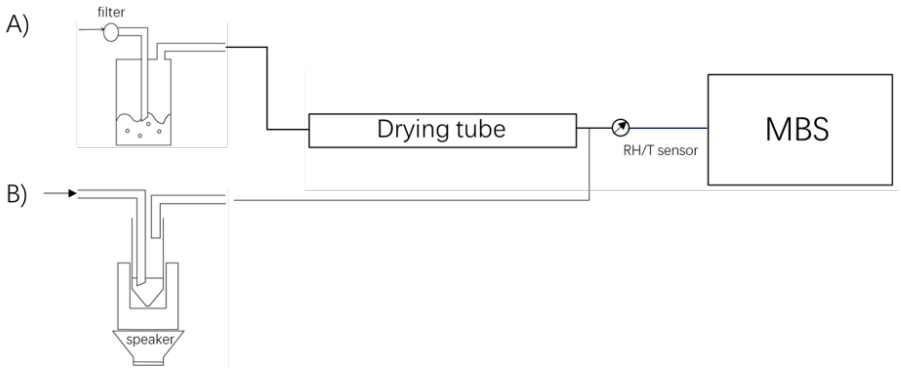

Figure S4. The laboratory setup for the source experiments conducted in this study, with (a) the dry air line where particles were aerosolized with a speaker in dry experiments, and (b) the particle nebulizer for wet experiments.

Tracer	Sampling frequency	Inlet type	Citation
Silicon	Daily	Whole air	Calzolai (2025)
Iron	Weekly	PM10	Aas (2024)
Aluminum	Weekly	PM10	Aas (2024)
Manganese	Weekly	PM10	Aas (2024)
Titanium	Weekly	PM10	Aas (2024)
Nickel	Weekly	PM10	Aas (2024)
Chromium	Weekly	PM10	Aas (2024)
Calcium	Daily	Whole air	Aas (2025)
Potassium	Daily	Whole air	Aas (2025)
Organic carbon	Weekly	PM10	Yttri (2023b)
Fructose	Weekly	PM10	Yttri (2023a)
Glucose	Weekly	PM10	Yttri (2023a)
Arabitol	Weekly	PM10	Yttri (2023a)
Mannitol	Weekly	PM10	Yttri (2023a)
Sodium	Daily	Whole air	Aas (2025)
Chloride	Daily	Whole air	Aas (2025)
Magnesium	Daily	Whole air	Aas (2025)
Levoglucosan	Weekly	PM10	Yttri (2023a)
Equivalent black carbon (eBC)	Minutely (resampled to daily)	Whole air	Freitas et al. (2023b)
Lead	Weekly	PM10	Aas (2024)

Table S3. Tracers used in this study, their sampling frequencies, the sampling inlet types used (PM10 denotes particulate matter under 10 μm), and their references.

