Supplement of Atmos. Chem. Phys., 21, 9223–9251, 2021
https://doi.org/10.5194/acp-21-9223-2021-supplement

Atmospheric Chemistry and Physics
Open Access
EGU

Supplement of

Towards understanding the characteristics of new particle formation in the Eastern Mediterranean

Rima Baalbaki et al.

Correspondence to: Rima Baalbaki (rima.baalbaki@helsinki.fi)

S1. Data availability

Table S1. Availability of hourly data (%) from the three particle measuring instruments.

Month	PSM	NAIS	SMPS
January	72.8	93.4	16.1
February	96.4	94.5	94.6
March	83.6	96.4	75.1
April	83.3	100.0	99.7
May	67.6	99.7	91.5
June	43.5	100.0	55.7
July	41.5	100.0	81.0
August	77.4	100.0	96.1
September	93.5	99.9	98.5
October	90.3	100.0	96.8
November	80.0	99.9	1.7
December	100.0	100.0	0.0

S2. PSM setup, operation and data handling

S2.1. PSM core sampling inlet

The PSM inlet design was first introduced by Kangasluoma et al. (2016). It is a simple design encompassing a 6-mm tube fitted inside a 10-mm tube using a Swagelok T piece (Figure S1). In normal operating conditions, the 3rd outlet of the T-piece is connected to vacuum, which enables drawing higher flow through the 10-mm tube than the PSM flow, allowing the PSM to sample from the middle of this flow and thus minimizing losses caused by diffusion to the inlet walls (Figure S1a). During the background measurements, the 3rd outlet is connected to particle-free pressurized air with a high enough flow rate allowing the PSM to sample this particle free air (Figure S1b)

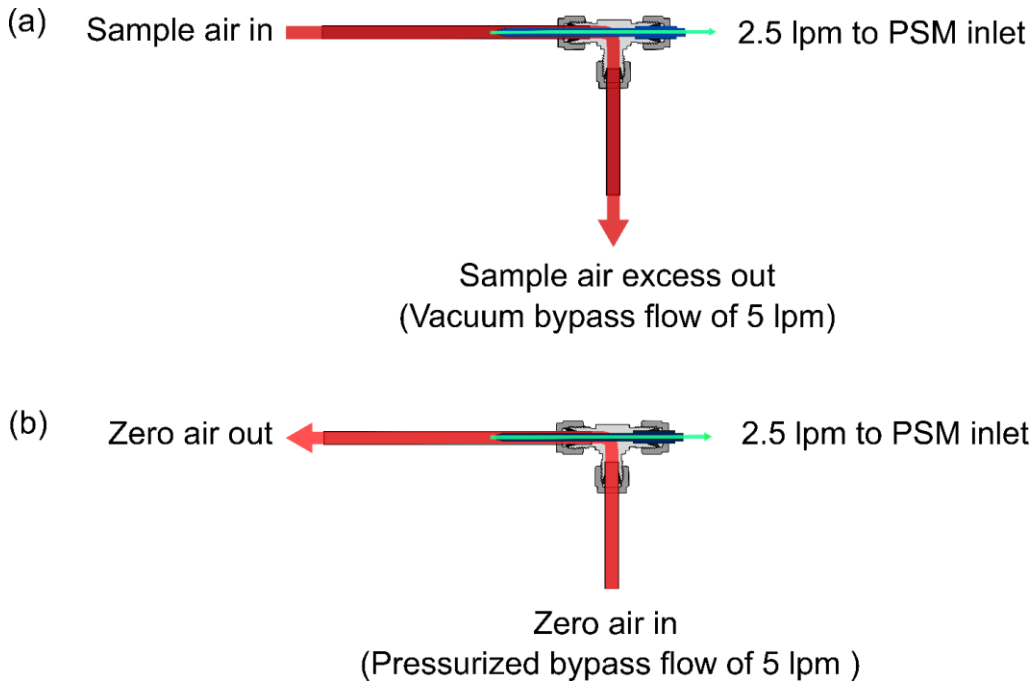

Figure S1. A schematic of the PSM core sampling inlet during normal operation (a) and during background measurements (b).

S2.2. PSM diluter

We used a prototype diluter designed at the University of Helsinki and later commercialized by Airmodus under the name "Airmodus nanoparticle diluter" (AND). The diluter has a cylindrical shape made of three modules. From the air-sampling side, the first module serves as a switchable ion filter that removes charged ions and particles up to a certain size and allows the measurement of neutral particles only. In this study, the ion filter was turned off. The second module is a core sampling piece radially connected to a vacuum source which draws 5 lpm excess flow from the sampling air. The third module constitutes the dilution module, where clean dry air is introduced radially into the sampled air flow. The differential pressure across the dilution unit is continuously monitored and is kept constant by a feedback mechanism to a PID controlled proportional valve which determines the dilution flow required to keep the dilution ratio constant. The design of the diluter was made as compact as possible to reduce losses and optimize penetration efficiency. Additionally, the dilution flow was monitored with a TSI flow meter and was used along with the pressure measurements to determine and correct for the real-time dilution factor.

The addition of the diluter has three different effects on the PSM measurements. The first effect is related to the penetration efficiency or line losses inside the diluter piece. The diluter's penetration efficiency was characterized in the laboratory and was found to be similar to that of the 6-cm-long core-sampling piece, which was used earlier in the study, so this effect is negligible. The second effect is related to possibly decreasing the water content in the sampled aerosols and thus making them smaller. However, we cannot correct for this effect because the hygroscopicity/dehydration of sub-3nm particles is not known. The third effect is related to the activation efficiency of particles at lower sample RH. The increased water content of the sample enhances both the DEG-water activation and DEG-aerosol-water activation. Since background zero measurements for the PSM were performed three times a day with filtered sample air, the effect of adding the diluter on the DEG-water interaction was indirectly monitored. The DEG-water counts were reduced after the addition of the diluter, as can be seen from Figure S2, but they were mostly in the range of 0 to 10, which is within normal operating conditions of the PSM. Concerning the DEG-aerosol-water activation, the uncertainty due to changing RH ranges between 0-0.3 nm on the PSM cutoff, and is smaller than the uncertainty due to change in particle composition (0-1 nm) (Kangasluoma et al., 2013), which also cannot be controlled.

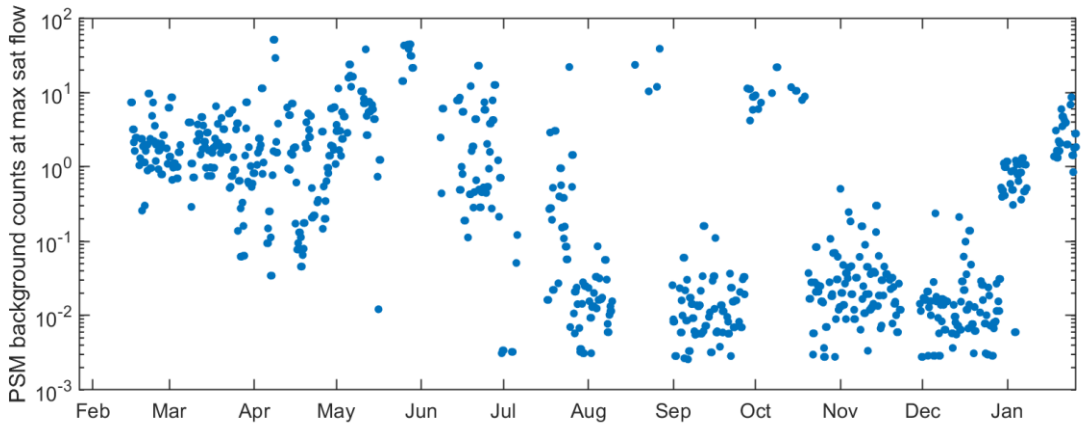

Figure S2. PSM counts at maximum saturator flow during zero measurements with filtered sampled air.

S2.3. nCNC (PSM+CPC) inversion

In principle, the PSM is a mixing-type condensation particle counter but without the measuring optics. It uses diethylene glycol (DEG) to grow nano-sized particles (~1-3 nm) up to around 90 nm. Subsequently, these particles enter the CPC and are further grown with butanol to sizes measurable by the CPC optical detector. In the first stage, the mixing ratio of DEG vapour with sample flow is scanned by continuously incrementing then

decrementing the saturator flow between 0.1 and 1.3 liters per minute (lpm) while keeping the sample flow constant. By varying the mixing ratio, the particle cut-off size is changed (i.e., at a higher mixing ratio, smaller particles are activated and grown; thus lower cut-off is achieved). Therefore, the nCNC measures the total particle concentration above a specific diameter, and inversion algorithms are required to retrieve the size distribution below 3 nm. The two most popular methods to invert PSM data are the kernel function method and the step inversion method. The expectation-maximization (EM) method has been recently recommended over the kernel method because it is less sensitive to random errors (Cai et al., 2018;Chan et al., 2020). Here, we compare the kernel method and the EM method using PSM data from the whole measurement period. Data pretreatment before inversion was done similarly for the two methods and included a:

1) Diagnostic check that identifies and removes erroneous data based on instrument diagnostics and flags.
2) Background subtraction: the instrumental background of the PSM was continuously monitored with daily automated random background (zero) checks. The background was subtracted from the measured data except when the background was very high (> 10% of the measured concentrations). Then, the corresponding data was deemed unusable until the background decreased to normal levels.
3) Correction for the time delay between PSM and CPC, which is typically ~5 seconds.
4) Noise filtering procedure achieved by applying a 6^{th} order median filter on the one second resolution data.
5) Quality check using the method suggested by Chan et al. (2020).
6) Minimization of the inversion matrix using a saturator flow inversion window of 0.08 lpm which minimized the saturator flow (corresponding to cut-off diameter) scans from ~120 to 16 per one-direction of the scan.
7) While pre-averaging before the inversion step is recommended for noisy data, here, we did not pre-average in order to capture the fast variations in the data.
8) The minimized cut-offs matrix is differentiated to retrieve the concentration in each size bin which is the input for the kernel inversion method. This step is unnecessary for the EM method, which takes the cut-off matrix as input (the varying total particle concentration at each saturator flow rate). Further explanation about the theoretical approach of each inversion method can be found in Cai et al. (2018).

During the inversion step, four kernels corresponding to four size channels (dp), with the following diameters: 1.1 nm, 1.3nm, 1.5 nm, and 2.4 nm were used with the kernel inversion method, whereas 50 kernels between 1.1 nm and 2.4 nm were used for the EM inversion method. The kernels are Gaussian-shaped and represent the derivative of the laboratory-derived detection efficiency curves with respect to the saturator flow rate. The median (μ) of the kernel function at each dp is equal to the saturator flow having half maximum detection efficiency at this diameter, whereas the width i.e. standard deviation (σ) is equal to $p_1/(d_p+q_1)$ where p_1 and q_1 are fitting parameters derived from the calibration curve. An example of PSM calibration curve data is shown in Figure 1 from Cai et al. (2018). Note that the actual input to the EM method is the detection efficiency curves rather than the kernels.

