Eight years of sub-micrometre organic aerosol composition data from the boreal forest characterized using a machine-learning approach

Liine Heikkinen[1], Mikko Äijälä[1], Kaspar R. Daellenbach[1], Gang Chen[2], Olga Garmash[1], Diego Aliaga[1], Frans Graeffe[1], Meri Räty[1], Krista Luoma[1], Pasi Aalto[1], Markku Kulmala[1], Tuukka Petäjä[1], Douglas Worsnop[1,3], and Mikael Ehn[1]

[1]Institute for Atmospheric and Earth System Research /Physics, Faculty of Science, University of Helsinki, Helsinki, FI–00014, Finland
[2]Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, Villigen, Switzerland
[3]Aerodyne Research Inc., Billerica, MA, USA

Correspondence to: Liine Heikkinen (liine.heikkinen@helsinki.fi) and Mikael Ehn (mikael.ehn@helsinki.fi)

Abstract.

The Station for Measuring Ecosystem Atmosphere Relations (SMEAR) II, located within the boreal forest of Finland, is a unique station in the world due to the wide range of long-term measurements tracking the Earth-atmosphere interface. In this study, we characterize the composition of organic aerosol (OA) at SMEAR II by quantifying its driving constituents. We utilize a multi-year data set of OA mass spectra measured *in situ* with an Aerosol Chemical Speciation Monitor (ACSM) at the station. To our knowledge, this mass spectral time series is the longest of its kind published to date. Similarly to other, previously reported efforts in OA source apportionment from multi-seasonal or –annual data sets, we approached the OA characterization challenge through Positive Matrix Factorization (PMF) using a rolling window approach. However, the existing methods for extracting minor OA components were found to be insufficient for our rather remote site. To overcome this issue, we tested a new statistical analysis framework. This included unsupervised feature extraction and classification stages to explore a large number of unconstrained Positive Matrix Factorisation (PMF) runs conducted on the measured OA mass spectra. Anchored by these results, we finally constructed a relaxed Chemical Mass Balance (CMB) run that resolved different OA species from our observations. The presented combination of statistical tools provided a data driven analysis methodology, which in our case achieved robust solutions with minimal subjectivity.

Following the extensive statistical analyses, we were able to divide the 2012–2019 SMEAR II OA data (mass concentration interquartile range (IQR): 0.7, 1.3, 2.6 μg m^{-3}) to three sub-categories: low-volatility oxygenated OA (LV-OOA), semi-volatile oxygenated OA (SV-OOA), and primary OA (POA) proving that the tested methodology was able to provide results consistent which literature. LV-OOA was the most dominant OA type (organic mass fraction IQR: 49, 62, and 73%). The seasonal cycle of LV-OOA was bimodal, with peaks both in summer and in February. We associated the wintertime LV-OOA with anthropogenic sources and assumed biogenic influence in LV-OOA formation in summer. Through a brief trajectory analysis, we estimated summertime natural LV-OOA formation of tens of ng m^{-3} h^{-1} over the boreal forest. SV-OOA was the second

highest contributor to OA mass (organic mass fraction IQR: 19, 31, and 43%). Due to SV-OOA's clear peak in summer, we estimate biogenic processes as the main drivers in its formation. Unlike for LV-OOA, the highest SV-OOA concentrations were detected in stable summertime nocturnal surface layers. Two nearby sawmills also played a significant role in SV-OOA production as also exemplified by previous studies at SMEAR II. POA, taken as a mix of two different OA types reported previously, hydrocarbon-like OA (HOA) and biomass burning OA (BBOA), made up a minimal OA mass fraction (IQR: 2, 6, and 13%). Notably, the quantification of POA at SMEAR II using ACSM data was not possible following existing rolling PMF methodologies. Both POA organic mass fraction and mass concentration peaked in winter. Its appearance at SMEAR II was linked to strong southerly winds. Similar wind direction and speed dependence was not observed among other OA types. The high wind speeds probably enabled the POA transport to SMEAR II from faraway sources in a relatively fresh state. In case of slower wind speeds, POA likely evaporated and/or aged into oxidized organic aerosol before detection. The POA organic mass fraction was significantly lower than reported by aerosol mass spectrometer (AMS) measurements two to four years prior to the ACSM measurements. While the co-located long-term measurements of black carbon supported the hypothesis of higher POA loadings prior to year 2012, it is also possible that short term (POA) pollution plumes were averaged out due to the slow time resolution of the ACSM combined with the further 3-hour data averaging needed to ensure good signal-to-noise ratios (SNR). Despite the length of the ACSM data set, we did not focus on quantifying long-term trends of POA (nor other components) due to the high sensitivity of OA composition to meteorological anomalies, the occurrence of which is likely not normally distributed over the eight year measurement period.

Due to the unique and realistic seasonal cycles and meteorology-dependences of the independent OA subtypes complemented by the reasonably low degree of unexplained OA variability, we believe that the presented data analysis approach performs well. Therefore, we hope that these results encourage also other researchers possessing several-year-long time series of similar data to tackle the data analysis via similar semi- or unsupervised machine learning approaches. This way the presented method could be further optimised and its usability explored and evaluated also in other environments.

1 Introduction

Despite the small sizes of atmospheric aerosol particles, they play an important role in the climate system. They interfere with solar radiation via direct absorption and scattering (direct aerosol radiative effect) and participate in cloud formation and processing thereby influencing the interactions between clouds and radiation (indirect aerosol radiative effect). In addition to the size of aerosol particles, their chemical composition plays an important role determining their direct or indirect radiative effects via composition-linked parameters such as aerosol hygroscopicity (water affinity), volatility and reflectivity.

The number concentrations of aerosol particles in the atmosphere range from a few particles per cubic centimetre to even millions, so they cannot be considered individually, but are typically divided into populations, groups or classes based on e.g.

some above mentioned characteristics. Thus, the classification of aerosol particles is a necessary and critical task preceding their further understanding. Real aerosol populations are spatially mixed, overlapping and smeared in the atmosphere and their physical and chemical characteristics are for the most part not discretely distributed but continuous. Therefore, practically all classifications of atmospheric aerosol are simplifications due to their complex interactions and change processes in the atmosphere, and any divisions between classes are to some extent arbitrary and debatable selections. Nevertheless, various statistical methods can be used to perform objective, well founded aerosol classifications, and construct aerosol models which strike a good balance between mathematical robustness, complexity (or simplicity) and usability for various purposes. In the following, some common classifications are discussed.

Organic aerosol (OA) is a major sub-micrometre aerosol constituent (Zhang et al., 2007). OA can be emitted directly as primary OA (POA) or it can form in the atmosphere via condensation or uptake of oxidized organic vapours. The latter OA fraction is termed as secondary organic aerosol (SOA). Various combustion processes are the main sources of POA. These combustion processes include for example diesel combustion in car engines, which emits hydrocarbon-like OA (HOA), or biomass burning in forms of residential heating or wild/agricultural fires, both of which emit biomass burning OA (BBOA). The number of SOA precursors in the ambient air is immense making the linking of ambient SOA observations to SOA precursors and detailed formation processes extremely challenging.

The utilization of Positive Matrix Factorization (PMF, Sect. 4.1) on OA mass spectra recorded by Aerosol Mass Spectrometers (AMS; Aerodyne Research Inc., MA, USA; Canagaratna et al., 2007) has linked SOA to two oxygenated organic aerosol (OOA) groups characterized by volatility: semi-volatile oxygenated OA i.e. SV-OOA, and low-volatility oxygenated OA i.e. LV-OOA. These groups are alternatively also named by their degree of oxygenation: less-oxygenated OA i.e. LO-OOA and more-oxygenated OA i.e. MO-OOA. In reality, atmospheric oxidation of aerosols is a continuum process and therefore such a division is mathematical, not clear cut and to some extent arbitrary. Due to the prominent link between OA degree of oxygenation and volatility, the SV-OOA and LO-OOA, and the LV-OOA and MO-OOA usually describe the same OA fractions, respectively (Jimenez et al., 2009;Ng et al., 2011a). LV-OOA is typically identified by an AMS OA mass spectrum dominated by a CO_2^+ (at *m*/*Q* 44 Th in LV-OOA mass spectrum) OA fragment (Jimenez et al., 2009;Ng et al., 2010). SV-OOA in turn typically has lower CO_2^+ mass fraction, but a high $C_2H_3O^+$ (at *m*/*Q* 43 Th in the SV-OOA mass spectrum) fragment (Jimenez et al., 2009;Ng et al., 2010). The CO_2^+ fragment has been linked to various organic acids (Duplissy et al., 2011), whereas the $C_2H_3O^+$ has been thought as a marker of non-acid oxygenates (Ng et al., 2011a). Importantly, a large amount of evidence suggests that photochemical aging of OA leads to an increasingly significant contribution of CO_2^+ in the OA mass spectrum (Alfarra, 2004;De Gouw et al., 2005;Aiken et al., 2008;Kleinman et al., 2008;Jimenez et al., 2009;Ng et al., 2010;Ng et al., 2011a). This indicates OA transformation to more oxygenated forms upon atmospheric aging, which ultimately yields OA of low volatility. Such OA processing (aging scheme) has shown to apply for several SOA and POA types.

While the direct POA emissions can nowadays often be quite well distinguished from SOA, perhaps due to the limitations in chemical information provided by AMS-type instruments and/or the overall similarity of SOA mass spectra regardless of the source, ambient SOA source apportionment is rarely successfully conducted. Source apportionment is also generally difficult due to complexity of atmospheric aerosol chemistry, meteorological and atmospheric transport processes and inherent methodological (both experimental and data analytical) limitations. However, SOA formation from various precursors has been a topic of numerous laboratory studies giving insights into the most dominant ambient SOA formation pathways. Biogenic volatile organic compounds (BVOC) have shown to have a high SOA formation potential upon oxidation (Hallquist et al., 2009). Although the number of different organic species in the atmosphere is enormous (10^4 – 10^5) (Goldstein and Galbally, 2007), isoprene and monoterpenes clearly distinguish themselves as the most emitted biogenic VOC (Guenther et al., 2012). While isoprene-derived SOA formation is hampered by the relatively high volatility distribution of isoprene oxidation products (Hallquist et al., 2009;Surratt et al., 2010;Shrivastava et al., 2017), monoterpenes stand out as one of the major biogenic SOA precursors, due to the production of readily condensable vapours upon oxidation (Donahue et al., 2011;Ehn et al., 2014). The boreal biome, which represents ~15% of the Earth's terrestrial area making up ~ 30% of the world's forests (Prăvălie, 2018), serves an example of a region with relatively high monoterpene emissions (Guenther et al., 2012;Rinne et al., 2009). Measurements from the boreal forests also provide evidence of high content of naturally produced biogenic SOA (Tunved et al., 2006;Yttri et al., 2011).

The current study is targeted on the analysis of OA composition at the well-established Station for Measuring Ecosystem Atmosphere Relations (SMEAR II; Sect. 2.1) located in the monoterpene-rich boreal forest of Finland. What makes this station unique is the large amount of long-term measurements conducted at the site. We recently reported the long-term phenomenology of sub-micrometre aerosol chemical composition seasonality at the site (Heikkinen et al., 2020). We reported a high OA mass fraction of the sub-micrometre particulate matter, ranging between 50 and 80%. The current work specifically focuses on this sub-micrometre particulate matter mass fraction with a goal to gain understanding of OA composition and specifically its seasonal variability at SMEAR II, which has never been reported for the site before. The data analysis includes PMF on the OA mass spectra recorded by an Aerosol Chemical Speciation Monitor (ACSM, Sect. 2.2), but due to the near-decade long mass spectral input from a rather remote measurement site, handling the data retrieved via PMF analyses required also the utilization of new analysis tools. Inspired by our previous work regarding statistical analyses of OA mass spectra (Äijälä et al., 2017;Äijälä et al., 2019), we tackled the analysis problem by combining and applying various advanced statistical methods and machine learning tools. After the extensive analyses, we not only report OA composition variability at SMEAR II, but equally highlight the development of the new framework for long-term OA mass spectral analysis.

2 Measurements

This chapter contains a brief description of the boreal SMEAR II measurement site and the ACSM measurements conducted. For a more comprehensive measurement and station meteorology descriptions, we direct the reader to Heikkinen et al. (2020).