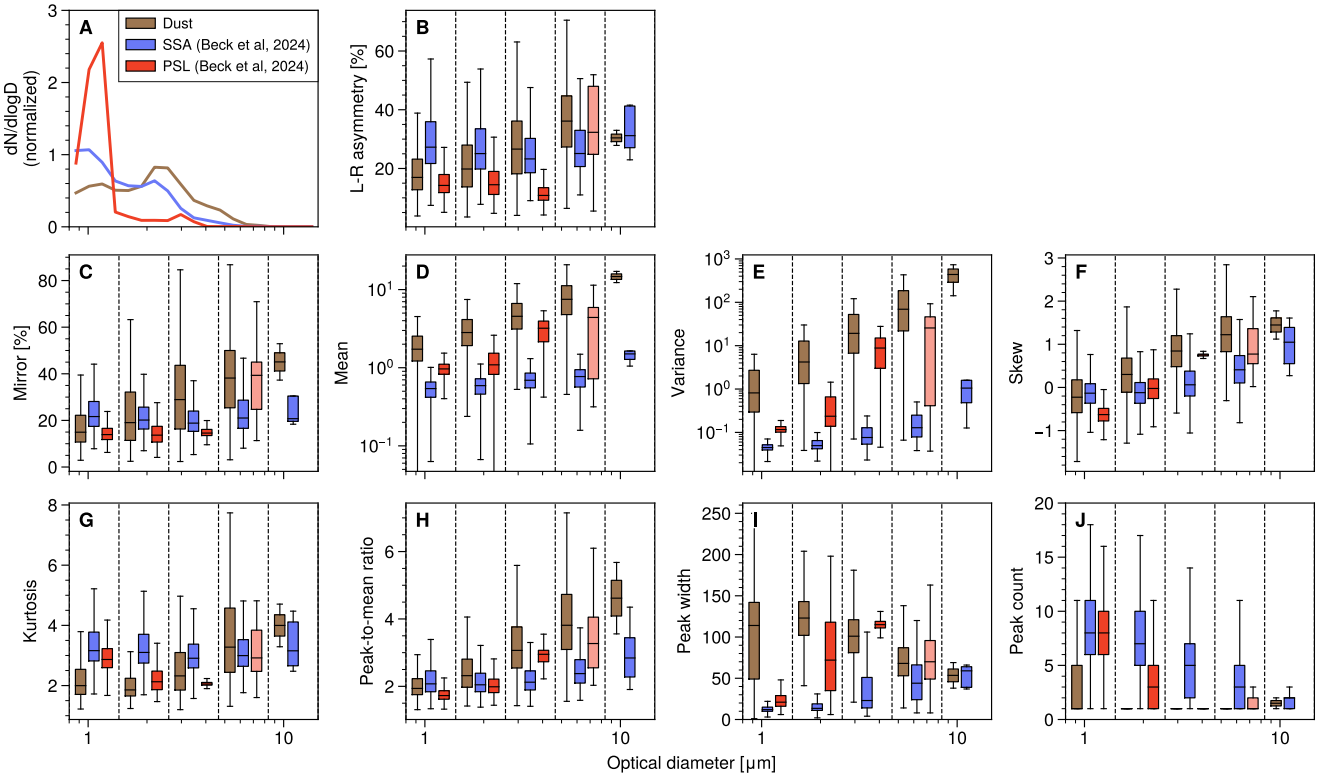

Figure S5. Optical morphology parameters, resolved by size, for all dust and sea spray aerosol (SSA) compared in this study, as well as the polystyrene latex (PSL) spheres from Beck et al. (2024). (a): Size distributions for all particles measured in these experiments. (b-j): Distributions for optical morphology parameters, binned by size; for PSLs, the larger size bin (lightly filled) is likely due to particle aggregation and is thus not representative of spherical morphology. (c) Left-right (L-R) asymmetry is based on both optical arrays. For single-array metrics (c-j), only the right array is used.