After the inversion step, inverted data was transformed from dN/dd_P to $dN/d\log d_P$ and averaged to longer times: five minutes and one hour. The comparison of the inversion methods was made by comparing the total $dN/d\log d_P$ concentration from the kernel and EM methods to each other. The two methods were reasonably comparable using the one hour resolution data (Figure S3). However, there is some scatter at low total concentrations, and the 5 min average data sometimes revealed considerable deviations. In this manuscript, we mainly use 1 hour resolution data for the presented analysis thus, we chose to use the data from the kernel inversion method because it gave better uniformity for the particle size distribution below 3 nm.

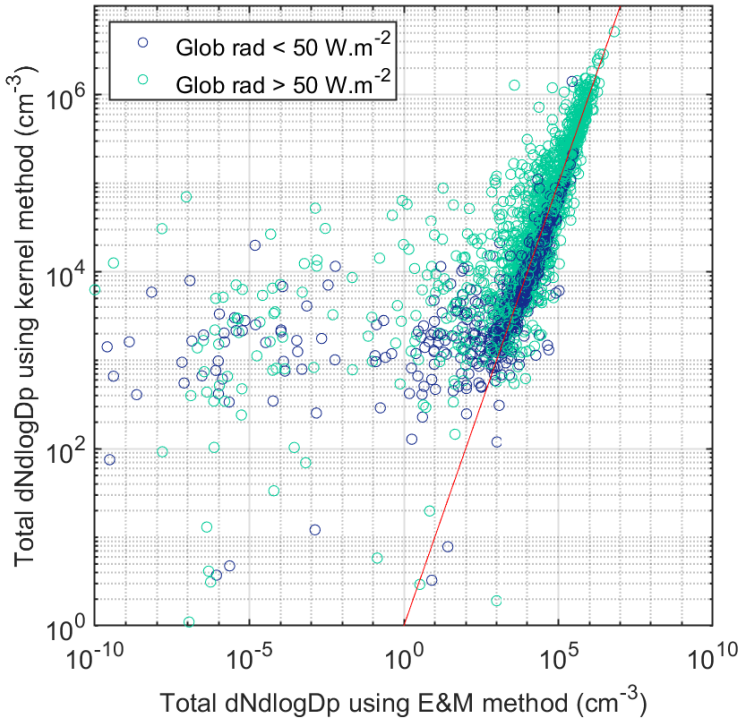

Figure S3. Comparison between total dN/dlogDp concentrations (cm^{-3}) between 1.1 and 2.4 nm computed from PSM data using the Kernel inversion method and the E&M method. Each data point represents one hour time resolution. Blue points represent data with global radiation lower than 50 $W.m^{-2}$ (night-time data). Green points represent data with global radiation higher than 50 $W.m^{-2}$ (day-time data). The red line represents the 1:1 line.

S3. NAIS inlet penetration efficiency

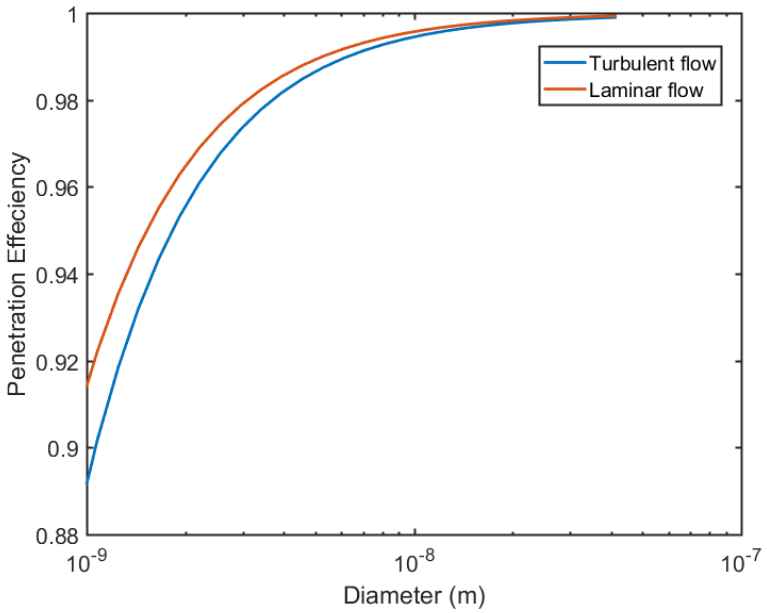

Figure S4. Penetration efficiency through the NAIS inlet based on turbulent or laminar flow calculations.

S4. SMPS hygroscopicity corrections

The "ambient" SMPS particle size distribution was back-calculated from the dry distribution using the hygroscopicity model of Petters and Kreidenweis (2007). This model relies on the Köhler theory, which

describes the equilibrium between the droplet phase and vapor phase. The traditional Köhler equation (Eq. S1) links the equilibrium size of the growing aerosol particle, its chemical composition and water content to the ambient water vapor saturation ratio (S) (Köhler, 1936).

$$S = \frac{P_{w,eq}}{P_{w,sat}} = \frac{RH(D)}{100} = a_w \exp\left(\frac{4\sigma M_w}{RT\rho_w D}\right) \qquad Eq.S1$$

Where:

- $P_{w,eq}$ *is the equilibrium vapor pressure of water over the droplet surface* (Pa)
- $P_{w,sat}$ *is the saturation vapor pressure over a pure flat water surface* (Pa)
- a_w *is the activity of water in solution* $(unitless)$
- M_w *is the molecular weight of water* $(kg.mol^{-1})$
- σ *is the surface tension of the solution* $-$ *air interface* $(\mathrm{N}.m^{-1})$
- ρ_w *is the density of water* $(kg.m^{-3})$
- D *is the diameter of the droplet* (m)

Petters and Kreidenweis (2007) introduced a single hygroscopicity parameter (κ) which described the water activity (a_w) and the difference in the densities and molar masses of water and the dry material:

$$\frac{1}{a_w} = 1 + \kappa \frac{V_{dry}}{V_w} \qquad Eq.S2$$

Where :

- V_{dry} *is the volume of the dry aerosol particle*
- V_w *is the volume of water*

Assuming additive volumes, the Köhler equation can be reformulated to the κ-Köhler equation, which can also be written in the form of hygroscopic growth factor (HGF), which is defined as the ratio between wet particle diameter ($D_{p,wet}$) and dry particle diameter ($D_{p,dry}$):

$$\frac{RH(D)}{100} = \frac{{D_{p,wet}}^3 - D_{p,dry}^3}{{D_{p,wet}}^3 - D_{p,dry}^3(1-\kappa)} \exp\left(\frac{4\sigma M_w}{RT\rho_w D_{p,wet}}\right) \qquad Eq.S3$$

In this study, average seasonal values of κ were retrieved from hygroscopic tandem differential mobility analyzer (HTDMA) measurements performed in parallel to our study (Table S2). The hygroscopic κ values for each SMPS size bin were extrapolated from the HTDMA size-resolved measurements by linear regression. The particle size distribution at ambient RH conditions was then calculated using equation S3, by incorporating the respective κ values per size bin, and the measured size distribution at dry conditions.

Next, the ambient (real) particle diameter was calculated from κ by solving equation S3, which was later used to calculate the real particle size distribution (before drying).

To show an example of the effect of humidity corrected particle size distribution on NPF-related parameters, we compared the dry condensation sink to that calculated when the particle sizes were assumed to be equilibrated to the ambient RH. This comparison shows that the actual condensation sink is sometimes up to 3.5 times higher than the dry condensation sink but on average it is between 1.1 and 1.3 times higher than the dry one (Figure S4).

Table S2. HTDMA derived kappa (κ) parameter.

	HTDMA derived Kappa				
Diameter (nm)	Spring	Summer	Fall	Winter	Average
30	0.19	0.23	0.14	0.16	0.18
80	0.19	0.28	0.17	0.15	0.2
160	0.22	0.26	0.21	0.22	0.23

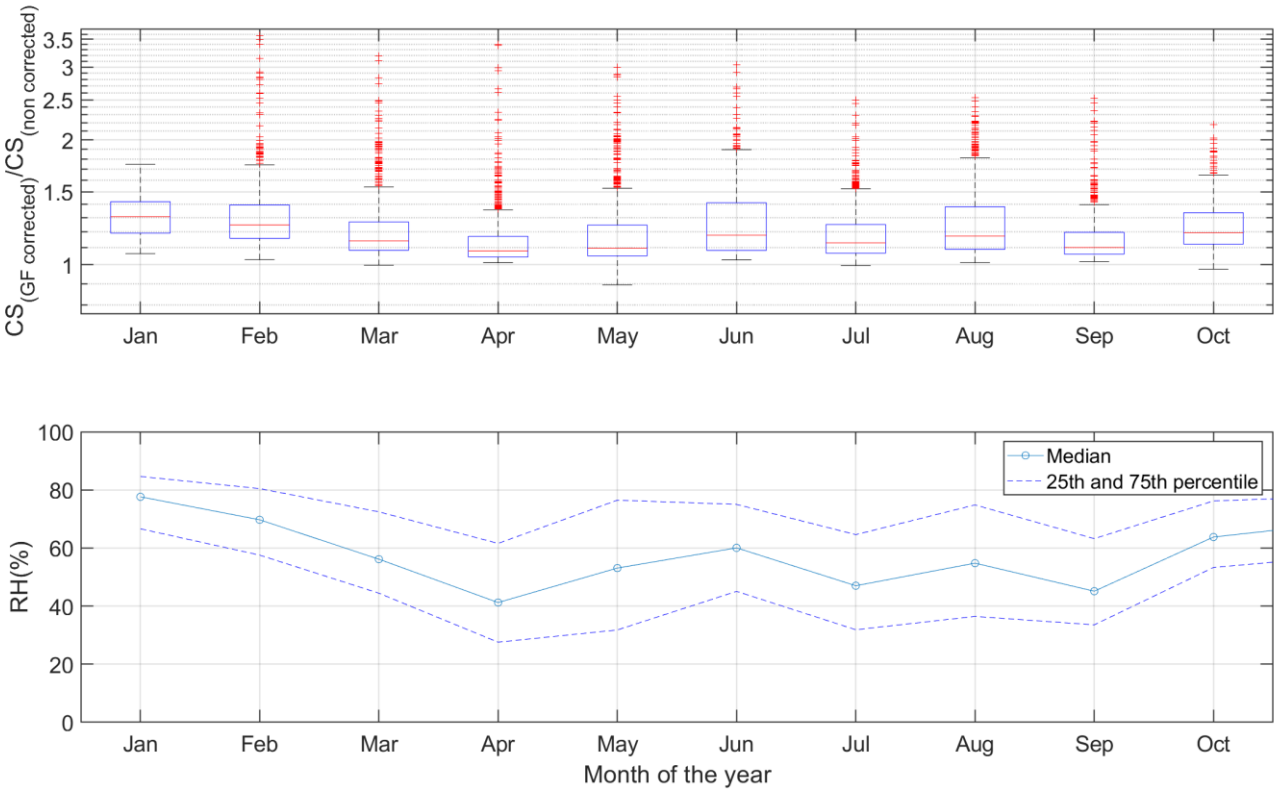

Figure S5. The top panel shows the effect of particle hygroscopic growth factor (GF) on condensation sink (CS) calculations presented as the ratio between condensation sink calculated from the "ambient" distribution and condensation sink calculated from the "dry" distribution. The bottom and top edges of the box plot represent 25% and 75% percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. The bottom panel shows median RH (%) with 25th and 75th percentiles.