2.1 Station Measuring Ecosystem Atmosphere Relations (SMEAR II)

The measurements were conducted at the SMEAR II station described in detail previously (Hari and Kulmala, 2005;Williams et al., 2011;Heikkinen et al., 2020). SMEAR II is well known due to the broad variety of measurements taking place at the station, tracking more than 1000 different environmental parameters within the Earth–atmosphere interface (Hari and Kulmala, 2005). The station is located in Southern Finland (61°51'N, 24°17'E, 181 m above sea level) in a ca. 60-year-old Scots pine (*Pinus sylvestris*) dominated forest. The station, recognized as a rural site, has low anthropogenic emissions, apart from two nearby sawmills situated 6–7 km to southeast from SMEAR II. In case of south-easterly winds, both monoterpene and OA concentration are elevated at SMEAR II (Eerdekens et al., 2009;Liao et al., 2011;Äijälä et al., 2017;Heikkinen et al., 2020). The dominant source of air pollutants at SMEAR II are air masses traveling from industrialized areas in Southern Finland, St. Petersburg (Russia) and continental Europe (Patokoski et al., 2015;Riuttanen et al., 2013;Yttri et al., 2011;Tunved et al., 2006). The surrounding forest emits multiple biogenic non-methane VOCs, dominantly monoterpenes (Hakola et al., 2012;Barreira et al., 2017). Monoterpenes have been recognized to yield condensable vapours at SMEAR II (Yan et al., 2016;Rose et al., 2018;Ehn et al., 2012) known to efficiently form SOA (Ehn et al., 2014).

2.2 Aerosol Chemical Speciation Monitor (ACSM)

The Aerosol Chemical Speciation Monitor (ACSM; Aerodyne Research Inc., USA), described in detail by (Ng et al., 2011c), serves as the key instrument in this study. The ACSM measurements at SMEAR II, together with the data processing techniques, are documented in detail in our earlier work (Heikkinen et al., 2020). Here, we utilize ACSM data recorded between April 2012 and September 2019. The 2019 measurements and data preparation were performed exactly the same way as for the 2012–2018 data (Heikkinen et al., 2020).

The ACSM, which is developed following the same technology as the AMS (Canagaratna et al., 2007), samples ambient air with a flow rate of 1.4 $cm^3\ s^{-1}$ through an aerodynamic lens having ~100% transmission of ca. 75–650 nm particles in vacuum aerodynamic diameter (D_{va}), but further passes through particles up to ca. 1 µm in D_{va} , albeit less efficiently (Liu et al., 2007). The particles are flash vaporized at 600 °C under high vacuum and ionized with 70 eV electron impact ionization. The resulting ions and their fragments are guided to a mass analyser that is a residual gas analyser (RGA) quadrupole, which scans through different mass-to-charge ratios (m/Q). The particulate matter detected by the ACSM is referred to as non–refractory (NR) sub–micron particulate matter (PM_1). The word 'non–refractory' is attributed to the instrument limitation to detect only material

flash evaporating at 600 °C and being unable to reliably measure extremely heat–resistant chemical components such as sea salt and black carbon. The word 'PM_1' is linked to the aerodynamic lens approximate cut-off at 1 µm.

The NR–PM_1 reported from ACSM measurements, is a difference (diff) between the signal of particle-laden air and signal recorded when the sampling flow passed a particle filter (filtered air). In addition to the diff measurement style, which is measured using a chopper instead of a filter in the AMS, the lack of particle sizing and the cheaper detector model are the major differences between the AMS and the ACSM. Indeed, while the AMS utilizes multichannel plate detector (MCP) gaining high signal-to-noise (SNR) ratios, the ACSM employs a secondary electron multiplier (SEM) that provides a longer lifetime at the cost of SNR. To improve the SNR, the ACSM data utilized here was 3-hour averages instead of the original sampling resolution of 30 min.

As explained previously (Heikkinen et al., 2020), the ACSM was measuring through the roof of an air conditioned container. The inlet system contained a $PM_{2.5}$ cyclone, and a 3 Lpm overflow to avoid inlet losses. From summer 2013 onwards, a Nafion drier was included in the sampling line, which kept the sample flow relative humidity (RH) below 30%. The instrument provides the NR-PM_1 chemical species' mass concentration every 30 min. The mass concentration calculations, namely the conversion from amperes to µg m^{-3} were based on ionization efficiencies, routinely calibrated using size selected ammonium sulphate and ammonium nitrate particles and a TSI Condensation Particle Counter (CPC; TSI 3772) as a reference instrument. A final collection efficiency (CE) correction was applied based on a two-month moving median comparison with a collocated differential mobility particle sizer as the commonly used composition-based CE correction (Middlebrook et al., 2012) was not applicable due to ammonium concentration being most of the time below the detection limit. A detailed description of the CE correction is presented previously in Heikkinen et al. (2020). The CE correction was applied to the OA mass spectra prior to the PMF analyses.

3 Openair and time-over-land (TOL) analyses

This chapter provides a brief description of wind and air mass trajectory analyses coupled to the analysis of OA composition at SMEAR II.

3.1 Openair polar plots

Openair polar plots are used in the paper to show how OA composition varied under different wind direction and speed combinations (Openair polar plots using R-based package presented by Carslaw and Ropkins (2012)). The concentration fields were calculated by binning the OA component concentration data into different wind direction and speed bins. The field was then smoothed by interpolation, which was performed between grid centres. These Openair polar plots are drawn utilizing ZeFir pollution tracker (Petit et al., 2017), which is an Igor Pro (Wavemetrics Inc., USA) graphical interface for producing

Openair polar plots (among other functionalities). The wind data used for Openair polar plots was recorded at the SMEAR II mast, above the forest canopy (16.8 to 67.2 m a.g.l.) with Thies 2D Ultrasonic anemometers. The wind roses are presented in Fig. S.1.

3.2 HYSPLIT trajectories and TOL

The time each air mass spent over land before reaching SMEAR II was calculated hourly using 96-hour-long HYSPLIT (Stein et al., 2016) air mass back trajectories, with arrival heights of 100 m above ground level. The model was run with NCEP/GDAS (Kanamitsu, 1989) as the meteorological input, with the 1° horizontal resolution dataset used for years 2012–2013 and the 0.5° resolution dataset for 2014–2018. Trajectories were grouped into three different source regions: clean sector, Europe-sector and Russia-sector (Fig. S.2). A source region criterion resembling our clean sector classification has been used before by e.g. Tunved et al. (2006), with similar calculations on the time spent over land. Europe- and Russia-sectors are considered polluted as mentioned earlier in Sect. 2.1. The grouping criterion was that the trajectory had to spend a minimum of 90% of the time in a sector. This means that all the trajectories grouped into the clean sector have spent minimum 90% of the time in the clean sector before arriving at SMEAR II. If the trajectory did not reach this criterion in any of the sectors, it was discarded, and not considered in any further analyses. Time spent over islands, other than the British Isles, is not considered in the time over land (TOL) value.

4 Statistical methods

This section provides an introduction to the statistical methods utilized in this study. The application of these tools is explained later in section 5. Here, we provide the basics of the main statistical tools utilized: Positive Matrix Factorization (PMF) and its application in aerosol mass spectrometry as well as K-Means clustering.

4.1 Positive Matrix Factorization (PMF) and the Multilinear Engine (ME-2)

Positive Matrix Factorization (PMF) (Paatero and Tapper, 1993;Paatero, 1997) is a widely used algorithm in chemometrics, which helps sorting complex measurement data into factors with altering abundances, with static factor profiles without prior knowledge regarding the factor features. More precisely, PMF approximates the measurement data matrix (**X**) as a linear combination of these constant factor profiles (**F**) and their temporal proportions (**G**), both **F** and **G** containing only non-negative elements ($g_{i,k} \geq 0$, $f_{k,j} \geq 0$). The PMF model iteratively minimizes uncertainty-weighted model residuals (Q) using a least squares algorithm, directing the model solution towards combinations of **F** and **G** best describing **X**. The PMF equation in matrix notation can be written as follows:

$$\mathbf{X}_{m\times n} = \mathbf{G}_{m\times p} \cdot \mathbf{F}_{p\times n} + \mathbf{E}_{m\times n},$$

where **E** equals to the model residual matrix. If written element-wise, this equation becomes:

$$x_{i,j} = \sum_{k=1}^{p} g_{i,k} f_{k,j} + e_{i,j}, \quad (1)$$

Here, the subscript *i* is the time column index, *j* the variable row index, and *k* the factor index in the PMF solution containing *p* factors (*p* defined by user). The following equation for *Q*,

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{e_{i,j}}{\sigma_{i,j}} \right)^2 \quad (2)$$

can then be written as

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{x_{i,j} - \sum_{k=1}^{p} g_{i,k} f_{k,j}}{\sigma_{i,j}} \right)^2, \quad (3)$$

where σ equals the measurement uncertainty.

Importantly, the PMF algorithm is frequently solved in robust mode, in which outliers are dynamically reweighted to prevent the PMF model fits to be pulled towards outliers. The outliers are defined as data cells, where the ratio between the model residual and uncertainty exceeds a user-defined threshold, α, usually set as $\alpha = 4$ (Paatero, 1997). The *Q* values given by the PMF model are calculated using the robust mode.

The reliability of one modelled *Q* minimum is not usually enough. Indeed, sometimes the PMF solutions are representative of only a local *Q* minimum instead of the global *Q* minimum. To avoid interpretations of a PMF solution representing a local *Q* minimum, it is recommended to start PMF from multiple different starting points, e.g. seeds. Increasing the number of seeds, preferably together with random resampling (bootstrap) (Efron, 1979), helps mapping the stability of the PMF solution. In the bootstrapping approach, the different PMF seeds have slightly different input matrices, which contain randomly chosen rows of the original matrix. Bootstrapping is a suitable tool for PMF statistical uncertainty evaluation, if sufficient amounts of resamples are conducted (Norris et al., 2008;Paatero et al., 2014).

Multilinear Engine (ME-2) is a popular PMF solver to reduce rotational ambiguity of PMF. One advantage of it is the possibility to introduce known **F** rows (or **G** columns) to PMF model during model initialization (Paatero and Hopke, 2009). This approach is traditionally conducted in three ways: via techniques named chemical mass balance (CMB), *a*-value, and pulling techniques (Paatero and Hopke, 2009). In CMB (Watson et al., 1984), all of the rows in **F** (i.e. all factor profiles) are known beforehand. It can be considered as a far extreme from the traditional PMF, where none of the factor profiles is known. The *a*-value approach falls somewhere between CMB and PMF. Now, certain elements of **F** or **G** can be constrained to the PMF, and the model output variability from the constraint is given by a scalar, *a*. *a* can be applied to the entire **F** row (or **G** columns), or alternatively to their individual elements. The more constraints and the tighter they are ($a \rightarrow 0$), the closer the *a*-

value approach is to CMB. Indeed, the case of having all p rows of **F** constrained with an a-value of zero equals the CMB method. If pulling equations are introduced to the PMF model, PMF pulls the $f_{j,k}$ (or $g_{i,k}$ in case of **G** pulling) towards a user-defined anchor during the iterative steps.

The evaluation of the appropriate number of factors in the PMF solution (p) can be (for example) estimated by observing the decrease of Q and the ratio between Q and the expected Q ($Q_{\text{exp.}}$, which is the Q normalized by the degrees of freedom of the model solution) (Paatero and Tapper, 1993). The decrease of Q/Q_{exp} as a function of p can be, to some extent, used to understand what the optimal number of factors in the solution could be. While Q/Q_{exp} tends to always drop as a function of p, the optimal p is typically where the Q/Q_{exp} drops stop being significant.

4.1.1 PMF application in aerosol mass spectrometry

The application of PMF was first utilized with the organic aerosol data matrix, obtained via aerosol mass spectrometer (AMS) measurements in 2007 (Lanz et al., 2007), and has since then become a widely used and popular method in OA source apportionment. PMF is conducted so that **F** equals the mass spectral profiles and **G** the time series, usually in μg m^{-3}. A comprehensive overview of AMS PMF studies and methodologies utilized between 2007–2011 has been given previously by (Zhang et al., 2011). (Ulbrich et al., 2009) introduced thorough AMS PMF interpretation guidelines and (Crippa et al., 2014) introduced guidelines for the ME-2 a-value approach. Since 2011, PMF with ME-2 has also been applied successfully to ACSM data (e.g. (Fröhlich et al., 2015;Canonaco et al., 2013;Zhang et al., 2019)).

Preparation of the PMF input (organic aerosol data matrix and a corresponding error matrix) for both AMS and ACSM data can be done with their data processing software. The preparations are based on PMF Evaluation Tool (PET) Wavemetrics Igor Pro functions (Ulbrich et al., 2009). Before initializing any PMF solver (such as the ME-2), certain preparations are often necessary for optimal modelling. The m/Q with low SNR (i.e. m/Q having more noise than signal) are down weighted by increasing their error. (Paatero and Hopke, 2003) suggested that m/Q having SNR<0.2 should be down weighted heavily or removed from the analysis, and m/Q with 0.2<SNR<2 down weighted by a factor of 2–3. Another noisy data down weighting approach was suggested by (Visser et al., 2015), where the errors are down weighted continuously with a penalty function SNR^{-1}, when SNR<1. These down weightings have been done either based on the average SNR across the data set or cell-wise. Another data input modification prior to PMF initialization, should be performed regarding CO_2^+ (m/Q 44 Th) -related variables (i.e. m/Q 16–20 Th and 28 Th) because the information stored at these m/Q are directly estimated from m/Q 44 Th. Such high correlation between these variables would be considered in the PMF modelling with too high importance. To avoid this, CO_2^+ -related variables are typically excluded or down weighted accordingly.