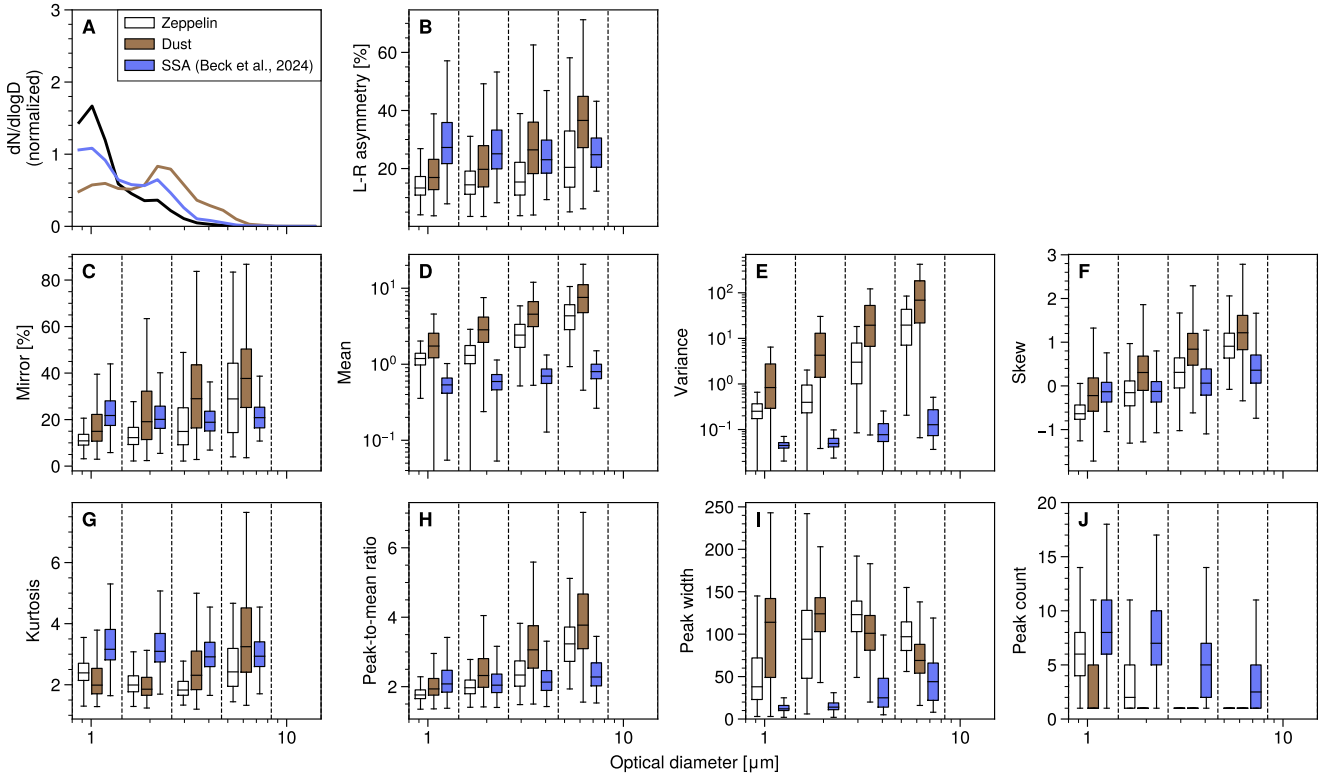

Figure S6. Optical morphology parameters, resolved by size, for all dust and sea spray aerosol (SSA) compared in this study, and for a particles observed on 2 March 2020, a day during the sampling period with least dust influence (selected by the wintertime minimum in dust tracer iron concentration measurements, with each sample taken over three days). (a): Size distributions for all particles measured in these experiments. (b-j): Distributions for optical morphology parameters, binned by size; for PSLs, the larger size bin (lightly filled) is likely due to particle aggregation and is thus not representative of spherical morphology. (c) Left-right (L-R) asymmetry is based on both optical arrays. For single-array metrics (c-j), only the right array is used.

Classification scheme algorithm	
Select all particles $\geq$0.8 µm.	
Fluorescent particle branch	**Non-fluorescence particle branch**
1. Select all highly fluorescent particles (HFP; $\geq 9\sigma$ fluorescence).	1. Select all non- or weakly fluorescent particles ($< 9\sigma$ fluorescence).
2. Feed the fluorescence and optical scattering variables to the pollution logistic regression model (LRM); particles with a probability of being pollution greater than a chosen confidence threshold are classed as interferents/pollution.	2. Select only particles that are of sufficient size for detecting optical property signals (≥ 2.5 µm).
3. Reduce the dimensionality of the input variables (fluorescence and morphology properties) to two dimensions using the uniform manifold approximation and projection (UMAP) transformer.	3. Feed the optical scattering properties as input to the dust LRM and attach the dust-likeness probability to each particle.
4. Feed the two-dimensional UMAP coordinates as input to the k-nearest neighbors (kNN) classifier trained on source characterization data and attach a most likely class identification, along with its probability estimate, to each particle.	4. Filter particles with a dust-likeness probability higher than a desired confidence threshold as dust; all others are considered SSA.
5. Filter the kNN-classified output by a desired confidence threshold. Particles positively identified as one of the three bioaerosol classes (pollen, bacteria, and fungal spores) according to the threshold are considered fluorescent primary biological aerosol particles (fPBAPs); all others are considered interferents/pollution.	