S5. Identification of days with high dust loading

The method proposed by Drinovec et al. (2020) permits the calculation of mineral dust concentrations with high time resolution using the following equation

$$Mineral\ dust_{PM_{10-1}} = \frac{b_{abs,VI} - b_{abs,PM_1}}{EF \times MAC} \qquad Eq.S4$$

Where $b_{abs,VI}$ is the absorption coefficient (at 370nm) measured by the aethalometer (model AE33, Magee Scientific, USA) coupled to a virtual impactor (VI), b_{abs,PM_1} is the absorption coefficient (at 370nm) measured by a second AE33 Aethalometer sampling through a PM_1 sharp-cut cyclone, EF is the enhancement factor of the VI and MAC is the mass absorption cross-section for dust. The last two coefficients were used as

determined experimentally by Drinovec et al. (2020) where additional information about the method and the instruments used can be found.
From the mineral dust daily time series, we defined a daily threshold above which a day is considered having high dust loading (Table S3). When aethalometer measurements were not available, coarse particle mass loading (PM_{10} - $PM_{2.5}$), determined by a Tapered Element Oscillating Microbalance (TEOM), was used to identify dust days. Additional information about the TEOM used can be found in Pikridas et al. (2018). The threshold for coarse PM was defined based on the linear regression between coarse PM and mineral dust concentration.

Table S3. List of dates with high dust loading

6-Feb-18	21-Mar-18	26-Apr-18	22-May-18	23-Oct-18
7-Feb-18	22-Mar-18	27-Apr-18	23-May-18	24-Oct-18
8-Feb-18	23-Mar-18	1-May-18	24-May-18	31-Oct-18
9-Feb-18	24-Mar-18	2-May-18	8-Jun-18	1-Nov-18
10-Feb-18	25-Mar-18	3-May-18	9-Jun-18	2-Nov-18
5-Mar-18	26-Mar-18	4-May-18	23-Jul-18	3-Nov-18
6-Mar-18	27-Mar-18	5-May-18	24-Jul-18	4-Nov-18
7-Mar-18	28-Mar-18	6-May-18	18-Oct-18	24-Jan-19
8-Mar-18	19-Apr-18	7-May-18	19-Oct-18	25-Jan-19
20-Mar-18	20-Apr-18	21-May-18	21-Oct-18	26-Jan-19

S6. Time range of Daytime conditions (global radiation > 50 W m^{-2})

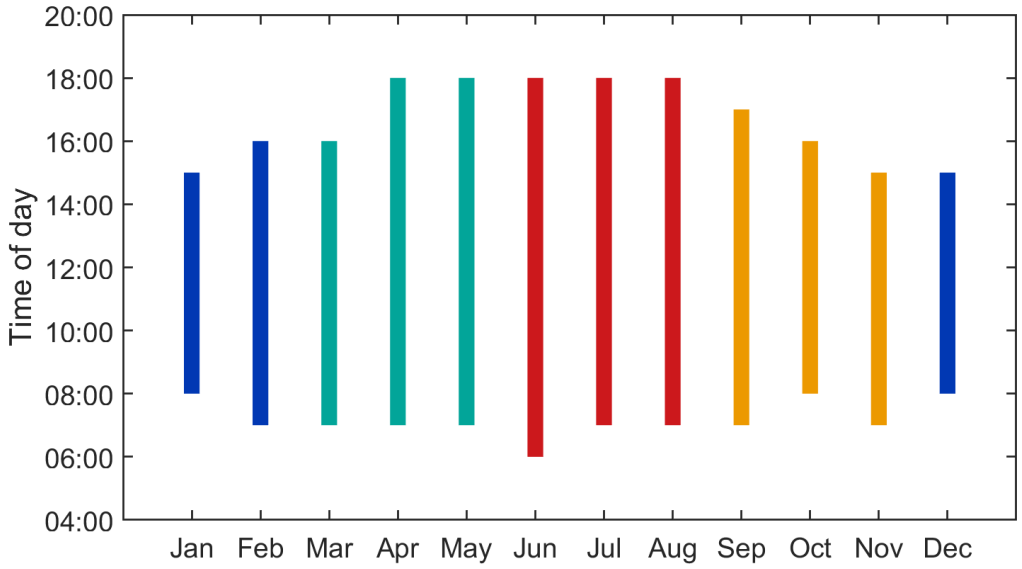

Figure S6. Monthly range of time of day having global radiation > 50 W. m^{-2}.

S7. Calendar of daily event day classification

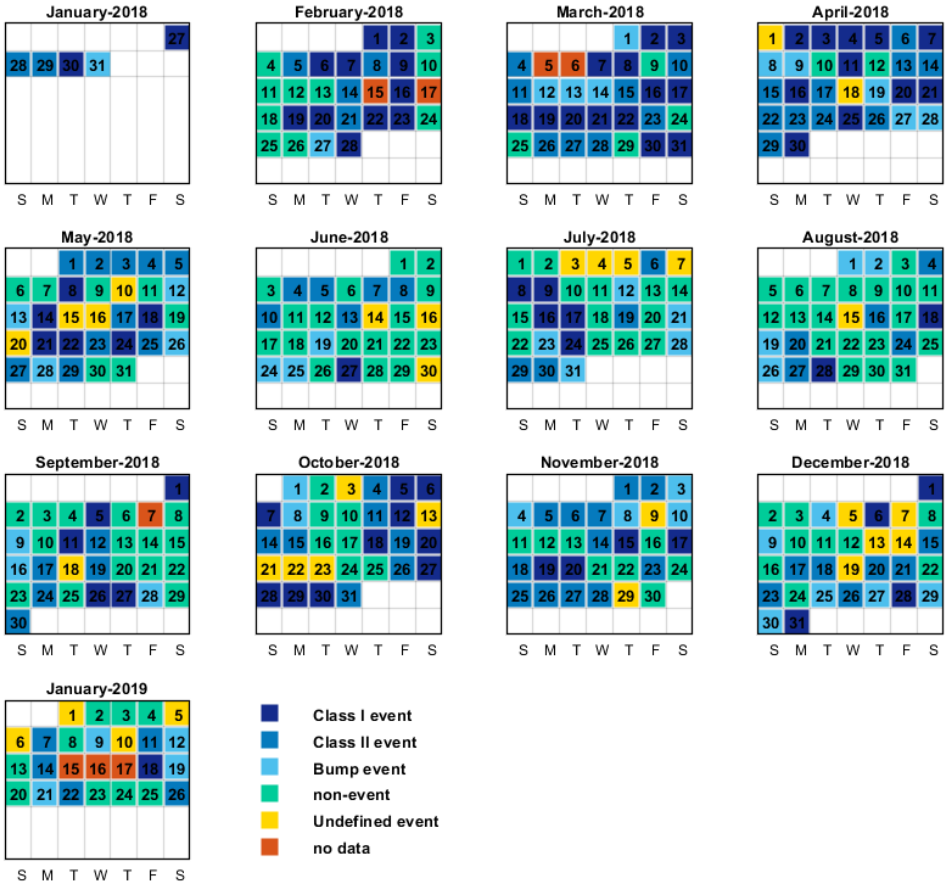

Figure S7. Calendar of daily event day classification between January 27, 2018 –January 26, 2019

S8. Example of event classes

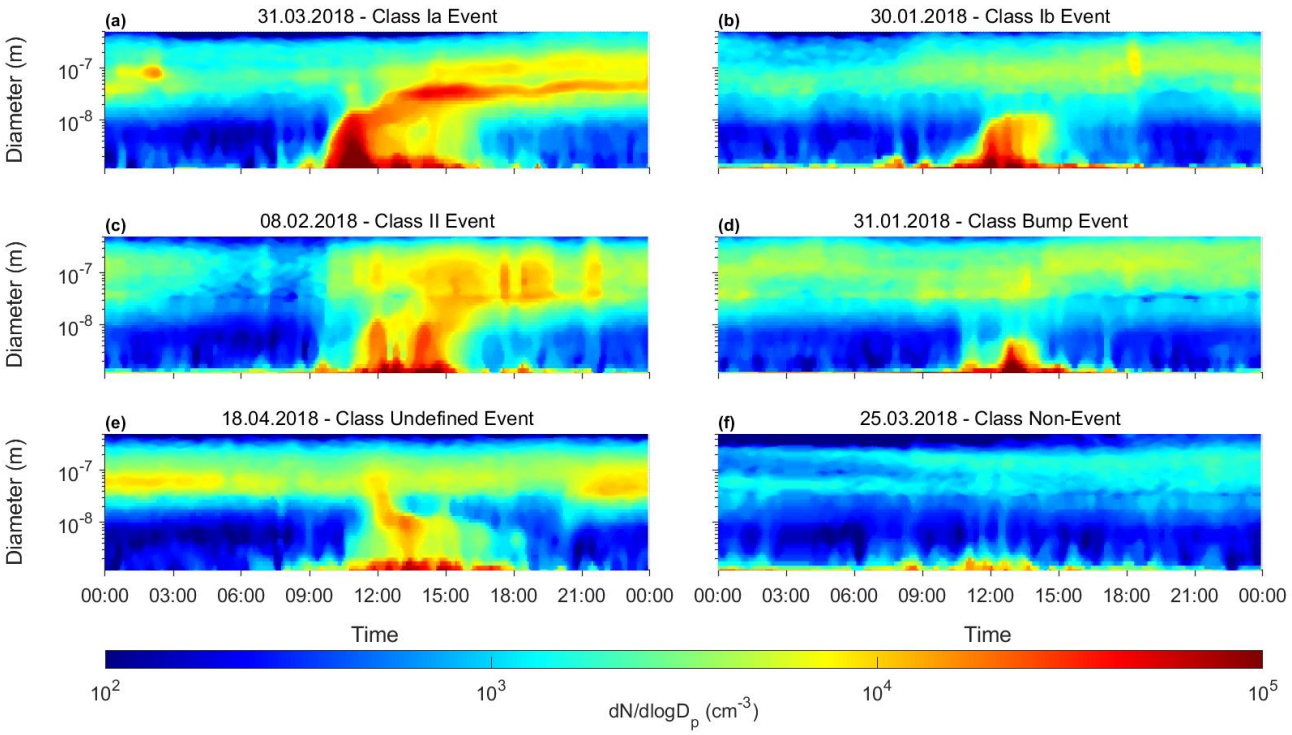

Figure S8. Examples of class Ia (a), class Ib (b), class II (c), bump (d), undefined (e) and non-events (f).