PMF analysis has become easily accessible for the whole AMS/ACSM community upon the development of Igor Pro (Wavemetrics inc, USA) based user friendly PMF analysis tools, such as the Source Finder (SoFi, Paul Scherrer Institute and

Datalystica Ltd., Switzerland) (Canonaco et al., 2013;Canonaco et al., 2021) and PET (Ulbrich et al., 2009). Recently, after the launch of the commercial SoFi Pro software (Datalystica Ltd., Switzerland) (Canonaco et al., 2021) also many advanced PMF methods, became available. These methods include rolling PMF (Paatero and Tapper, 1994;Parworth et al., 2015) and PMF resampling (bootstrap).

The assumption of static factor profiles serves one of the questions of the atmospheric representativeness of the PMF output. A rolling PMF approach was suggested (Parworth et al., 2015) to account for such factor profile temporal variability. In the rolling PMF approach, a PMF run is conducted a short time window at a time (the time scale for which the static factor profile is assumed valid). This time window is shifted across the data set in even smaller time steps creating overlap between PMF windows. In practise this means choosing an n day time window in an m day data set ($n << m$), and shifting the window q days at a time ($q < n$) chronologically along the time axis, until all the m days are covered.

As the rolling PMF approach results in a large amount of PMF runs, and the amount grows even larger in case of incorporating bootstrapping (typically 100–1000 seeds per PMF window), manual investigation and conclusion-making becomes very challenging. The challenge of sorting as well as accepting good rolling PMF runs and/or rejecting unrealistic rolling PMF runs is addressed in SoFi Pro via criteria-based selection of PMF runs (Canonaco et al., 2021). The user-defined criteria, best describing each PMF factor (for example correlation between NO_x and HOA, which both are emitted from traffic), are evaluated for each PMF run, and their scores (for example the Pearson correlation coefficient R between NO_x and HOA) are presented. The user can then select all the PMF runs above certain thresholds (for example R>0.5), or select all of the PMF runs. Such criteria-based selection of PMF runs was first introduced by Daellenbach et al. (2017) and Visser et al. (2019). Selection and averaging all of the PMF runs without criteria-based sorting would work only in the case of having all, or all but one, factor constrained. In the case of having two or more free PMF factors, it is likely that their positions in the PMF output matrices are frequently changing, i.e. being situated in different columns in **G**. In the case of constrained PMF factors, they will always appear in their pre-designated **G** columns.

4.2 K-Means clustering

K-Means (Ball and Hall, 1965;MacQueen, 1967;Steinhaus, 1956;Jain, 2010) is the most popular unsupervised machine learning approach utilized in data classification. It works particularly well (computationally efficient) for large data sets with a small number of well-definable clusters (k). The K-Means algorithm works as follows:

1. Picking k number of centroids (i.e. cluster centre points), and assigning each sample (for example a mass spectrum) to its nearest centroid based on a selected distance metric, usually the squared Euclidean distance. This step is nowadays performed following the (Arthur and Vassilvitskii, 2007) K-Means++ algorithm, proven to not only speed up the clustering process, but also significantly improve its accuracy.
2. Moving the centroids to represent the new mean of the cluster.

3. Reassigning the all the points to their closest centroids (this sometimes moves points from one cluster to another).
4. Repeating steps 2 and 3 until convergence is achieved (i.e. data points stop moving between clusters and the centroids stabilize).

The goal of the K-Means clustering algorithm is to minimize the following objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i - c_j \right\|^2, \tag{4}$$

where k is the number of clusters, n the number of data points, x_i the i^{th} data point and c_j the centroid of cluster j, and $\left\| x_i - c_j \right\|^2$ represents the Euclidean squared distance function. Hence, this makes the object function, J, the average squared Euclidean distance between points in the same cluster.

4.2.1 Silhouette score

Silhouette score (Rousseeuw, 1987) is one of the many metrics available for evaluating the number of clusters present in the data set. It is calculated both based on intra-cluster distances of data points (cohesion, a) and their distances to points assigned in other clusters (separation, b). The silhouette score for the i^{th} sample can be expressed as:

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \ . \tag{5}$$

The silhouette scores range between [-1, 1]. The scores for the i^{th} sample can be interpreted as follows:

1. $s_i = -1$; The sample is (likely) assigned to a wrong cluster,
2. $s_i = 0$; The sample is at the decision boundary between clusters,
3. $s_i = 1$; The sample is well clustered.

The silhouette score (s_i) is calculated for an individual sample in Eq. 5, but can also be defined for clusters ($\bar{s}$) as the average over all silhouette scores of samples belonging to the cluster, or for the entire solution (average over all samples), yielding diagnostic information on point, cluster and solution level. (Kaufman and Rousseeuw, 2009) further suggested an average cluster silhouette $\bar{s} = 0.25$ as a lower limit for weak structure and $\bar{s} = 0.50$ as a lower limit for strong cluster structures. Strong structures indicate of a good clustering result, where the samples in the cluster are very similar to each other while being very different from the samples assigned to other clusters.

5 The application of PMF and K-Means in the current study

The current study focuses on conducting rolling PMF on 8 years of OA data recorded by an ACSM at the SMEAR II station. First, we performed unconstrained rolling PMF runs. We used these runs to determine the common OA factor profiles through K-Means clustering. The ultimate goal of the unconstrained PMF and K-Means clustering was to provide mass spectral profiles as *a priori* input for a PMF run in which all of the profiles are constrained with reported intra-cluster mass spectral variabilities. This PMF approach is therefore termed as rolling relaxed CMB, i.e. rolling rCMB. This section contains a detailed description of this framework. A written overview of the method is given below and the work flow is summarized Fig. 1.

5.1 Rolling PMF

The initial rolling PMF was conducted using the 2012–2019 ACSM data (Fig. 2a), prepared with the ACSM data processing software, i.e. the Wavemetrics (USA) Igor Pro-based ACSM Local 1.6.0.3 toolkit, as PMF input. No down weighting, based on low SNR or relation to CO_2^+ was conducted with the ACSM Local software. The data matrices were imported to an Igor experiment with the SoFi Pro (6.A1) toolkit, and averaged from the initial half an hour time resolution to three-hour time resolution in order to improve the SNR. The error propagation was accounted for during averaging (linear terms of the squared Taylor series expansion on the measurement data). Upon the initialization of the PMF matrices, all the CO_2^+-related variables (i.e. *m/Q* 16, 17, 18 and 28 Th) were excluded from the analysis. Then, the errors of the noisy variables (SNR<1) were weighted cell-wise by SNR^{-1}.

Only the *m*/*Q* range of 12-100 Th was included in the rolling PMF. This mass range has been typically chosen for the ACSM PMF analysis, and it avoids introducing the ACSM internal standard, naphthalene at *m*/*Q* 128, to the PMF run. *m*/*Q* 29, 31 and 38 Th were excluded from the analysis due to unknown interferences, likely from air and instrumental issues time to time affecting these signals, and yielding mass spectra not resembling any known aerosol type.

The rolling PMF was initialized with a constant factor number of three. The decision was made based on several (standard, i.e. rolling mechanism disabled) PMF runs, having time series lengths ranging from few months to years. Three factors were considered as an upper limit of the number of factors, as a greater number would not significantly reduce Q/Q_{exp} nor produce meaningful factor profiles. This step required a subjective decision.

The rolling window width was set to 30 days with 10 days window shifts. Previous studies conducted by Parworth et al. (2015) and Canonaco et al. (2021) set the window width to two weeks and the shift to one day, which is much shorter than selected here. However, as shown by Canonaco et al. (2021) the PMF solutions were seemingly equally good for window widths higher than two weeks (tests up to window width of 28 days). Only window widths shorter than two weeks led to a less good PMF result. As the time span of our data was nearly eight times greater than utilized in the previous studies, we speeded up the PMF

modelling process by choosing a longer window width and shift. More testing could be conducted on appropriate lengths. However, if the number of PMF runs were to increase significantly from the amount performed here, it would be feasible to perform the PMF modelling on a server. With the current settings, the rolling PMF run performed in this study using a PC, lasted 48 hours.

Finally, also, the bootstrap mechanism (resampling) was enabled, and a hundred iterations were conducted at each window. A subset of the rolling PMF input is visualized in Fig. 2b. The rolling PMF yielded 62700 factor profiles (20 900 three factor solutions) and time series, respectively, distributed in 209 PMF windows.

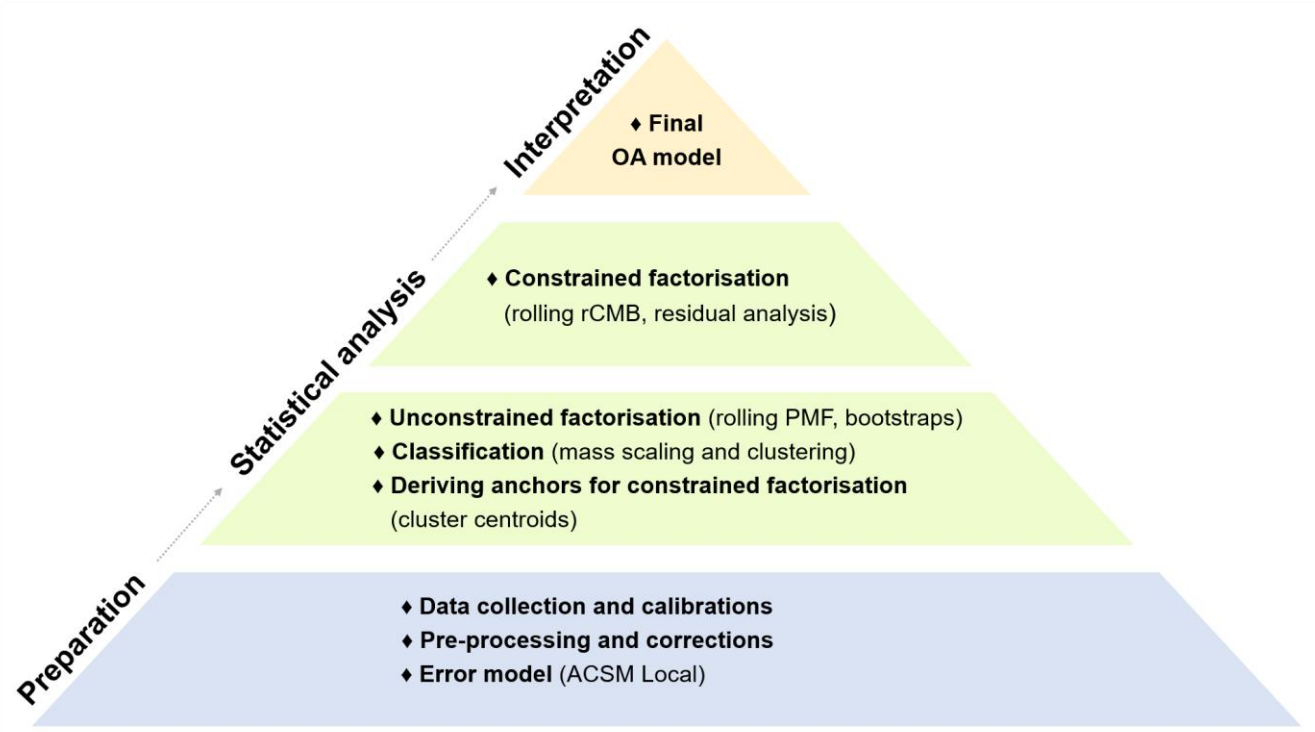

Figure 1 A pyramid flow chart, roughly describing the steps from data collection to the final OA model (i.e. the time series of OA sub-species making up the total OA signal). The statistical analysis steps (in green) are explained further in detail in sections 4 and 5 in the paper as well as listed in Appendix A.