Table S4. A written form of our particle classification algorithm illustrated in Fig. 5.

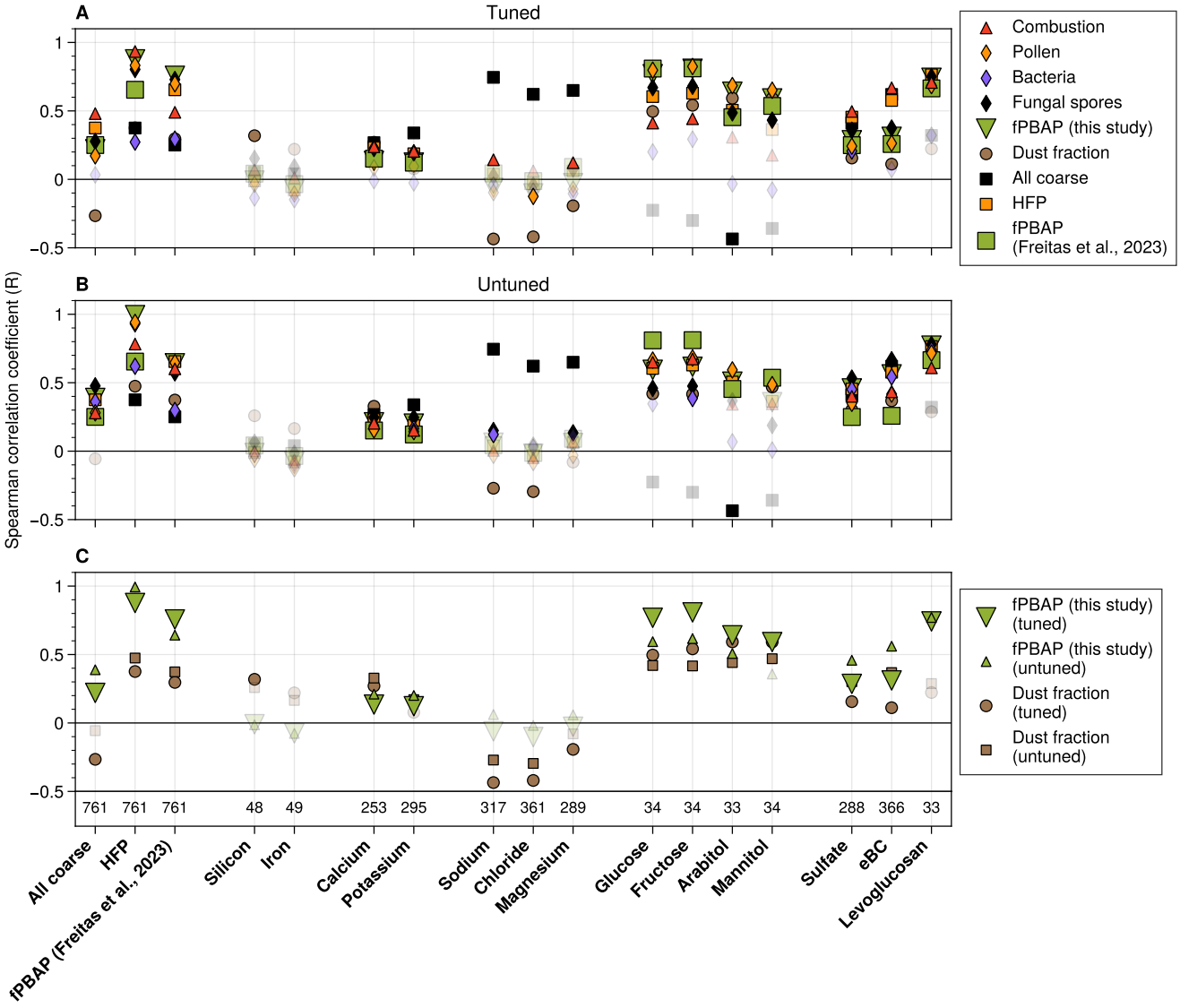

Figure S7. Spearman correlation coefficients between particle concentration estimates for all coarse particles, highly fluorescent particles (HFPs), fluorescent biological aerosol particles (fPBAP) according to the methods of Freitas et al. (2023a) and our classification algorithm (this study), individual fPBAP classes from our classifier, and chemical tracers at Zeppelin Observatory during 2020, for both the tuned (a) and untuned (b) classification algorithms. A comparison of the tuned and untuned main classification algorithm components are shown in (c). Correlations that are not significant with $p < 0.05$ are made transparent. Tracers used in the domain adaptation steps are marked with an asterisk. The panel on the bottom includes the sample size for each tracer.