S9. Formation rates and growth rates with respect to high dust loading

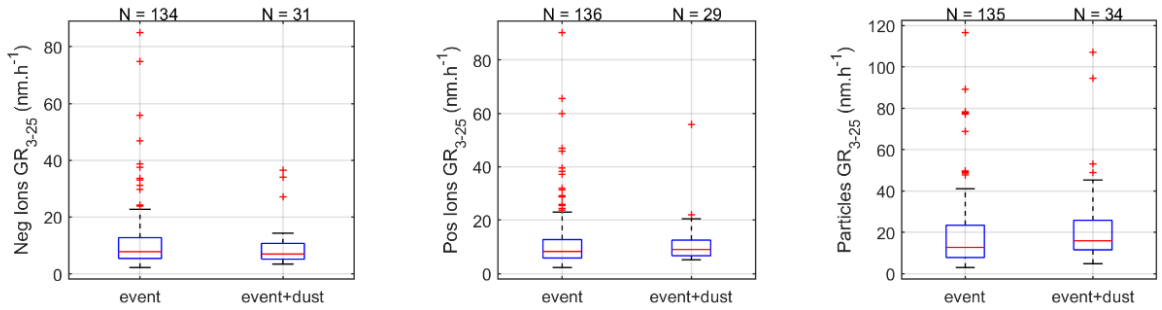

Figure S9. Boxplots of $GR_{3\text{-}25}$ of negative ions, positive ion and particles during events not affected by high dust loading and events affected by high dust loading

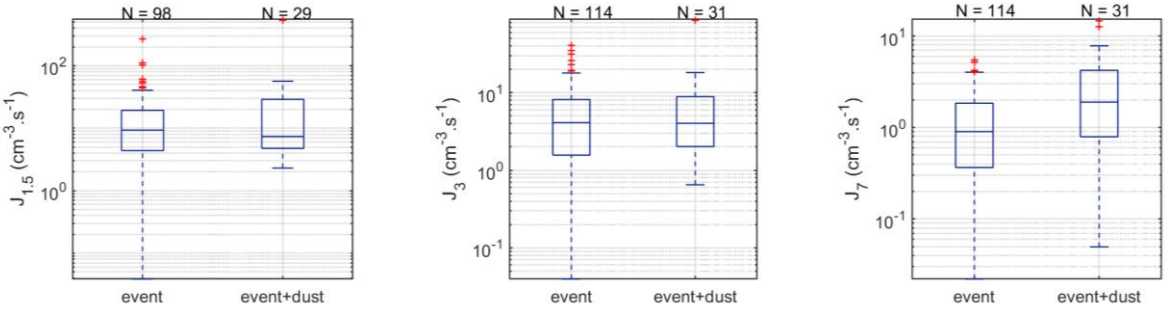

Figure S10. Boxplots of formation rates ($J_{1.5}$, J_3, J_7) during events not affected by high dust loading and events affected by high dust loading

S10. NPF specific parameters

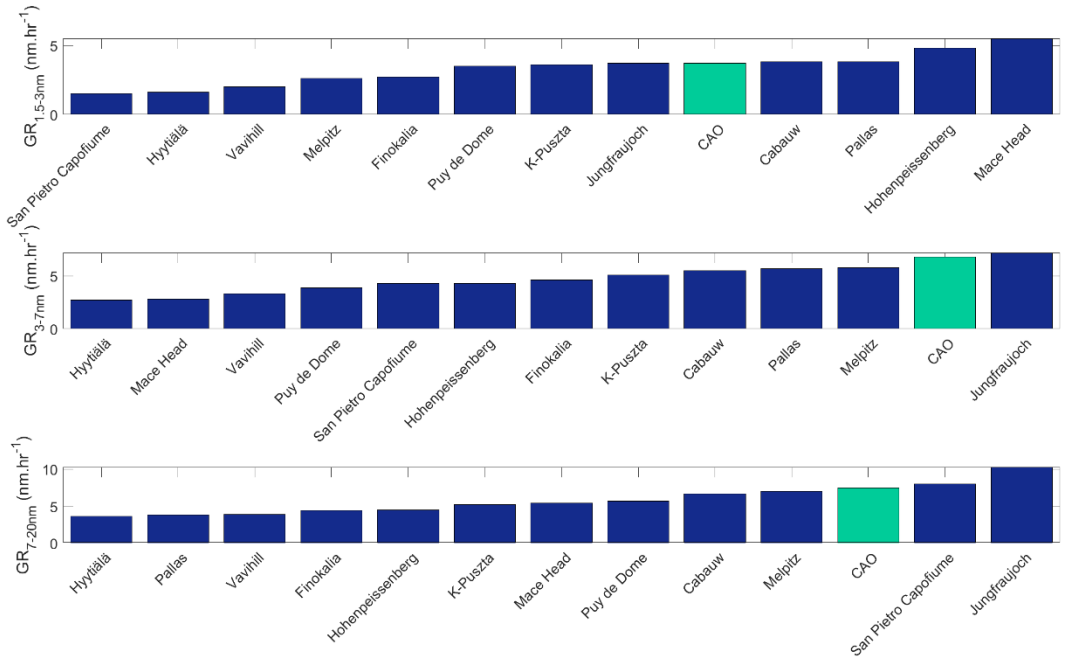

Figure S11. Comparison of ion mode growth rates measured in this study to growth rates measured at 12 European sites (Manninen et al., 2010).

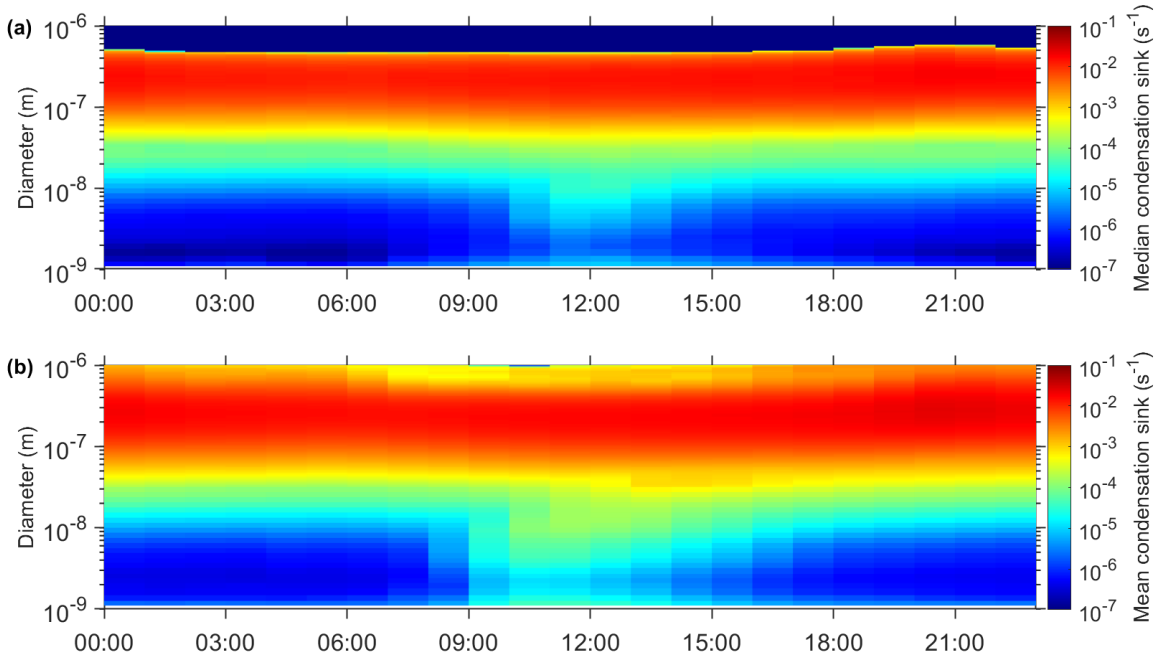

Figure S12. The median (a) and mean (b) averages of the diurnal size segregated condensation sink (s^{-1}) computed over the whole measurement period of this study.

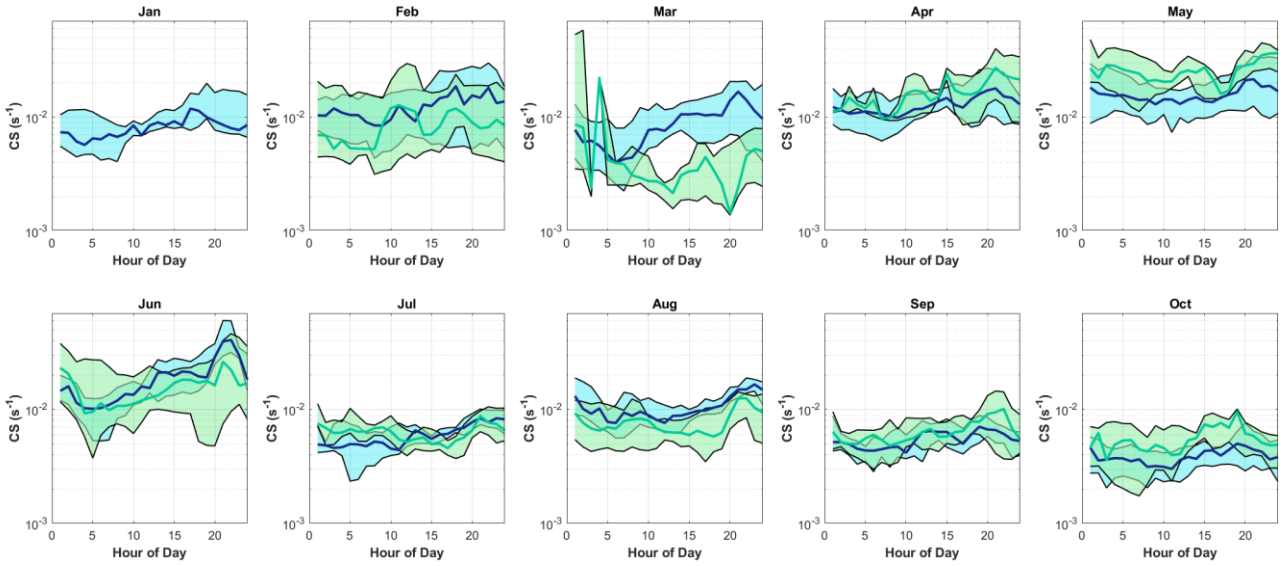

Figure S13. The monthly diurnal cycle of condensation sink (s^{-1}) during event (blue) and non-event (green) days. The shaded areas represent 25th to 75th percentile while the solid line represents the median