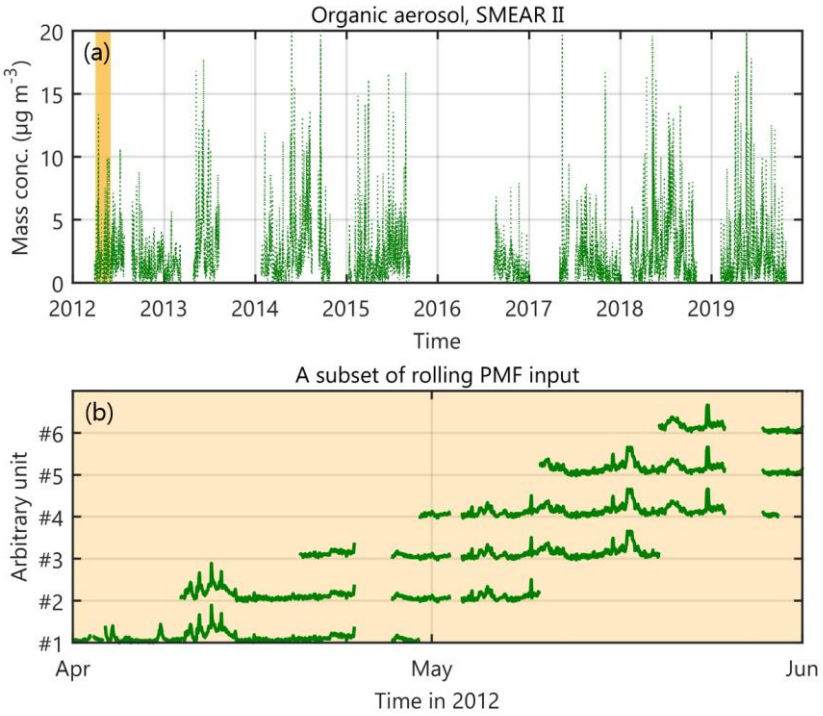

Figure 2 (a) The 3-hour averaged time series of OA measured at SMEAR II and utilized in the current study. The y-axis represents OA mass concentration in μg m^{-3} and the x-axis the time. The figure also depicts the data coverage within the eight years. The yellow shaded region represents the first two months of measurement data, which are further shown in panel b. (b) Schematic figure visualizing the rolling window approach. Now, *x*-axis spans from April 1st to June 1st, 2012 and the six OA time series represent the timespans of successive rolling PMF windows. With the settings used in the current study, this two-month period would be part of six rolling PMF windows.

5.2 K-Means clustering PMF profiles

Selecting and sorting the rolling PMF output via various criteria into three factors would have required a significant understanding of the PMF output beforehand. Choosing solid criteria can be straightforward near known pollution sources, but in case of multiple unknown factors and distant sources such becomes complicated. SMEAR II represents a station with minimal anthropogenic sources. To exemplify the challenges in correlation-based criteria at SMEAR II, we can take the correlation between NO_x and HOA as an example. Both of these species are emitted from traffic and known to correlate well near traffic sources. However, in the case of transported traffic emissions, many things can affect the life time of the emitted species, which affects the correlation between the emissions at SMEAR II. If we pick the effect of wet deposition as an example, it will remove the particulate HOA much more efficiently than gaseous NO_x. If HOA and BBOA were constrained within a SMEAR II OA PMF run, it would not be surprising that the PMF output would suggest that 10% of the OA mass was made up of HOA and 30% of BBOA. As shown later on in this paper, these number are highly unrealistic. Due to the difficulty in interpreting correlations between HOA and BBOA and their markers, correlation analyses do not directly answer when constraining HOA or BBOA would have been appropriate. This is why traditional rolling PMF techniques would prevent us

from HOA/BBOA quantification. This complexity motivated us to 1. use mass spectral clustering to explore the types of OA resolved within the unconstrained rolling PMF runs (i.e. answering when HOA/BBOA were present) and 2. performing rolling rCMB (Sect. 5.3) to explore the temporal behaviour of these OA types. The clustering-based exploration of the unconstrained PMF profiles was conducted PMF window-by-window across various bootstrapped PMF iterations (Phase I; See detailed description in Sec. 5.2.1). This step was followed by exploring the number of clusters across all PMF windows by further clustering all the Phase I cluster centroids (Phase II; See detailed description in Sec. 5.2.2). All the clustering procedures conducted in this study were performed within MATLAB 2017a using the *kmeans* algorithm, which utilizes K-Means++. K-Means was selected as the clustering algorithm due to previous successful OA mass spectral classification performed by Äijälä et al. (2017, 2019). Future work could be conducted in exploring the potential of other clustering algorithms.

5.2.1 Solutions for rolling windows (K-Means clustering Phase I)

The rolling PMF output was uploaded into MATLAB from Hierarchical Data Format (HDF) -files created for each PMF window, respectively, during the ME-2 modelling process. Prior to clustering, we scaled the PMF output with the following function suggested by (Stein and Scott, 1994):

$$\text{weight}_{\frac{m}{Q}} = \left(\frac{m}{Q}\right)^{s_m}, \quad (6)$$

where m/Q equals the mass-to-charge ratio ranging from 12–100 Th, and s_m = 1.36 (recommendation by Äijälä et al., 2017). We previously showed the information value gains of mass scaling in conjunction with AMS data (Äijälä et al., 2017). Indeed, if not applied, several OA types could not be classified (Äijälä et al., 2017). Following Eq (6), each signal at each m/Q was multiplied by its m/Q-corresponding weight-value. As recommended by Äijälä et al., 2017, the usage of this scaling factor gives gradually more weight to the patterns at the end of mass spectrum, containing a lot of information regarding OA sources.

Importantly, the following clustering of bootstrap iterations one rolling window at a time was conducted using cosine (dis)similarity (Sokal and Sneath, 1963) as the K-Means distance metric as opposed to the commonly used squared Euclidean distance. This decision was again based on our earlier work in which various K-Means distance metric alternatives were explored, and best classification outcomes (i.e. highest number of mathematically well-structured clusters, the centroids of which resembled well-known OA types found in the literature) resulted from clustering efforts utilizing cosine angles along with correlations (Äijälä et al., 2017). While nearly equally good clustering outcomes were achieved between these two metrics, we decided to report the cosine (dis)similarity results due to the popularity of cosine angles in mass spectral comparisons (Stein and Scott, 1994). Cosine (dis)similarity describes the similarity between two n-dimensional (n, i.e. the number of m/Q, which was 70 in our study) vectors (**A** and **B** in the equation below) via the cosine of the angle between them. Hence, the metric is not magnitude but orientation dependent. In our case this also meant that normalization of the weighted mass spectra was not necessary. The cosine (dis)similarity is defined as follows:

$$\text{Cosine (dis)similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}, \quad (7)$$

where **A** and **B** are n-dimensional vectors, which in the current case would correspond to two mass spectra.

Silhouette values were utilized to evaluate the clustering outcome similarly to Äijälä et al. (2017). Other metrics were not tested within this work as they would operate only by using squared Euclidean distance measures within our analysis software, MATLAB 2017a.

Finally, the PMF window-by-window clustering of bootstrap iterations was conducted as follows:

1. Clustering (MATLAB 2017a *kmeans* function using cosine (dis)similarity as the distance metric) and calculating mean silhouette values (MATLAB 2017a *silhouette* function using cosine (dis)similarity as the distance metric) for 2–4 clusters per PMF window. This step was performed using the 300 mass scaled (Eq. 6) mass spectral profiles (3 factor profiles, 100 iterations) given by the 30-day rolling window.
2. Finding the number of clusters achieving the highest mean silhouette value in the PMF window. Only this clustering result was used in the following steps as it was considered as the "best solution".
3. Un-doing the mass scaling and calculating silhouette-weighted cluster centroids (here: the median of all mass spectra belonging to the cluster, each multiplied by their spectra-specific silhouettes) for each PMF window. The weighting of the cluster centroid calculation by silhouette scores was performed similarly to Äijälä et al. (2017, 2019) studies: all mass spectra possessing a negative silhouette score were discarded from the cluster centroid calculation and the rest of the mass spectra were multiplied by their spectra-corresponding silhouettes. This way, the spectra with the highest silhouette scores would influence the cluster centroid the most, and the spectra with the lowest silhouette score were either discarded (if silhouette score is zero or negative) or having minimal weight on the final cluster centroid. This step helps to alleviate possible K-Means susceptibility to outliers in clusters.
4. Appending the silhouette-weighted cluster centroids in a matrix ($\mathbf{F_I}$). If the PMF window was clustered with three factors in the 3rd step listed here, then $\mathbf{F_I}$ would gain three new rows: one for each cluster centroid mass spectrum.
5. Moving to the next PMF window and repeating steps 1-6 until all PMF window are clustered and matrix $\mathbf{F_I}$ contains all the silhouette-weighted centroids from each PMF window.

All the steps presented above, were done programmatically in MATLAB. The final number of mass spectra stored in $\mathbf{F_I}$ was 479. The overall mean silhouette values for 2-4 clusters were high, strongly indicating segregation of strong cluster structures in the PMF window-by-window clustering of bootstrap iterations (Fig. 3a). The optimal number of clusters in the PMF windows was 2 in ca. 80% of the PMF windows (Fig. 3b), which meant that only ca. 20% of the PMF windows contained 3–4 different resolvable PMF factors.

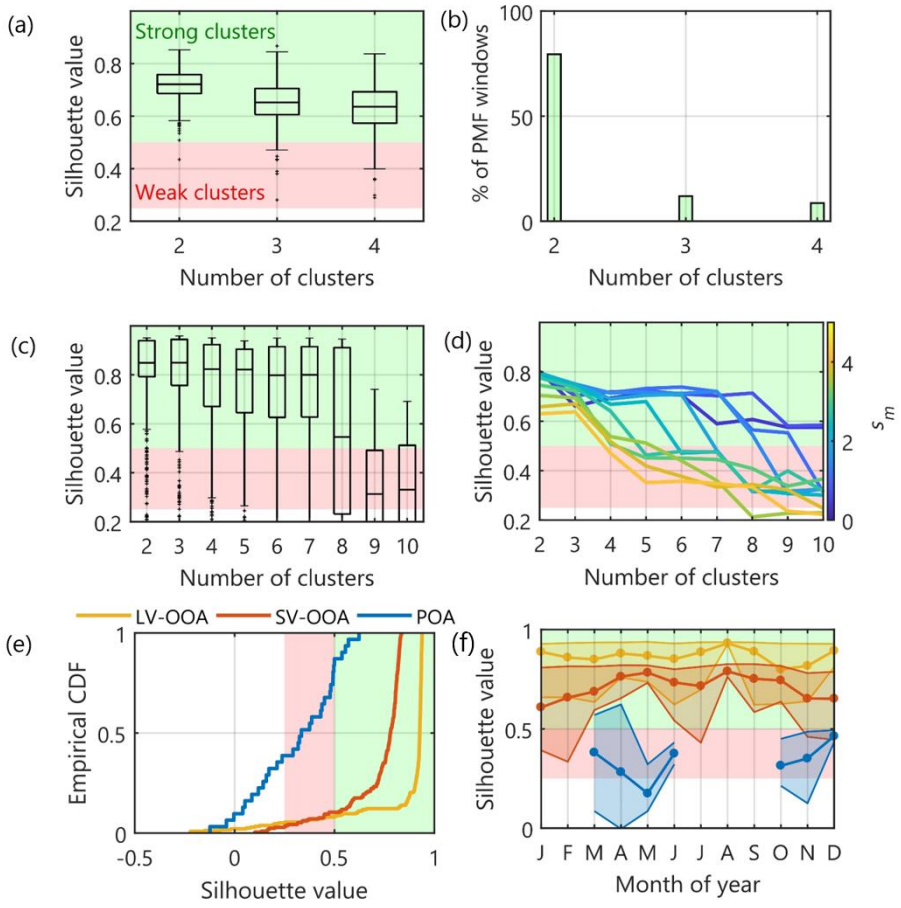

Figure 3 (a) Box-and-whisker diagram displaying the silhouette score distribution for *k* (number of factors) = [2, 4] representing all 209 PMF windows (Phase I). The green and red shadings indicate the ranges of strong and weak cluster structures, respectively. (b) Fraction of PMF windows achieving the highest silhouette score when the number of clusters (*k*) was 2, 3 or 4. (c) Silhouette score distribution for *k* = [2, 10] for Phase II (i.e. clustering the 479 profiles obtained from the 209 PMF windows in Phase I). (d) Evolution of the median silhouettes in *k*-space as a function of the mass scaling (Eq. 6) factor, s_m, which gives dynamically more weight to the end of the mass spectrum. The colour scale presents the s_m value for each line. (e) Cumulative distribution function (CDF) of the *k* = 3 Phase II silhouette scores for the three clusters (named LV-OOA, SV-OOA and POA), respectively. This subplot shows that POA has the weakest cluster structure, and LV-OOA the strongest. (f)Temporal behaviour of the median silhouette score of each cluster in the *k* = 3 Phase II solution. Here, each month displayed must contain a minimum of 30 days of cluster appearance, explaining the gap in the POA seasonal cycle, as it is not as frequently resolved as the other clusters.