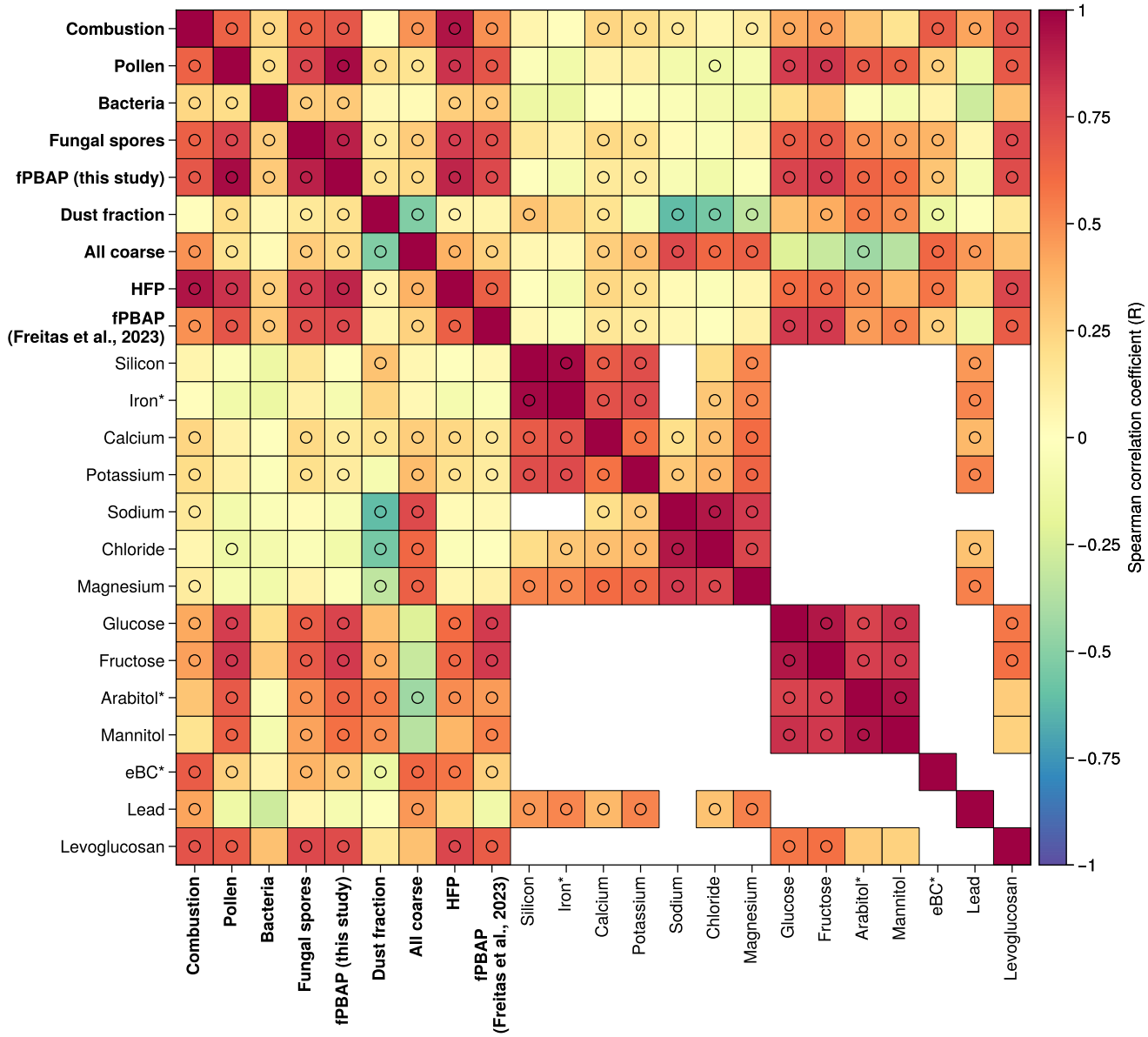

Figure S8. Spearman correlation matrix between particle concentration estimates and chemical tracers at Zeppelin Observatory. *All coarse* includes highly fluorescent particles (HFPs) and fluorescent primary biological aerosol particles (fPBAPs). HFP and fPBAP (Freitas et al., 2023a) are decision tree-based definitions according to fluorescence signals using the methods of Freitas et al. (2023a), and fPBAP (this study) is based on the classification algorithm presented in this study. Circles denote where the Spearman correlation coefficient is significant with $p < 0.05$. Empty boxes denote where collocated data are not present. Tracers used in the domain adaptation steps are marked with an asterisk.