S11. The relation between some parameters and NPF events

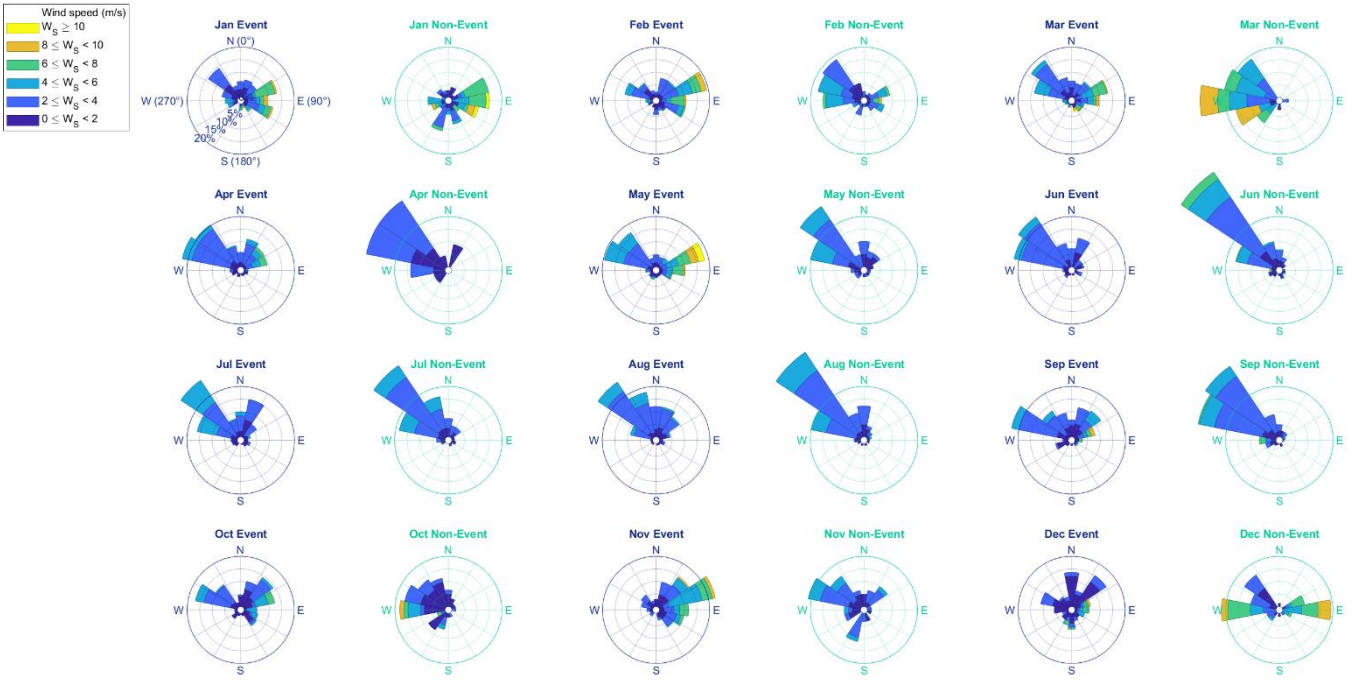

Figure S14. Month wind roses during event and non-event days using data corresponding to global radiation greater than 50 W.m^{-2} .Data presented have hourly time resolution.

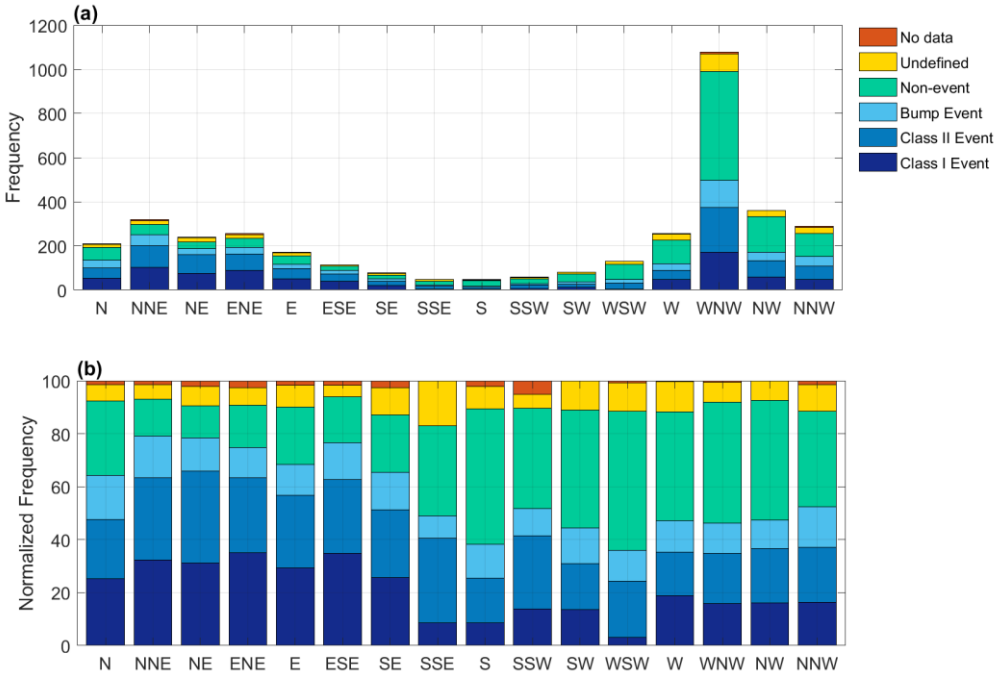

Figure S15. The frequency (a) and normalized frequency (b) of hourly wind direction data divided to 16 sectors color-coded by the event classification

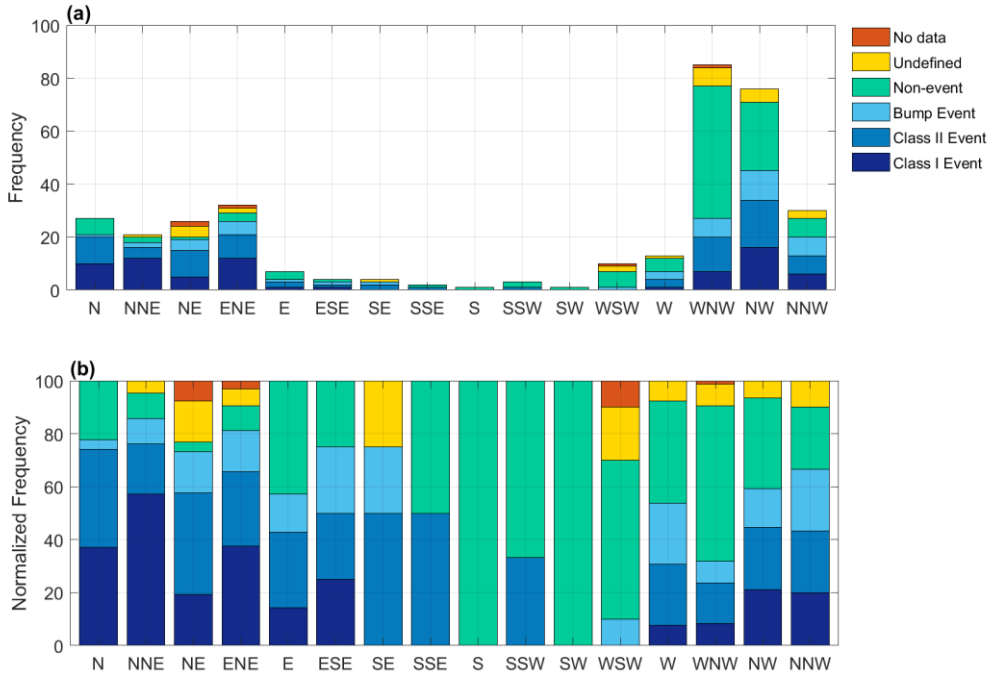

Figure S16. The frequency (a) and normalized frequency (b) of daily wind direction data divided to 16 sectors color-coded by the event classification

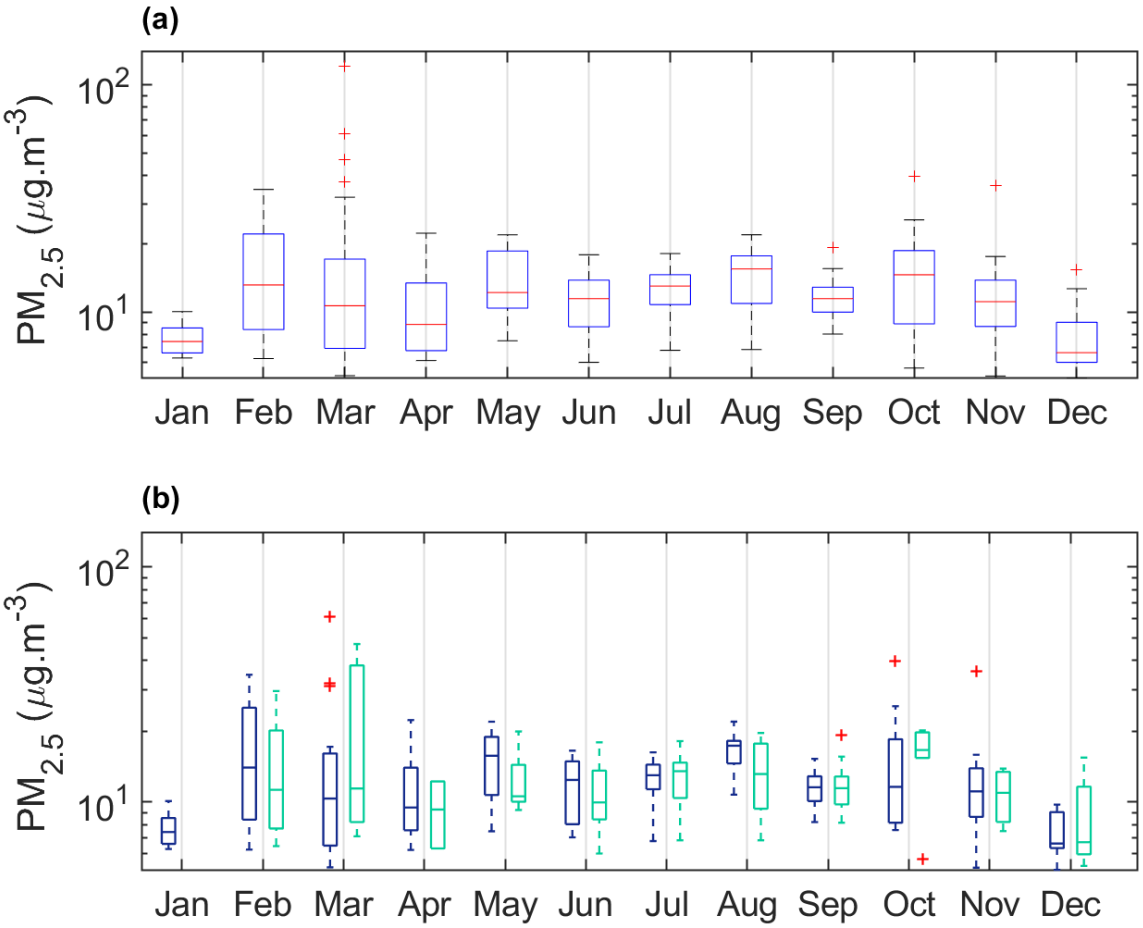

Figure S17. (a) Monthly variation of $PM_{2.5}$ (µg.m^{-3}). (b) Monthly variation of $PM_{2.5}$ (µg.m^{-3}) separated between event (blue) and non-event (green) days. The bottom and top edges of the box plots indicate the 25th and 75th percentiles, respectively. The central mark indicates the median. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. Data presented have daily time resolution.