5.2.2 Overall classification of mass spectra (K-Means clustering Phase II)

The next step was to explore the dominant mass spectral clusters in the whole data set. Phase II contained the following steps:

1. Performing mass scaling (Eq. 6) for $\mathbf{F_I}$ mass spectra, as performed earlier in the PMF window-by-window clustering of bootstrap iterations (Phase I; Sec. 5.2.1).
2. Calculating mean silhouette scores (MATLAB 2017a *silhouette* function using cosine (dis)similarity as the distance metric) for 2–10 clusters.
3. Exploring how many clusters are needed to gain the highest mean silhouette score. In case of a vague difference between silhouettes (as shown in Fig. 3c), the step is followed by performing steps 1 and 2 again with different mass scaling s_m-values. The optimal number of clusters should preserve the high silhouette score even at high s_m-values. We explored k = [3, 6] solution space with different s_m-values (s_m = [0, 5]). By increasing s_m, the silhouette value for k = 3 increased to the same level as k = 2, while k > 3 solution silhouettes decreased below the strong cluster limit (Fig. 3d). We thus selected three clusters for the following steps.
4. Clustering (MATLAB 2017a *kmeans* function using cosine (dis)similarity as the distance metric) the mass weighted mass spectra (s_m = 1.36) with the number of clusters defined in the previous step.
5. Un-doing the mass scaling, and calculating silhouette-weighted, normalized cluster centroids (cluster median) and the cluster mass spectral variability (lower and higher quartiles). These cluster centroids represent the prevailing OA types in SMEAR II sub-micrometre aerosol.

The three different OA clusters found by this method were named low-volatility oxygenated organic aerosol (LV-OOA), semi-volatile oxygenated organic aerosol (SV-OOA) and primary organic aerosol (POA). The LV-OOA and SV-OOA clusters had generally high silhouette scores whereas the POA cluster had a weaker structure (Fig. 3e). More discussion on the mass spectral features is provided in the results section (Sect. 6.1).

5.3 Rolling rCMB

After gaining the prevailing OA types mass spectral features via the above explained clustering processes, we wanted to gain understanding of the temporal features and mass loading of each OA type. As the HDF-files for each rolling PMF window also contain time series information for each factor profile, we were able to calculate cluster-specific time series utilizing these time series connected to each cluster member spectra. The time series of the OA types were discontinuous since factors were not resolved in every window. Therefore, we utilized the silhouette-weighted cluster interquartile ranges (IQRs) gained in Sect. 5.2.2. to constrain a rolling rCMB run to gain continuous time series for each OA type. These cluster-specific time series extracted from the initial PMF were afterwards used to evaluate the rolling rCMB run (Sect. 5.3.1), but also enabled us to explore the silhouette score temporal behaviour. The silhouette score monthly medians are visualized in Fig. 3f. Only SV-

OOA showed some seasonality, which could hint that SV-OOA composition has some, yet little, inter-annual variability. Due to the stability in the monthly median silhouettes, we consider the mass spectral classification robust.

The rolling rCMB run was conducted via rolling PMF using the cluster centroids of the OA factor profiles as *a priori* information. After extracting the governing mass spectral features across the data set, we exported the silhouette weighted and normalized mass spectra to SoFi Pro 6B. We set up a PMF run with three factors, all of them constrained with our silhouette-weighted cluster centroids (median factor profiles). However, differing from the traditional CMB approach, we passed ME-2 the allowed limits within which the factor profiles should vary. These limits were the 25th percentile (lower limit) and 75th percentile (higher limit) of the silhouette-weighted cluster centroid spectra. The rolling rCMB was otherwise initialized exactly like the initial rolling PMF run. The CO_2^+ related variables were excluded, and the errors of the weak variables were treated similarly (cell-wise SNR^{-1} penalty function). The rolling window length was again 30 days with a 10 day shift, and resampling was enabled with 100 seeds. *m*/*Q* 29, 31 and 38 Th were still discarded from the analysis. The final rolling rCMB results for each factor, respectively, were obtained by averaging over the 20 900 PMF runs for each time point (in total: 3 × 20 900 = 62 700 factor profiles and time series). As all the factor positions in rolling rCMB were fixed (LV-OOA profile was constrained at the **F** matrix first row, SV-OOA at the second and POA at the third), such averaging was appropriate.

5.3.1 Rolling rCMB residual analysis and output evaluation

To evaluate the averaged rolling rCMB output, we first compared the Q/Qexp values between the initial rolling PMF and rolling rCMB. The comparison of the Q/Qexp retrieved from each iteration in each rolling window is visualized in Fig. S.3. As expected, the mean rolling rCMB Q/Qexp value was higher (38% increase) than that of the initial rolling PMF Q/Qexp. This is typical as Q/Qexp tends to increase whenever constraints are added to the PMF run. However due to the relaxed approach, the Q/Qexp increase is for example much less dramatic than shown in Canonaco et al. (2013) CMB tests. We find the observed Q/Qexp increase acceptable, considering the higher information value (interpretability) provided by the rCMB solution.

To continue the rolling rCMB result evaluation via residuals, we investigated rolling rCMB model uncertainty-scaled residuals (**R** matrix, $r_{i,j}$ in cell notation in Eq. (8)). **R** elements were calculated with SoFi Pro using the following equation:

$$r_{i,j} = \frac{e_{i,j}}{\sigma_{i,j}}, \tag{8}$$

where σ_{ij} indicates the measurement error provided in the initial PMF input error matrix and e_{ij} the model residual (i.e. the difference between model input and model output: x_{ij} (meas.) – x_{ij} (mod.)). A normalized scaled residual histogram is presented in Fig. 4. The scaled residual histogram, presented in figure 4 in green, is fairly unimodal and spreads between [-4, 4] (most data between [-3, 3]) as desired (Paatero and Hopke, 2003), but tends to have high frequency of slightly negative, near-zero

readings. We connected this behaviour to periods with high SNR (i.e. summers; Fig, S.4). As downweighting of the noisy and weak variables made as a function of SNR-1 (Sect. 4.1.1 and 5.1) which further influences σij in Eq. (8), the seasonality in SNR was seemingly driving the scaled residual seasonal cycles. This was visible, yet to a lesser extent, in a test rCMB run conducted without downweighting (Fig. 4 in purple) and with a more traditional average step wise downweighting procedure (not shown), which further brings us to the conclusion that the PMF input matrix errors are also SNR-dependent (Ulbrich et al., 2009.) and could perhaps be further optimised. However, it should be kept in mind that the scaled residuals in general speak for a good performance of rolling rCMB in modelling the input data, and the scaled residual time series shown in Fig. S.5 reveal no evident patterns/trends except the negative values in summers. An additional investigation into the real unexplained variation within the data (shown later on in Fig. 8) revealed no correlations with temperature or sub-micrometre PM components.

Annual median scaled residual mass spectra are visualised in Fig. S.5. Even clustering attempts on the scaled residual matrix do not reveal clear structures in the scaled residual matrix although an overall median scaled residual mass spectrum calculated using the negative residuals alone would hint towards some resemblance with POA at m/Q > 50 Th. We note that this could indicate minor POA overestimation in the rolling rCMB and speculate whether introducing time-dependent profile variation limits to ME-2 could help us overcome the issue. With the method presented here, we could easily extract time-dependent limits for ME-2 variability. However, introducing such to dynamic approach to the ME-2/SoFi Pro analysis software is not yet possible.

The comparisons between rolling rCMB time series (Fig. S.6) to the cluster-specific time series serve as the final step in rolling rCMB validation. The overall Pearson correlation coefficient between the mean-cluster time series and the sum of rolling rCMB factor time series is approaching unity (R = 0.99), and the correlations between different OA classes are 0.97, 0.94 and 0.78 for LV-OOA, SV-OOA and POA, respectively (Fig. S.6). In fact, such high degree of agreement indicates very good rolling rCMB performance in retrieving time series for the different OA classes. As a final note, as discussed previously, the POA appearance in the time series retrieved after the Phase II clustering was likely depending on the POA mass fraction in different PMF windows. We evaluated that 95% (3σ) of the PMF windows where POA was not classified, had a POA mass fraction (i.e. the mass fraction of POA in relation to the total rolling rCMB OA mass; f_{POA}) of 6% (Fig. S.7a), when POA explained variation (i.e. rolling rCMB-derived variability explained by POA compared to the total measurement variability) was 7% (Fig. S.7b). Such numbers resemble the PMF "rule-of-thumb" detection limit of ca. 5% estimated by Ulbrich et al. (2009). This final note indicates simply that the POA cluster was not found when the POA concentration was near-zero in rolling rCMB. Such behaviour is certainly a factor explaining the slopes between the cluster-specific time series and rolling rCMB time series presented in Fig. S.6.

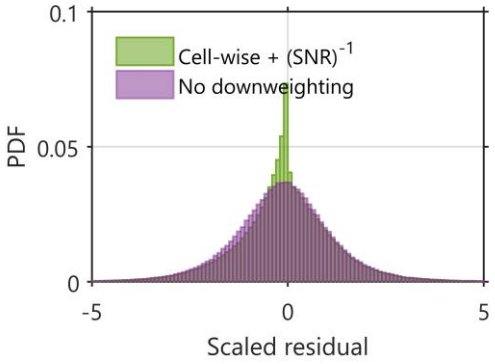

Figure 4 Normalized histograms (probability density function, PDF) of the scaled residuals obtained from rolling rCMB. The effect of downweighting weak/bad variables is visible by the high scaled residual frequencies at negative near-zero readings. If rolling rCMB was conducted without downweigting the scaled residual distribution behaves in a highly normal manner.

6 Results and discussion

In this section, we introduce the key features of the LV-OOA, SV-OOA and POA clusters' mass spectra (Sec. 6.1). After the detailed mass spectral investigation, which explains the naming of each cluster, we further discuss the temporal behaviour of these OA classes (data retrieved via rolling rCMB; Sec. 6.2). The section then includes a brief analysis of wind direction and speed dependences of the OA classes (Sec. 6.3.1) via Openair polar plots (Carslaw and Ropkins, 2012; Petit et al., 2017). As a final section in this chapter we explore LV-OOA, SV-OOA and POA loading as a function of time over land in the clean sector (Sec. 6.3.2) to yield understanding on natural OOA production over the NW quadrant of Europe.

6.1 Mass spectral features of OA clusters

The cluster centroids resulting from the overall classification of SMEAR II mass spectra serve as one of the key results of the current study (Fig. 5). The three OA classes were named already previously as low-volatility oxygenated organic aerosol (LV-OOA), semi-volatile oxygenated organic aerosol (SV-OOA) and primary organic aerosol (POA), but we start this chapter by motivating the decisions behind each OA cluster name.

The naming of LV-OOA was based on the dominance of m/Q 44 Th in the mass spectrum, and the naming of SV-OOA was done due to the high m/Q 43 Th (higher than m/Q 44 Th). The naming of the POA was motivated based on the resemblance of the POA mass spectrum with both hydrocarbon-like OA (HOA) and biomass-burning OA (BBOA). The cosine (dis)similarities between POA and HOA or BBOA (both references from (Ng et al., 2011b); spectra downloaded from http://cires1.colorado.edu/jimenez-group/AMSsd/, last access June 3rd, 2020; Ulbrich et al., 2009) were 0.85 and 0.80, respectively. If a mass scaling (Eq. 6 with various s_m) was applied to all spectra, the cosine (dis)similarities between POA and HOA and BBOA, respectively, fast exceeded 0.90. This possibly happened, because less weight was given to m/Q 44 (and 43 Th), which is higher in our POA than in typical fresh HOA or BBOA spectra (see for example Ng et al., 2011b) likely meaning

that our POA cluster is more oxidized than fresh POA. As we expect HOA and BBOA to be primary in origin, and our cluster centroid spectrum resembles both of them, we decided to call this OA class POA.

To further motivate our selection of names for the three clusters (as well as to visualize the cluster structures for the readers), we displayed all the different mass spectra belonging to each cluster in an *m*/*Q* 43 Th vs *m*/*Q* 44 Th organic signal contribution space (f_{44} vs f_{43} space; Fig. 6a). Ng et al. (2010) first introduced this projection, also called the 'triangle plot'. This perspective separates well the LV-OOA, SV-OOA and POA clusters. They are placed in each corner of the triangle in Fig. 6a. LV-OOA lies on the top of the triangle, exhibiting the highest OA mass fraction of *m*/*Q* 44 Th (i.e. f_{44}; hereafter this same nomenclature logic is used also for other OA mass fractions of various different *m*/*Q*), whereas SV-OOA and POA lie at the bottom of the graph possessing nearly equally low f_{44}. The f_{43} on the other hand, is highest for SV-OOA, and lowest for POA (nearly equally low as for LV-OOA).

By using a parametrization provided by (Canagaratna et al., 2015), we converted the f_{44} vs f_{43} plot into a hydrogen-to-carbon ratio ($\mathrm{H:C} = 1.12 + 6.74 \times f_{43} - 17.77 \times f_{43}^2$) vs oxygen-to-carbon ratio ($\mathrm{O:C} = 0.079 + 4.31 \times f_{44}$) space (Van Krevelen (VK) diagram (Van Krevelen, 1950); Fig. 6b). The bulk OA data from AMS measurements has been shown to follow a -1 slope on the VK diagram (Heald et al., 2010), where the most fresh OA has the highest H:C and lowest O:C and the aged OA the opposite. The evolution of OA in the VK space following different lines results mainly from OA functionalization. In case of a slope of 0, OA functionalization would occur mostly by addition of alcohol or peroxide groups. In case of a slope of -1, carboxylic acid groups are being added and the slope of -2 would indicate additions of ketone or aldehyde groups. Factorized OA data were previously visualized in the VK diagram by (Ng et al., 2011a), where the slope for OOA data was ca. -0.5. They suggested that ambient OOA aging would result from addition of alcohol and peroxide functional groups without introducing fragmentation and/or the addition of carboxylic acid groups with fragmentation. Here, we visualize only SV-OOA and LV-OOA, as they provide better statistics than POA as number of objects in POA cluster was small, and these points would be highly scattered in the VK diagram. Furthermore, it is also mentioned in Canagaratna et al. (2015) that the parameterization works less well for POA.