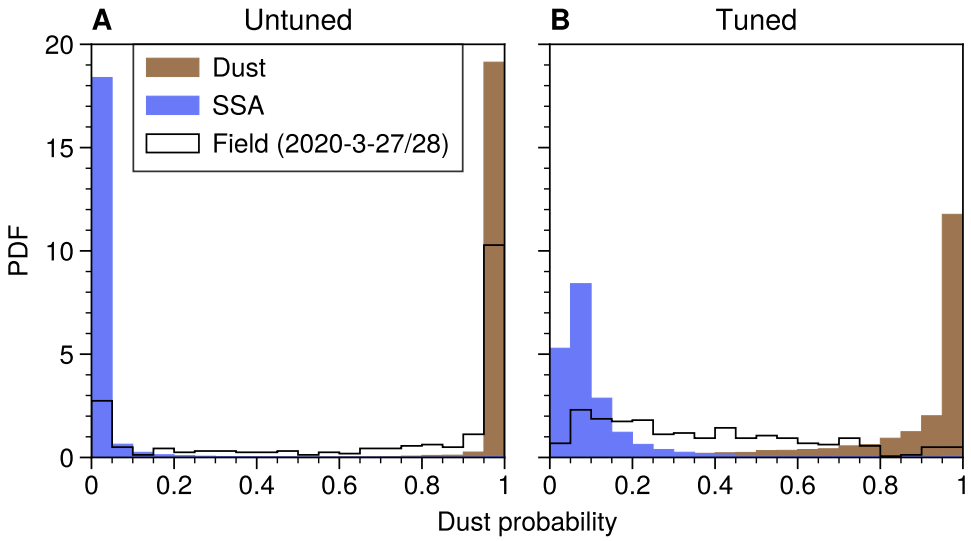

Figure S9. Probability density functions (PDFs) of logistic regression model (LRM) output for probability of being classified as dust for laboratory-measured dust (shown is the Svalbard sample from Tobo et al. 2019, characterized in this study) and sea spray aerosol (SSA; shown is the artificial sea salt sample from Beck et al. 2024), as well as non-fluorescent coarse particles observed at Zeppelin Observatory during an example period with minimal dust influence (27-28 March 2020). Results are shown for both the untuned (a) and tuned (b) LRM.