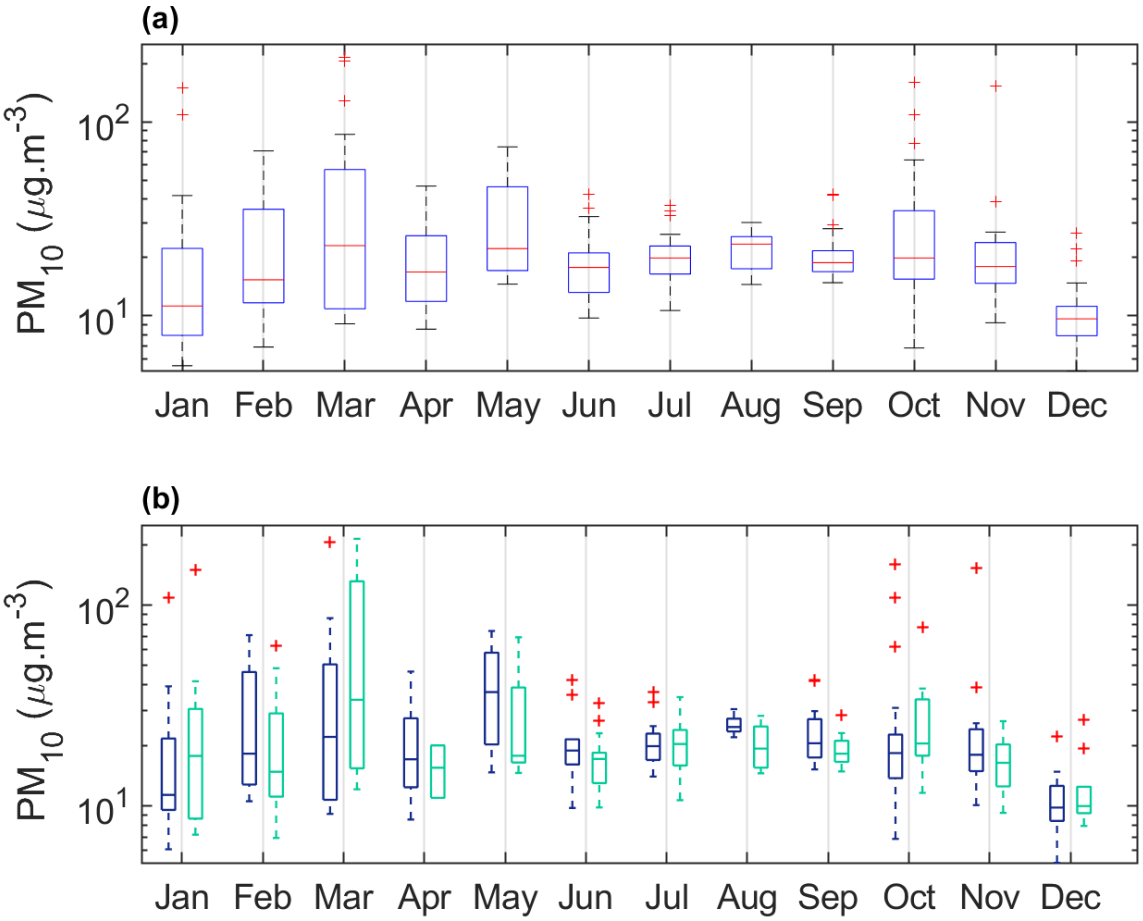

Figure S18. (a) Monthly variation of PM_{10} (µg.m^{-3}). (b) Monthly variation of PM_{10} (µg.m^{-3}) separated between event (blue) and non-event (green) days. The bottom and top edges of the box plots indicate the 25th and 75th percentiles, respectively. The central mark indicates the median. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. Data presented have daily time resolution

S12. Regression and classification analysis

S12.1 Stepwise linear regression analysis:

Data pretreatment:

We used hourly data for the regression analysis. Before performing the analysis, we applied a logarithmic transformation to the predictor variables having a skewed distribution. Then, we used Belsley collinearity diagnostics for assessing the strength and sources of collinearity among the predictor variables (Belsley et al., 1980). The remaining predictor variables after removing the variables that exhibited collinearity were NO, NO_2, CO, RH, temperature, solar radiation, wind direction, wind speed, $PM_{2.5}$ and sulfuric acid. We transformed the wind direction data into a categorical variable with four levels (N to E; E to S; S to W; W to N) to avoid data circularity. We removed the data corresponding to nighttime hours (solar radiation < 50 W/m^2), and undefined days from the analysis for a better separation between events and non-events. We further excluded any observation with any missing variable. Finally, we normalized all variables to make sure that all variables are of equal weight. Figure S19 shows the available data presented as correlation matrices.

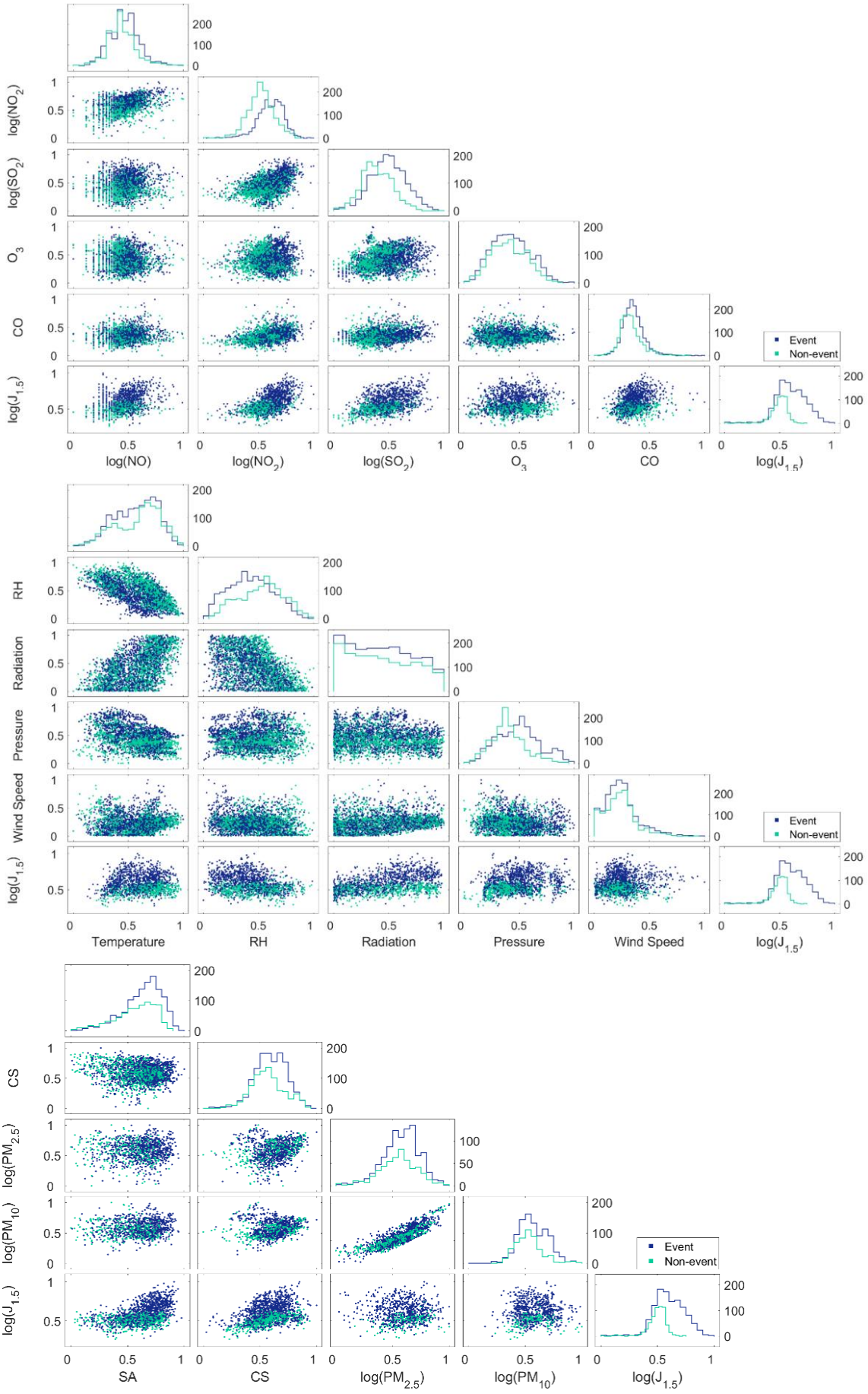

Figure S19. Correlation matrix of hourly atmospheric variables during event and non-event days.

Analysis:

We performed stepwise regression on the hourly data set to fit a linear model that would best describe $J_{1.5}$, which is the response variable. The steps were bidirectional starting from a model having no predictor terms and at each step, searching for terms to add to the model or remove from the model based on a pre-specified optimization criterion. Here we used the Akaike information criterion (AIC) as the optimization criterion, which is an estimator of prediction error and thereby the relative quality of the statistical model (Yamashita et al., 2007). First, we ran the stepwise linear regression by setting the upper bounds of the model to have an intercept and a linear term for each predictor (basic linear model; Model 1). Based on AIC criteria, NO_2, RH, solar radiation, H_2SO_4 and wind direction are the most important terms for the model. The importance of the aforementioned variables is also reflected by the coefficients for these terms and their p-values; Figure S20.a). CO and NO had a positive influence on $J_{1.5}$ but they were not as important as the aforementioned variables based on AIC criteria. Temperature and wind direction from the east to south sector had very little effect and were excluded from the final model. Wind speed, $PM_{2.5}$, RH and wind from the south to west sector had a negative influence on $J_{1.5}$, but the coefficients for the last two terms did not pass the significance level.