Before interpretation of the VK diagram, we revisit results from European ACSM inter-comparisons conducted at Aerosol Chemical Monitor Calibration Center (ACMCC). A large variability within f_{44} was observed between different ACSM units (Crenn et al., 2015;Fröhlich et al., 2015;Freney et al., 2019). Furthermore, the observed f_{44} were systematically higher than the f_{44} measured with a co-located high resolution AMS, which was shown to give consistent O:C for a suite of organic samples with known O:Cs. While the f_{44} variability was not significantly propagated in OA class mass fractions retrieved with PMF analyses of co-located ACSM data sets (Fröhlich et al., 2015), the O:C ratios of different classes were naturally affected (as O:C parameterization for AMS-type instruments is directly f_{44} dependent). The f_{44} variability has been to some extent explained by an AMS/ACSM vaporizer artefact, which leads to a release of CO_2^+ in the presence of high nitrate mass fractions (Pieber

et al., 2016;Freney et al., 2019). Even though the presence of *m*/*Q* 44 Th has been minor in our ammonium nitrate calibrations, and the nitrate mass fraction is generally low at SMEAR II, we cannot be sure whether the f_{44} and thus the O:C-ratios presented in the VK are overestimated. Thus, the absolute O:C-values should be interpreted with caution. However, if comparing the VK diagram to the VK diagram drawn by (Ng et al., 2011a) representing from 43 ambient AMS datasets, we can see that our SV-OOA O:C is similar to the SV-OOA O:C retrieved by Ng et al. (2010), but our O:C for LV-OOA is higher. Still, our LV-OOA values do resemble those retrieved by Äijälä et al. (2019) with an AMS.

In general, the separation of SV-OOA and LV-OOA in the VK is distinct: the O:C of SV-OOA is ca 30% of the LV-OOA O:C. The SV-OOA H:C is highest, and stays rather constant in the SV-OOA cluster data cloud (slope = 0, slope of adding alcohol or peroxide groups), whereas the H:C decreases as a function of O:C in the LV-OOA cluster data cloud. Due to the scatter in the LV-OOA data cloud we do not aim on quantifying a slope for it.

The second row of projections visualized in Fig. 6 focuses on visualizing key POA characteristics. The f_{44} vs f_{60} visualization used in Fig. 6c is common to distinguish fresh BBOA from aged OA (Cubison et al., 2011). The lower the f_{44} is, the more fresh the OA is expected to be, and the higher the f_{60} is, the higher the fresh BBOA fraction. The POA captured most of the high f_{60} cases (i.e. cases with f_{60} above determined background of 0.003 (Cubison et al., 2011)), and the rest (which also had the highest f_{44}) were included in the LV-OOA cluster. These were clear LV-OOA cluster outliers as these spectra silhouette scores were all below 0.20. Owing to their high f_{60}, these outlier spectra likely originate from biomass burning, but are mixed within the LV-OOA cluster due to the high humic-like substance content of the BBOA (e.g. Ng et al., 2010). If moving to Fig. 6d, i.e. an $f_{55}\times f_{57}$ vs f_{60} diagram, we can see that these high f_{60}-containing LV-OOA points are situated at the bottom of the plot, and all POA objects score a much higher $f_{55}\times f_{57}$. f_{57} has been associated with HOA (Zhang et al., 2005), while f_{55} is present in HOA mass spectra usually at equally high contributions. However, f_{55} is not a good HOA marker alone, as it is present in all of the mass spectra (Fig. 5). Thus, the *y*-axis in Fig. 6d was chosen to be a product of the two instead of a sum of the two, as in this way a high f_{55} (often the case with biogenic SOA) with marginal f_{57} would not be classified as a HOA marker. To conclude, Figs. 6c&d visualize how POA contains both HOA and BBOA features.

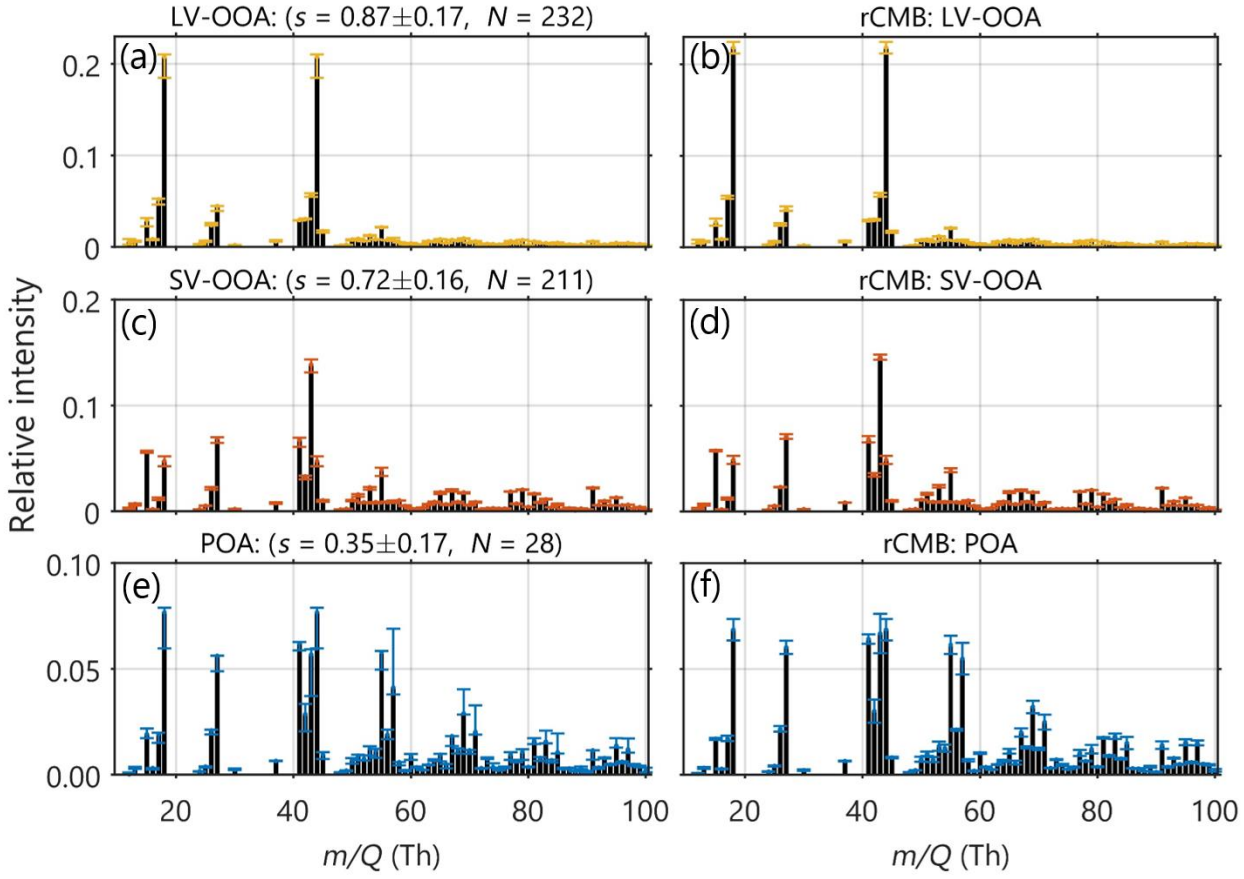

Figure 5 The left panels (a, c, e) represent silhouette-weighted median cluster centroid mass spectra obtained when the number of clusters (*k)* equals 3 in Phase II K-Means clustering (final result). Here, *y*-axis indicates the relative signal intensity and *x*-axis the mass-to-charge ratio (*m*/*Q*) the cluster centroid mass spectra identified as low volatility oxygenated organic aerosol (LV-OOA), semi-volatile oxygenated organic aerosol (SV-OOA) and primary OA (POA). The panel titles include the mean ± standard deviation of the cluster silhouette score (*s*), and the number of spectra belonging to each cluster (*N*). The error bars visualize the 25th and 75th percentiles (i.e. the lower and higher quartiles). The right panels (b, d, f) show the mean LV-OOA, SV-OOA and POA mass spectra obtained from rolling rCMB. The error bars visualize the standard deviation of each *m*/*Q* signal fraction.

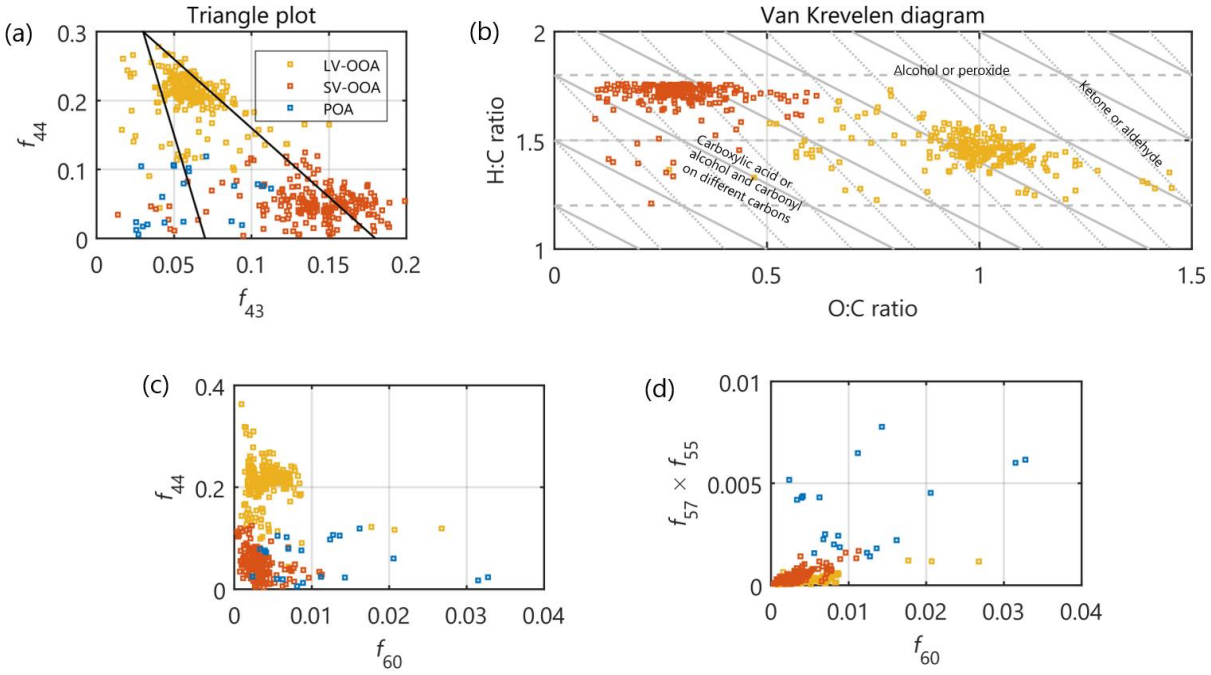

Figure 6 (a) A triangle plot visualizing the mass spectra distribution in each cluster in f_{44} vs f_{43} space, (b) Van Krevelen diagram visualizing the mass spectra in H:C vs O:C space for LV-OOA and SV-OOA, (c) mass spectra in f_{44} vs f_{60} space for indications of fresh BBOA, (d) $f_{55} \times f_{57}$ vs f_{60} space for indications of HOA and BBOA.

6.2 Temporal variability of OA composition

This section contains the analysis of the OA components' time series retrieved via rolling rCMB. These time series are visualized in monthly resolution in Fig. 7. While some of the OA composition variability could be visually extracted from Fig. 7, we focus on the description of Figs 8–10, which summarize the temporal behaviour of each OA component. The three components explained ca. 70–80% of the OA variation at SMEAR II (Fig. 8a). The unexplained variation can be split into data with low SNR (noisy) and data with high SNR. The unexplained fraction due to high noise (low SNR) was lowest in summer, ca. 10%, otherwise ca. 20%. The rest of the unexplained OA variability (data with high SNR) was nearly constant at 10–12%. This fraction is termed as the "real unexplained variation" and includes only the variation made up by variables having the unexplained variation fraction less than 25% (Paatero, 2004). As mentioned before, the unexplained variation did not correlate with any external data nor show seasonal or diel patterns.