References

Aas, W.: CAMP, NILU, EMEP, AMAP, 2017-2025, Heavy metals at Zeppelin mountain (Ny-Ålesund), data hosted by EBAS at NILU, https://doi.org/10.48597/6PXH-3ESD, accessed 2025-8-11, 2024.

Aas, W.: EMEP, CAMP, NILU, AMAP, 2001-2025, Inorganics in air and particle phase at Zeppelin mountain (Ny-Ålesund), data hosted by EBAS at NILU, https://doi.org/10.48597/H7PV-77C5, accessed 2025-6-16, 2025.

Beck, I., Moallemi, A., Heutte, B., Pernov, J. B., Bergner, N., Rolo, M., Quéléver, L. L. J., Laurila, T., Boyer, M., Jokinen, T., Angot, H., Hoppe, C. J. M., Müller, O., Creamean, J., Frey, M. M., Freitas, G., Zinke, J., Salter, M., Zieger, P., Mirrielees, J. A., Kempf, H. E., Ault, A. P., Pratt, K. A., Gysel-Beer, M., Henning, S., Tatzelt, C., and Schmale, J.: Characteristics and sources of fluorescent aerosols in the central Arctic Ocean, Elementa: Science of the Anthropocene, 12, 00 125, https://doi.org/10.1525/elementa.2023.00125, 2024.

Calzolai, G.: NILU, 2017-2020, Heavy metals, inorganics in air and particle phase, silicon and sulphur at Zeppelin mountain (Ny-Ålesund), data hosted by EBAS at NILU, https://doi.org/10.48597/8SEU-C5UP, accessed 2025-8-12, 2025.

Freitas, G. P., Adachi, K., Conen, F., Heslin-Rees, D., Krejci, R., Tobo, Y., Yttri, K. E., and Zieger, P.: Regionally sourced bioaerosols drive high-temperature ice nucleating particles in the Arctic, Nature Communications, 14, 5997, https://doi.org/10.1038/s41467-023-41696-7, 2023a.

Freitas, G. P., Conen, F., Adachi, K., Yttri, K. E., Krejci, R., and Zieger, P.: Concentration of bioaerosols and ice nucleating particles at Zeppelin Observatory, Svalbard, 2017-2020, Dataset version 1. Bolin Centre Database., https://doi.org/10.17043/zeppelin-freitas-2023-bioaerosols-1, accessed: 2025-07-09, 2023b.

Gamage, D. K., Hasenecz, E., Dorcé, G. P., Mayer, K. J., Sauer, J. S., Lee, C., Prather, K. A., and Stone, E. A.: Molecular Characterization of Nascent and Aged Sea Spray Aerosol, ACS Earth and Space Chemistry, 9, 1453–1464, https://doi.org/10.1021/acsearthspacechem.4c00412, 2025.

Tobo, Y., Adachi, K., DeMott, P. J., Hill, T. C. J., Hamilton, D. S., Mahowald, N. M., Nagatsuka, N., Ohata, S., Uetake, J., Kondo, Y., and Koike, M.: Glacially sourced dust as a potentially significant source of ice nucleating particles, Nature Geoscience, 12, 253–258, https://doi.org/10.1038/s41561-019-0314-x, 2019.

Yttri, K. E.: ACTRIS, GAW-WDCA, EMEP, 2019-2021; Organic tracers at Zeppelin mountain (Ny-Ålesund), data hosted by EBAS at NILU, https://doi.org/10.48597/VYK3-XQEN, accessed 2025-6-16, 2023a.

Yttri, K. E.: ACTRIS, GAW-WDCA, EMEP, 2020-2021; OC/EC at Zeppelin mountain (Ny-Ålesund), data hosted by EBAS at NILU, https://doi.org/10.48597/9WYQ-6KVF, accessed 2025-6-16, 2023b.