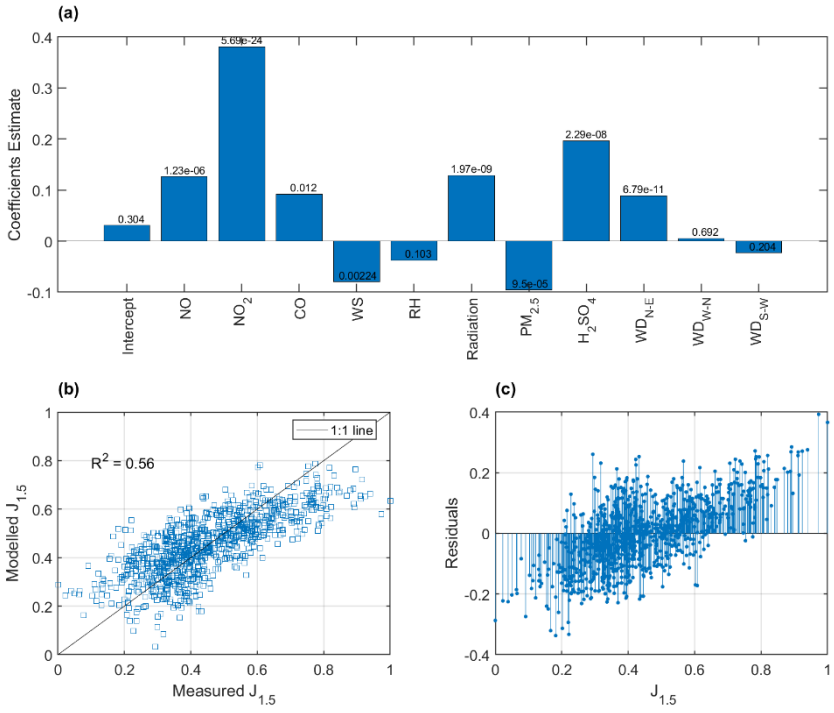

Figure S20. Model 1: (a) coefficients of model terms with p-values presented above the bars. (b) modelled versus measured $J_{1.5}$. (c) distribution of model residuals with respect to measured $J_{1.5}$.

The basic model had a coefficient of determination (R^2) of 0.56, a root mean square error (RMSE) of 0.1135 and it was not able to fit the data at high $J_{1.5}$ properly (Figure S20.b). The residuals of the model (Figure S20.c) indicate that there might be a missing predictor variable. We tried optimizing the regression model by allowing interaction terms (Model 2) and quadratic terms (Model 3). These models show enhanced RMSE and R^2 but they were also not able to fit the lowest and highest $J_{1.5}$ values properly. Figure S21 shows the optimized interactions model (Model 2; RMSE=0.0966). In general, this model is more complex to analyze because of overlapping terms. However, it shows that H_2SO_4 (represented as SA in the figure) has a positive influence when coupled with NO_2 and solar radiation and a negative influence when coupled with RH and temperature. It also shows that H_2SO_4, NO_2, and wind direction are the most important variables to explain $J_{1.5}$. In fact, we get a reasonable response when including these terms only in the regression (Model 4; Figure S22). However,

none of the models gave a good response for the lowest and highest $J_{1.5}$ values, which could be yet another indication that there is a missing variable not included yet in the regression.

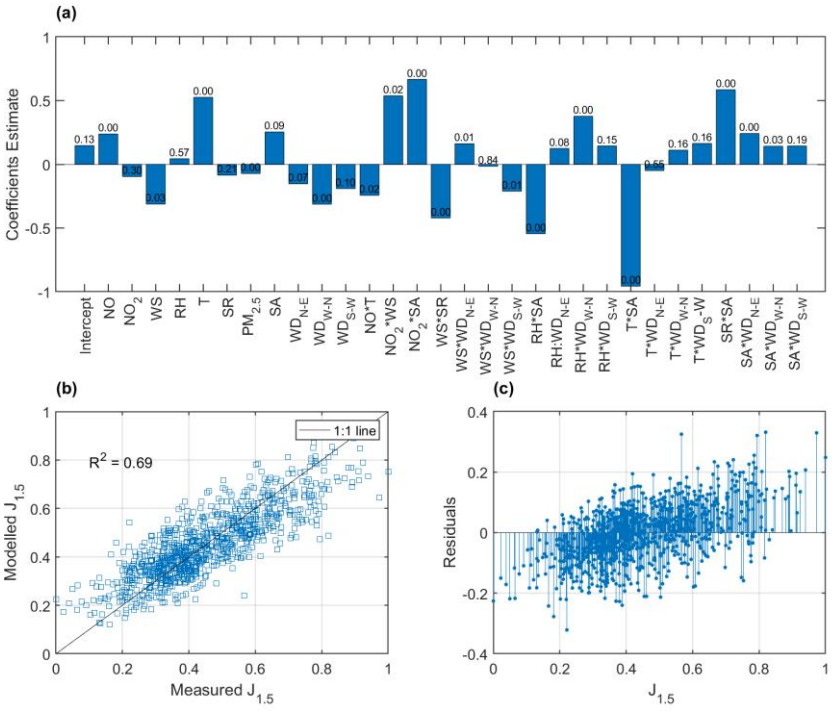

Figure S21. Model 2: (a) coefficients of model terms with p-values presented above the bars. (b) modelled versus measured $J_{1.5}$. (c) distribution of model residuals with respect to measured $J_{1.5}$.

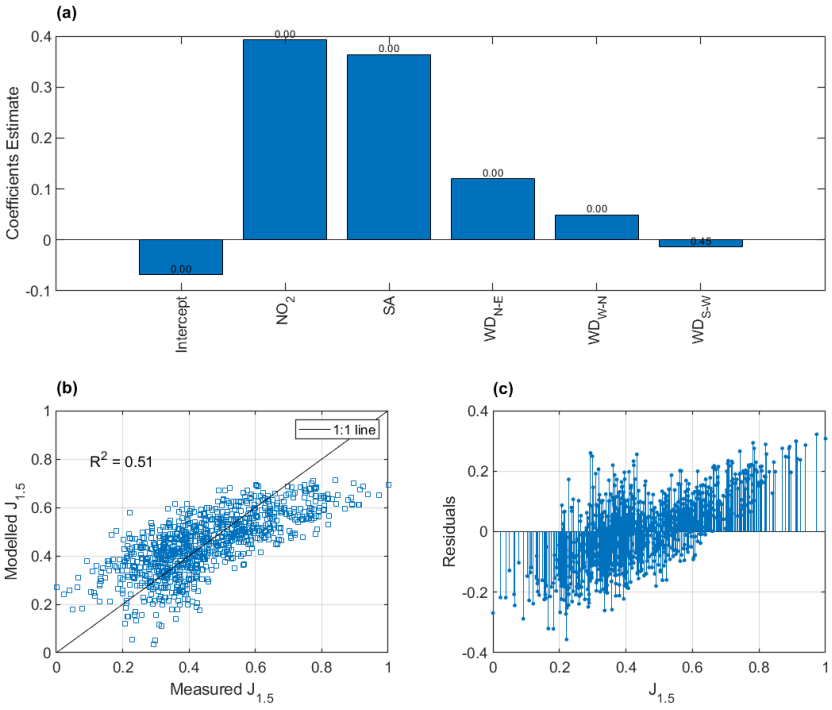

Figure S22. Model 4: (a) coefficients of model terms with p-values presented above the bars. (b) modelled versus measured $J_{1.5}$. (c) distribution of model residuals with respect to measured $J_{1.5}$.

S12.2 Classification decision trees:

Classification trees comprise one of the most commonly used non-parametric classification approaches in machine learning and data mining. They recursively partition the feature space into a set of leaves with the most homogeneous collection of outcome possible (Breiman et al., 1984).

Data pretreatment:

We used daily data for the classification analysis. The daily data was computed as the mean of daytime (solar radiation < 50 W/m^2) hourly observations and was only calculated if there are more than 75% of hourly observations within the appropriate time window. Similar to the regression analysis, we removed the variables that exhibited multicollinearity. We also removed both PM_{10} and $PM_{2.5}$ data because they exhibited many missing values. The remaining predictor variables were NO, NO_2, CO, O_3, RH, temperature, solar radiation, wind direction, wind speed and H_2SO_4. We further excluded the bump events from the analysis because the decision trees usually misclassified them. The number of days after removing undefined events and bump events was 279. Finally, we excluded any observation with any missing variable. The total observation days were thus reduced to 184 days (115 event days and 79 non-events days). We did not normalize or log-transform the data because these procedures are not necessary for decision trees.

Analysis:

The classification tree hyperparameters (maximum depth, minimum number of samples required to split an internal node, minimum number of samples required to be at a leaf node and the split criterion) were tuned until the best performance was reached. The performance of the trees was evaluated using performance metrics (accuracy, sensitivity, specificity and precision), 10-fold cross-validation error, and re-substitution error. The outcome decision tree is shown in Figure S23, while the confusion matrix and predictor importance is shown in Figure S24, and the statistics for each node are shown in table S4. The decision tree model had an accuracy (ratio of the correctly labeled days to the whole pool of days) of 89%, a sensitivity (percentage of the labeled events to true events) of 89.6%, a specificity (percentage of labeled non-events to true non-events) of 88.4%, a resubstitution error of 0.1087, and a 10-fold cross-validation error of 0.2337. The decision tree shows that when NO_2 is greater than 0.88 ppb, event days occurred when H_2SO_4 was greater than $1.4e^6$ $molecules.cm^{-3}$. When NO_2 is less than 0.88 ppb, the event occurrence could be split based on a combination of wind direction, solar radiation, RH and H_2SO_4 concentration. When the wind is from the N-E or S-W direction, events coincided with RH<54.4%. While at E-S and W-N wind direction, events coincided with solar radiation between 584 and 620 $W.m^{-2}$ and H_2SO_4 concentration>$3.1e^6$ $molecules.cm^{-3}$. The analysis also showed that events did not occur when the solar radiation was > 620 $W.m^{-2}$. It is important to note here that the decision tree tries to find the best variable that could split the underlying days into event and non-event days. Since a physical explanation cannot be given for why event days are not occurring at solar radiation above 620 $W.m^{-2}$, it is most probable that the analysis is missing other variables that have an association with solar radiation.