6.2.1 LV-OOA

LV-OOA was always the dominating OA type at SMEAR II, both in terms of OA mass fraction ($f_{\text{LV-OOA}}$; Fig. 8b) and absolute concentration (Fig. 7a&9a). LV-OOA is understood to form as a result of OA aging in the atmosphere (e.g. Jimenez et al.,

2009). Indeed, several OA types have been shown to chemically transform to LV-OOA in relatively short time scales (e.g. Jimenez et al., 2009). This makes the dominance of such aged OA product perfectly reasonable at a rural background site, such as SMEAR II. LV-OOA made up ca. 60% of OA mass concentration, and the median absolute LV-OOA loading was 0.74 µg m^{-3} (overall LV-OOA IQR 0.35, 0.74, 1.46 µg m^{-3}).

LV-OOA loading had a bimodal seasonal cycle. The first peak occurred in February (February LV-OOA IQR: 0.30, 0.64, 1.28 µg m^{-3}), similarly as previously reported SMEAR II NR-PM_1 inorganics (Heikkinen et al., 2020). We previously speculated that this February peak of NR-PM_1 inorganics could result from a combination of meteorology-driven phenomena, such as more southerly winds compared to other winter months, the enhanced amount of solar radiation enabling photochemistry, or relatively dry conditions (in terms of less precipitation) diminishing wet deposition of aerosol particles upon transport from more polluted areas. Similar phenomena could certainly favour also higher LV-OOA loading in February. While LV-OOA mass spectrum does not offer insights of possible LV-OOA sources (spectrum comprises mostly of m/Q 44 Th; Fig. 5a), we can still assume the wintertime LV-OOA sources to be mostly anthropogenic due to reduced biogenic activity in the wintertime boreal environment. Wintertime LV-OOA could be to a large extent for example aged wood-burning organic aerosol as wood burning is expected to be the most dominant wintertime OA source in Europe (Jiang et al., 2019). Also anthropogenic SOA formation in urban plumes is a potentially high source of wintertime OOA (Shah et al., 2019). Despite the less efficient oxidation (OH radical concentration much lower in wintertime compared to summer), the cold wintertime temperatures enable condensation of less oxidized organic vapours (e.g. (Stolzenburg et al., 2018)), which could favour wintertime SOA formation. Due to aging processes, it is likely that such wintertime (anthropogenic) SOA would be detected as LV-OOA at SMEAR II due to OOA aging during transport from the far-away urban plumes. The diel cycle of wintertime LV-OOA showed no diel pattern (Fig. 10a). Such behaviour is typical for long-range transported, i.e. not locally produced air pollutants, as boundary layer dynamics will not influence their concentration in the surface layer. More discussion on LV-OOA sources, supporting the abovementioned statements on the anthropogenic and biogenic influences on LV-OOA, is presented later in the paper in conjunction with wind and air mass trajectory analyses (Sect. 6.3).

The second, yet most significant peak of LV-OOA loading occurred in summer (summertime LV-OOA IQR: 0.65, 1.18, 2.01 µg m^{-3}; Fig. 9a), when biogenic emissions rapidly produce SOA in ambient air. It is likely that in summertime biogenic processes were the dominating sources of LV-OOA. LV-OOA possessed a diel cycle clearly only in summer, where the LV-OOA reached a maximum concentration during daytime (Fig. 10a). It is likely that in contrary to wintertime, LV-OOA was produced also locally via photochemical pathways during daytime.

6.2.2 SV-OOA

The highest SV-OOA OA mass fraction ($f_{SV\text{-}OOA}$ Fig. 8b) and loading (Fig. 7b&9b) were observed in summer (unimodal seasonal cycle). The summertime $f_{SV\text{-}OOA}$ was ca. 40% (summertime SV-OOA IQR: 0.33, 0.59, 1.07 µg m^{-3}), otherwise ca. 25–

30% (wintertime SV-OOA IQR: 0.10, 0.17, 0.28 μg m^{-3}). The seasonal cycle of SV-OOA could be explained by the surrounding forest's enhanced biogenic activity in summer months, which leads to biogenic SOA formation. However, we are not able to confirm whether all of the SV-OOA is of biogenic origin. This is because the nearby sawmills in Korkeakoski (ca. 7 km NE of SMEAR II; Sec. 2.1) represent significant SV-OOAF sources (e.g. Äijälä et al., 2017). It is likely that SV-OOA production from terpenes emitted from the Korkeakoski sawmills also express seasonality following the air's oxidation capacity. In addition, it is also possible that terpene emissions from the Korkeakoski sawmills are also temperature-dependent.

SV-OOA possessed a diel cycle in all months but December and January. The SV-OOA diel cycle was typical for semi-volatile species: the maximum loading was achieved in early mornings (Fig. 10b), when atmospheric mixing layer is typically the shallowest and temperature the lowest. We previously reported a similar seasonal cycle for NR-PM_1 nitrate at SMEAR II (Heikkinen et al., 2020). The SV-OOA formation is likely strongly linked to the accumulation of monoterpenes in these shallow nocturnal boundary layers in forests. During calm, stable nights radiative cooling promotes formation of inversion layers hindering vertical dispersion of the forest's emissions. The cooling of the air enables partitioning of less-oxygenated gaseous species yielded from monoterpene oxidation to the condensed phase enhancing also SV-OOA formation. SV-OOA formation via condensation of highly oxidized organic molecules (HOM, which commonly originate from monoterpene oxidation; (Bianchi et al., 2019)), has been previously suggested to occur at SMEAR II's nocturnal boundary layer(s) (Hao et al., 2018).

It is important to mention here that if these ACSM measurements were conducted in a higher altitude, perhaps even a few tens of metres above ground level, such strong diel cycle would likely not have been captured. In addition, upon the development of the turbulent daytime boundary layer the SV-OOA yielded during the night does likely not play any major role in the SV-OOA loading within this daytime boundary layer. The BVOC oxidation in the boreal forest is more efficient during daytime compared to night time (e.g. (Peräkylä et al., 2014)), which would mean a higher production of condensable vapours potentially forming SV-OOA during daytime.

When summing up SV-OOA and LV-OOA, we can see that summertime OA was nearly exclusively OOA (which is typically a good approximation of SOA), and even in wintertime OOA organic mass fraction was ca 80%. High OA mass fractions of OOA in PM_1 have been observed all over the Northern mid-latitudes (Zhang et al., 2007).

6.2.3 POA

The f_{POA} seasonal cycle was opposite to that of SV-OOA, with highest f_{POA} achieved in wintertime (13%; Fig. 8b). The summertime f_{POA} was 3% and the overall median ca. 6%. Interestingly, when comparing the overall median to f_{POA} estimated previously at SMEAR II, we observe much lower fractions. For example, Äijälä et al. (2019) report a HOA OA mass fraction of 6% and BBOA OA mass fraction of 21%. The sum of them, which should somewhat represent POA, is 21 percentage points higher than the mean f_{POA} reported here. As the Äijälä et al. (2019) study was conducted with an AMS the data set should

certainly better capture short-term pollution plumes compared to the ACSM, which has significantly lower time resolution and higher noise level. Another important fact to consider is that the Äijälä et al. (2019) study period is situated between years 2008 and 2010. It is possible that POA emissions have reduced since then, or the emissions were for some reason higher than usual between 2008 and 2010. Hints of such long-term reduction or higher concentrations in 2008-2010 at SMEAR II can be observed in the equivalent black carbon (eBC) concentrations. The eBC concentration between years 2008 and 2011 was nearly twice as high as between years 2013–2018 (Luoma et al., 2020). This could certainly explain some of the discrepancy between these studies.

In addition to the f_{POA}, also the absolute POA concentration peaked in winter (Fig. 7c&9c). The seasonal cycle resembles that of NO_x shown in our previous work (Heikkinen et al., 2020), which in turn follows the cycles of atmospheric boundary layer height and temperature. Several phenomena can explain a larger wintertime POA loading: wintertime POA dispersed in a shallower atmospheric mixing layer compared to summer, and sources of POA are possibly greater in winter due to enhanced need for residential heating and less of POA evaporation due to cold temperatures. In addition, POA wintertime aging to LV-OOA is possibly hindered compared to summertime, due to less efficient photochemical oxidation. The wintertime POA diel cycle showed most of the time a minor afternoon maximum and a minor night-time elevation was slightly visible only in late January/ early February (Fig. 10c). Typical HOA diel cycles in populated areas show an extremely distinct diel pattern following morning and evening rush hours (e.g. (Zhang et al., 2005)). In residential areas, BBOA in turn typically clearly peaks in the evening, when domestic heating takes place and the emissions are dispersed in the nocturnal boundary layer (e.g. (Canonaco et al., 2013)). Due to SMEAR II's distance from major HOA and BBOA sources, we did not observe such clear POA diel cycles in neither summer nor winter. The summertime POA diel cycle resembled a diel cycle of the sum of LV-OOA and SV-OOA. As discussed earlier in Sec. 5.3.1, it is likely that summertime POA loading was overestimated by the rolling rCMB-model (Fig. S.3).

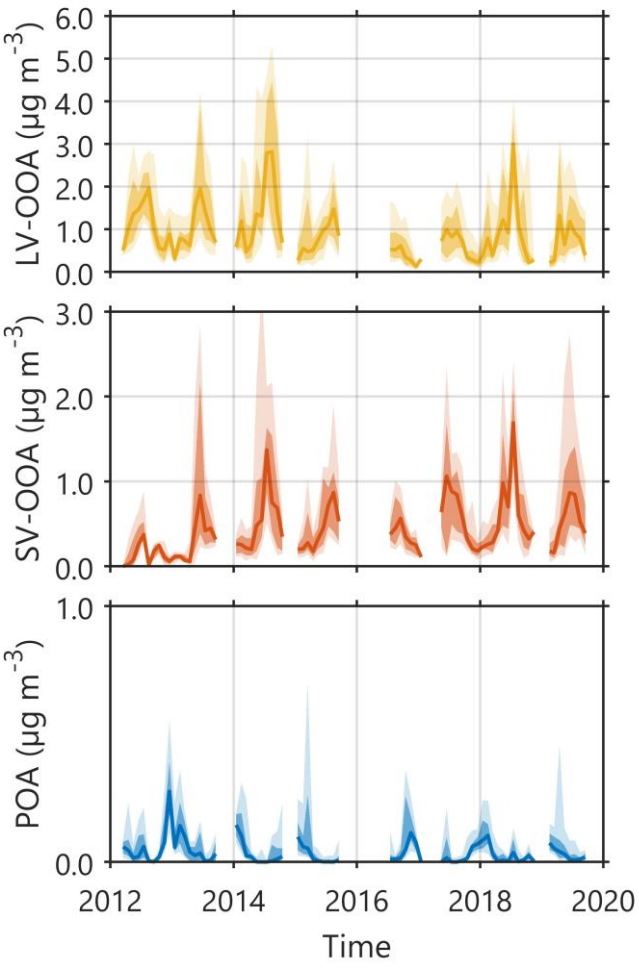

Figure 7 Monthly resolution time series of LV-OOA (panel a), SV-OOA (panel b) and POA (panel c) mass concentrations obtained with rolling rCMB. The light shadings indicate the area between the 10th and 90th percentiles, and the dark shadings the area between the 25th and 75th percentiles. The solid line represents the monthly medians for each month of measurements in 2012 – 2019. Note the different *y*-axes scales (grid lines are drawn every 1 µg m^{-3}).

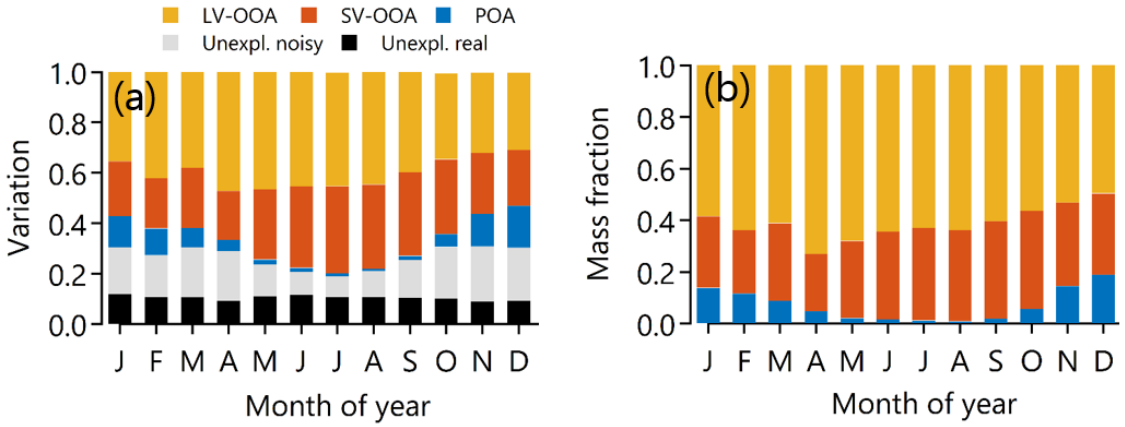

Figure 8 The panel (a) depicts the variability of the rolling rCMB compared to measurement variability (scaled by uncertainty). The unexplained fraction is ca. 30% outside summer, when its ca. 25%. This variation in the unexplained variation is due to increased noisy fraction (light grey) outside summer. The real unexplained fraction (in black) stays at rather constant of ca. 11%. Panel (b) shows $f_{\text{LV-OOA}}$, $f_{\text{SV-OOA}}$ and f_{POA} in different months. This panel only visualizes their variability in rolling rCMB.