Final remarks:

While the analysis shown in this section provides an important insight into the parameters governing NPF events, it is important to present them with caution because causality cannot be inferred by this analysis. Additionally, the sample size and limitations of available predictors heavily affected the output of both models. A yearlong dataset with missing values cannot provide adequate counting statistics for every NPF case. Therefore, this analysis is useful to discern the importance of specific chemical atmospheric components or physical properties on NPF but cannot be used to predict future nucleation events characteristics

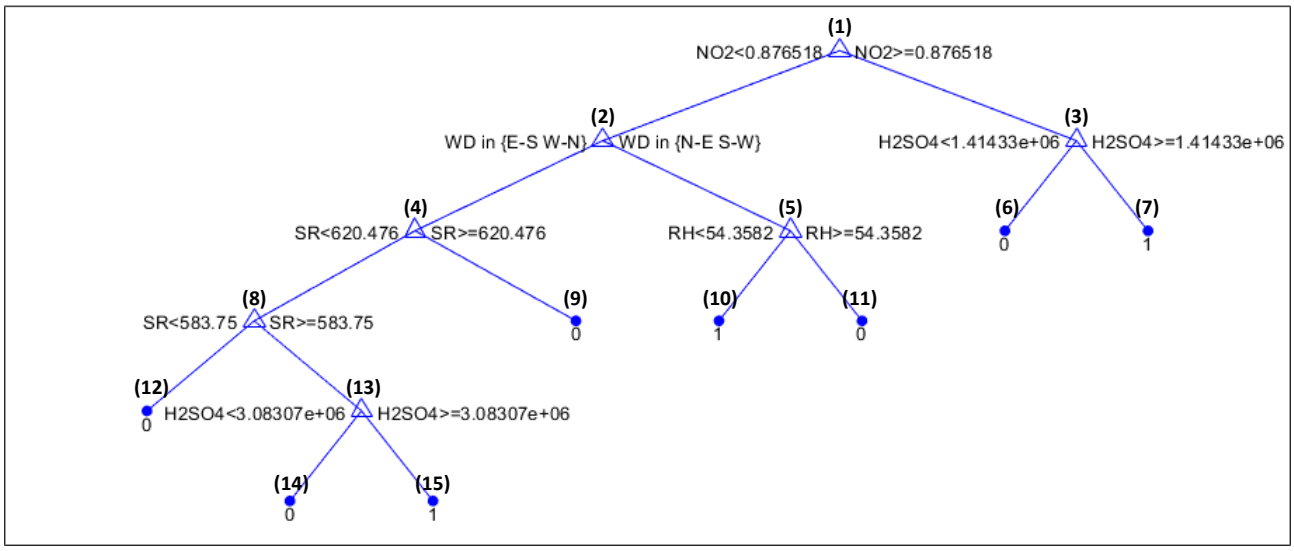

Figure S23. Daily NPF occurrence decision tree. The number of each node is displayed between parentheses above the node. Branch nodes are represented with triangles, while leaf nodes are presented with circles. Non-events are represented by 0, and events are represented with 1. WD is wind direction; SR is solar radiation in ($W.m^{-2}$); NO_2 is in ppb; RH is in %, and H_2SO_4 is in $molecules.cm^{-3}$.

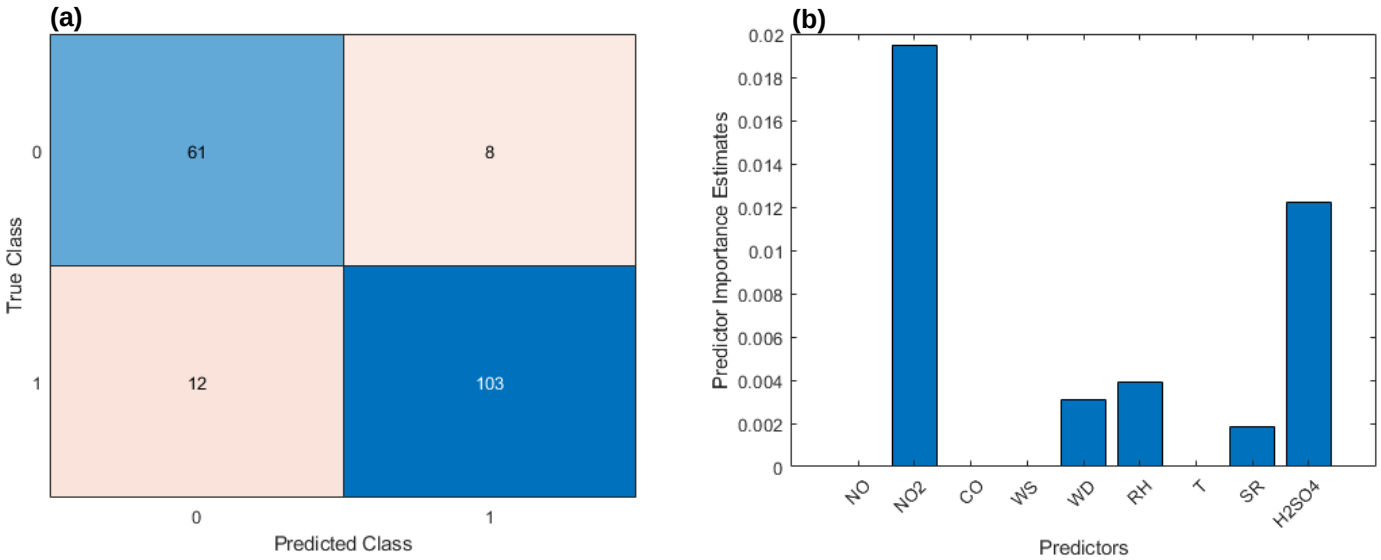

Figure S24. (a) The classification tree confusion matrix where class 0 represents non-event days and class 1 represent event days. (b) The importance estimate of predictors.

Table S4. Description and content of the decision tree nodes

Node no	Node type	Node class	Node size	No of non-events	No of events	Misclassified	Node error
1	branch	event	184	69	115	-	0.38
2	branch	non-event	78	53	25	-	0.32
3	branch	event	106	16	90	-	0.15
4	branch	non-event	60	46	14	-	0.23
5	branch	event	18	7	11	-	0.39
6	leaf	non-event	14	11	3	3	0.21
7	leaf	event	92	5	87	5	0.05
8	branch	non-event	48	35	13	-	0.27
9	leaf	non-event	12	11	1	1	0.08
10	leaf	event	11	1	10	1	0.09
11	leaf	non-event	7	6	1	1	0.14
12	leaf	non-event	33	27	6	6	0.18
13	branch	non-event	15	8	7	-	0.47
14	leaf	non-event	7	6	1	1	0.14
15	leaf	event	8	2	6	2	0.25

References

Belsley, D., Kuh, E., and Welsch, R.: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, in, John Wiley & Sons, Inc., New York, 1980.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A.: Classification and regression trees, CRC press, 1984.

Cai, R., Yang, D., Ahonen, L. R., Shi, L., Korhonen, F., Ma, Y., Hao, J., Petäjä, T., Zheng, J., Kangasluoma, J., and Jiang, J.: Data inversion methods to determine sub-3 nm aerosol size distributions using the particle size magnifier, Atmospheric Measurement Techniques, 11, 4477-4491, https://doi.org/10.5194/amt-11-4477-2018, 2018.

Chan, T., Cai, R., Ahonen, L. R., Liu, Y., Zhou, Y., Vanhanen, J., Dada, L., Chao, Y., Liu, Y., Wang, L., Kulmala, M., and Kangasluoma, J.: Assessment of particle size magnifier inversion methods to obtain particle size distribution from atmospheric measurements, Atmospheric measurement Techniques Discussions, 2020, 1-21, https://doi.org/10.5194/amt-2019-465, 2020.

Drinovec, L., Sciare, J., Stavroulas, I., Bezantakos, S., Pikridas, M., Unga, F., Savvides, C., Višić, B., Remškar, M., and Močnik, G.: A new optical-based technique for real-time measurements of mineral dust concentration in PM10 using a virtual impactor, Atmospheric Measurement Techniques, 13, 3799-3813, https://doi.org/10.5194/amt-13-3799-2020, 2020.

Kangasluoma, J., Junninen, H., Lehtipalo, K., Mikkilä, J., Vanhanen, J., Attoui, M., Sipilä, M., Worsnop, D., Kulmala, M., and Petäjä, T.: Remarks on Ion Generation for CPC Detection Efficiency Studies in Sub-3-nm Size Range, Aerosol Science and Technology, 47, 556-563, http://dx.doi.org/10.1080/02786826.2013.773393, 2013.

Kangasluoma, J., Franchin, A., Duplissy, J., Ahonen, L., Korhonen, F., Attoui, M., Mikkilä, J., Lehtipalo, K., Vanhanen, J., Kulmala, M., and Petäjä, T.: Operation of the Airmodus A11 nano Condensation Nucleus Counter at various inlet pressures and various operation temperatures, and design of a new inlet system, Atmospheric Measurement Techniques, 9, 2977-2988, https://doi.org/10.5194/amt-9-2977-2016, 2016.

Köhler, H.: The nucleus in and the growth of hygroscopic droplets, Transactions of the Faraday Society, 32, 1152-1161, https://doi.org/10.1039/TF9363201152, 1936.

Manninen, H. E., Nieminen, T., Asmi, E., Gagné, S., Häkkinen, S., Lehtipalo, K., Aalto, P., Vana, M., Mirme, A., Mirme, S., Hõrrak, U., Plass-Dülmer, C., Stange, G., Kiss, G., Hoffer, A., Törő, N., Moerman, M., Henzing, B., de Leeuw, G., Brinkenberg, M., Kouvarakis, G. N., Bougiatioti, A., Mihalopoulos, N., O'Dowd, C., Ceburnis, D., Arneth, A., Svenningsson, B., Swietlicki, E., Tarozzi, L., Decesari, S., Facchini, M. C., Birmili, W., Sonntag, A., Wiedensohler, A., Boulon, J., Sellegri, K., Laj, P., Gysel, M., Bukowiecki, N., Weingartner, E., Wehrle, G., Laaksonen, A., Hamed, A., Joutsensaari, J., Petäjä, T., Kerminen, V. M., and Kulmala, M.: EUCAARI ion spectrometer measurements at 12 European sites – analysis of new particle formation events, Atmospheric Chemistry and Physics, 10, 7907-7927, https://doi.org/10.5194/acp-10-7907-2010, 2010.

Petters, M. D., and Kreidenweis, S. M.: A single parameter representation of hygroscopic growth and cloud condensation nucleus activity, Atmospheric Chemistry and Physics, 7, 1961-1971, https://doi.org/10.5194/acp-7-1961-2007, 2007.

Pikridas, M., Vrekoussis, M., Sciare, J., Kleanthous, S., Vasiliadou, E., Kizas, C., Savvides, C., and Mihalopoulos, N.: Spatial and temporal (short and long-term) variability of submicron, fine and sub-10 μm particulate matter (PM1, PM2.5, PM10) in Cyprus, Atmospheric Environment, 191, 79-93, https://doi.org/10.1016/j.atmosenv.2018.07.048, 2018.

Yamashita, T., Yamashita, K., and Kamimura, R.: A Stepwise AIC Method for Variable Selection in Linear Regression, Communications in Statistics - Theory and Methods, 36, 2395-2403, https://doi.org/10.1080/03610920701215639, 2007.