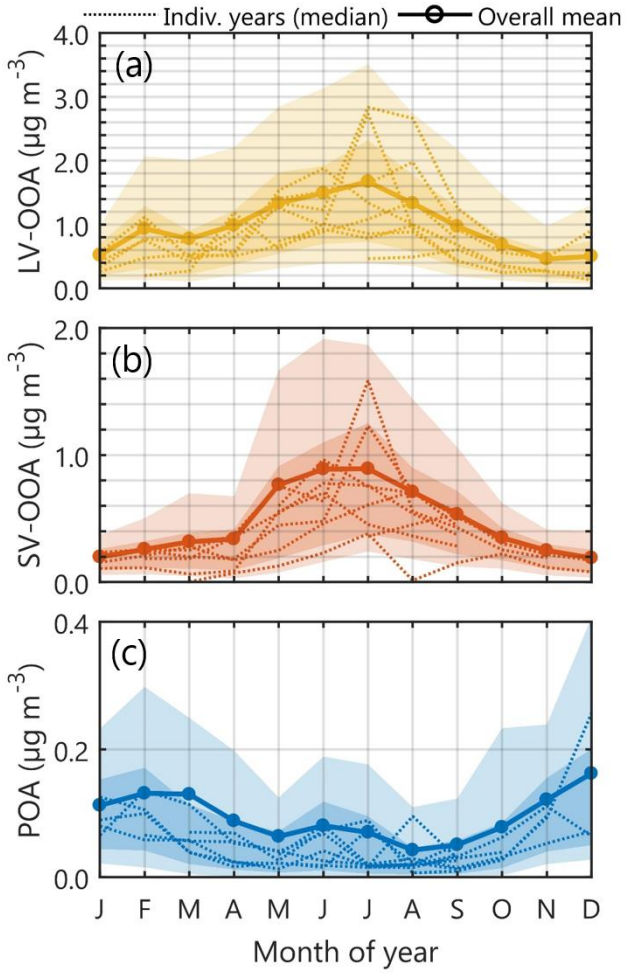

Figure 9 The monthly mass concentrations of LV-OOA (panel a), SV-OOA (panel b) and POA (panel c) obtained with rolling rCMB. The light shadings indicate the area between the 10th and 90th percentiles, and the dark shadings the area between the 25th and 75th percentiles. The narrow dotted lines represent monthly medians for individual years and the dark lines with circled markers represent the overall monthly mean concentrations. Note the different *y*-axes scales (grid lines are drawn every 0.2 μg m^{-3}).

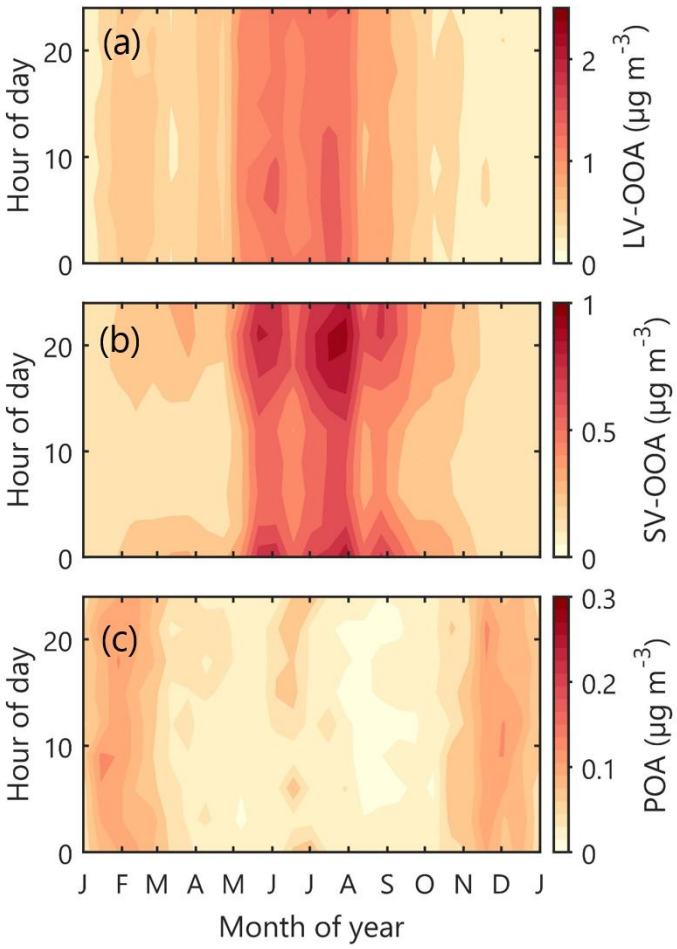

Figure 10 The median diel cycles of LV-OOA (panel a), SV-OOA (panel b) and POA (panel c) obtained via rolling rCMB. The y-axes represent the local time of day (UTC+2) and x-axes the month. The colour scales represent the mass concentration of each OA type. Note the different scales for each plot. Each grid point represents a 14d × 3h period, visualized with the MATLAB 2017a *contourf* function.

6.3 Wind and air mass trajectory influence on OA composition

In this section we will discuss the wind direction and speed dependencies of OA composition, which provide useful insights in estimating whether OA is locally produced or transported. After this analysis we briefly examine the OA types' behaviour as a function of time over land (Sec. 3) to understand the potential magnitude of natural aerosol formation over the boreal forest.

6.3.1 Wind direction and speed dependency of OA composition

The Openair polar plot for LV-OOA is displayed in Fig. 11a. Based on this figure, elevated LV-OOA concentrations could be expected from SE (polluted sectors) regardless of the wind speed. In case of easterly winds, the LV-OOA concentrations were generally the highest if wind speeds stayed below 20 km h^{-1} (ca. 5.6 m s^{-1}). On the contrary, in the case of NW winds (winds from the clean sector) with wind speeds exceeding 20 km h^{-1}, the LV-OOA concentration approached zero implying clean air transport. The LV-OOA Openair polar plot resembles greatly the overall NR-PM_1 organics' Openair polar plot visualized previously in Heikkinen et al. (2020), which was also expected due to LV-OOA being the dominant OA component. The LV-OOA Openair polar plot had more southerly influence in wintertime (Fig. 11b) and significantly less LV-OOA was detected with SE winds compared to the overall picture (Fig. 11a). The summertime LV-OOA Openair polar plot (Fig. 11c) in turn was nearly identical to the median plot including all months.

The SV-OOA concentration was highest with low wind speeds (below 10 km h^{-1}, i.e. ca. 2.8 m s^{-1}; Fig. 11d). In addition, SE winds favoured SV-OOA presence. As SV-OOA loading peaked at night (Fig. 9c), the low wind speed dependence of SV-OOA indicates that calm nights are most suitable for SV-OOA detection. Low nocturnal wind speeds promote the formation of shallow nocturnal boundary layers, as the mixing is not enhanced by mechanically produced eddies. Thus, both the SV-OOA diel cycle and the SV-OOA formation boost at low wind speeds support the hypothesis that SV-OOA is produced locally and it builds up in the night time surface air. However, the Korkeakoski sawmills probably explain why SV-OOA concentration field is darker at the SE side of the Openair plot origin (Fig. 11d). The wintertime SV-OOA Openair polar plot still showed highest SV-OOA loading with low wind speeds, however having less SE influence in the concentration field (Fig. 11e). The summertime polar plot (Fig. 11f) again resembled the overall plot (Fig. 11f). This summertime concentration field of SV-OOA greatly resembled the summertime LV-OOA concentration field (Fig. 11c). The Pearson correlation coefficient between these fields was $R = 0.87$. This similarity supports the previously stated hypothesis that summertime LV-OOA was likely of biogenic origin (also with possible sawmill influence).

Finally, the POA Openair polar plot (Fig. 11g) exemplifies how specific wind direction and speed combinations were required for POA detection: POA was resolvable only if the wind direction was S –SE and wind speed ca. 20 km h^{-1} (rarely the case at

SMEAR II; Fig. S.1). While such high wind speeds ultimately reduce the time the air masses spend over populated areas with potentially high POA emissions, the high wind speeds also enable fast transport of the POA types making their detection at fresh state possible (before POA has evaporated/aged).

The wintertime POA Openair polar plot had also SE influence with less high wind speeds (Fig. 11h). It greatly resembled the wintertime LV-OOA Openair polar plot (Fig. 11b). The Pearson correlation coefficient between the wintertime POA concentration field and LV-OOA concentration field was R = 0.93. The high agreement between these concentration fields supports the previously stated hypothesis that wintertime LV-OOA was likely of anthropogenic origin. The summertime POA Openair polar plot (Fig. 11i) was not greatly differing from the other POA Openair polar plots, which gives some confidence in summertime POA quantification: if most summertime POA was overestimated, the summertime POA Openair polar plot would likely have similar wind dependence as OOA.

6.3.2 Time over land analysis

Tunved et al. (2006) showed how (organic) aerosol mass concentration increased as a function of time over land (TOL; i.e. the number of hours the air mass spent over the forested land surface upwind of SMEAR II) when the land surface had little anthropogenic influence (e.g. in the clean north-westerly sector; Fig. S.2). This increase was attributed to natural (biogenic) OA production in the boreal boundary layer. Here, we observe a similar increase in the clean sector (Fig. 12a), LV-OOA loading being the most sensitive to TOL (Fig. 12). The lower increase shown for SV-OOA (Fig. 12) in comparison to LV-OOA supports our hypothesis of SV-OOA sources being also local and SV-OOA aging into LV-OOA. The relationship between POA and TOL was not significant (Fig. 12). The increase of LV-OOA loading as a function of TOL indicates OA formation in the boreal boundary layer, its build-up in the air mass, and aging into LV-OOA prior to arrival at SMEAR II. Such phenomenon is not visible when investigating the OA types' behaviour as a function of TOL in polluted sectors (Fig. S.8). Indeed, none of the OA-types indicate links between OA loading and TOL in neither air masses of European (southerly sector) nor Russian (easterly sector) origin. We are not surprised of such lack of correlation between OA and TOL as the picture is greatly hampered by anthropogenic emissions. As the anthropogenic emissions are minor in the clean sector, and as suggested by Tunved et al. (2006), the OA production in the clean sector is dominated by biogenic SOA formation.

The biogenic SOA hypothesis is supported also by the seasonality of the OA vs TOL relationship (Fig. 12b): a highest correlation between the two and the steepest OA increase as a function of TOL is observed in July, which held the greatest temperatures during the measurement period (Heikkinen et al., 2020). Such temperature dependence is typically associated with biogenic SOA production (e.g. (Daellenbach et al., 2017;Stefenelli et al., 2019)) as the emission rates of several SOA precursors (such as monoterpenes) increase as a function of temperature (Guenther et al., 1993). The linear regression slopes for a LV-OOA vs TOL scatter plot would suggest LV-OOA formation of ca. 42 ng m^{-3} h^{-1} in July, which is twice the SV-OOA vs TOL slope (Fig. 12b). To exemplify these numbers, three days over the boreal forest in July would yield ca. 3 µg m^{-3} of

LV-OOA and 1.6 μg m^{-3} of SV-OOA. The slopes for LV-OOA stay below 10 ng m^{-3} h^{-1} between October and April (values similar to the slopes for SV-OOA at the same time; Fig. 12b), when there is less of biogenic plant activity. These slopes were similar in magnitude to those derived previously for SMEAR II data (Tunved et al., 2006;Liao et al., 2014). Another interesting feature extracted from this analysis was that if the OA type vs TOL slopes were calculated using data only below TOL = 40 h, the SV-OOA and LV-OOA slopes would be identical, and only after TOL exceeds 40 h, LV-OOA loading keeps increasing while the SV-OOA loading shows a minor decreasing trend (Fig. 12c). More analysis and perhaps investigations of similar plots from other boreal research stations could give us insights whether the figure informs more of time scales of OA chemistry or whether it is linked to meteorology and/or distance to the ocean from the measurement station. Additionally, also cloud processing and subsequent precipitation will influence aerosol size distribution during the transport to the observation site. However, in this study we did not take these interactions and precipitation processes into account. Our aim was to explore the net effect of TOL in sub-micrometre aerosol chemistry at a fixed site. Therefore a need to explore these features in a systematic manner in the future also exists.

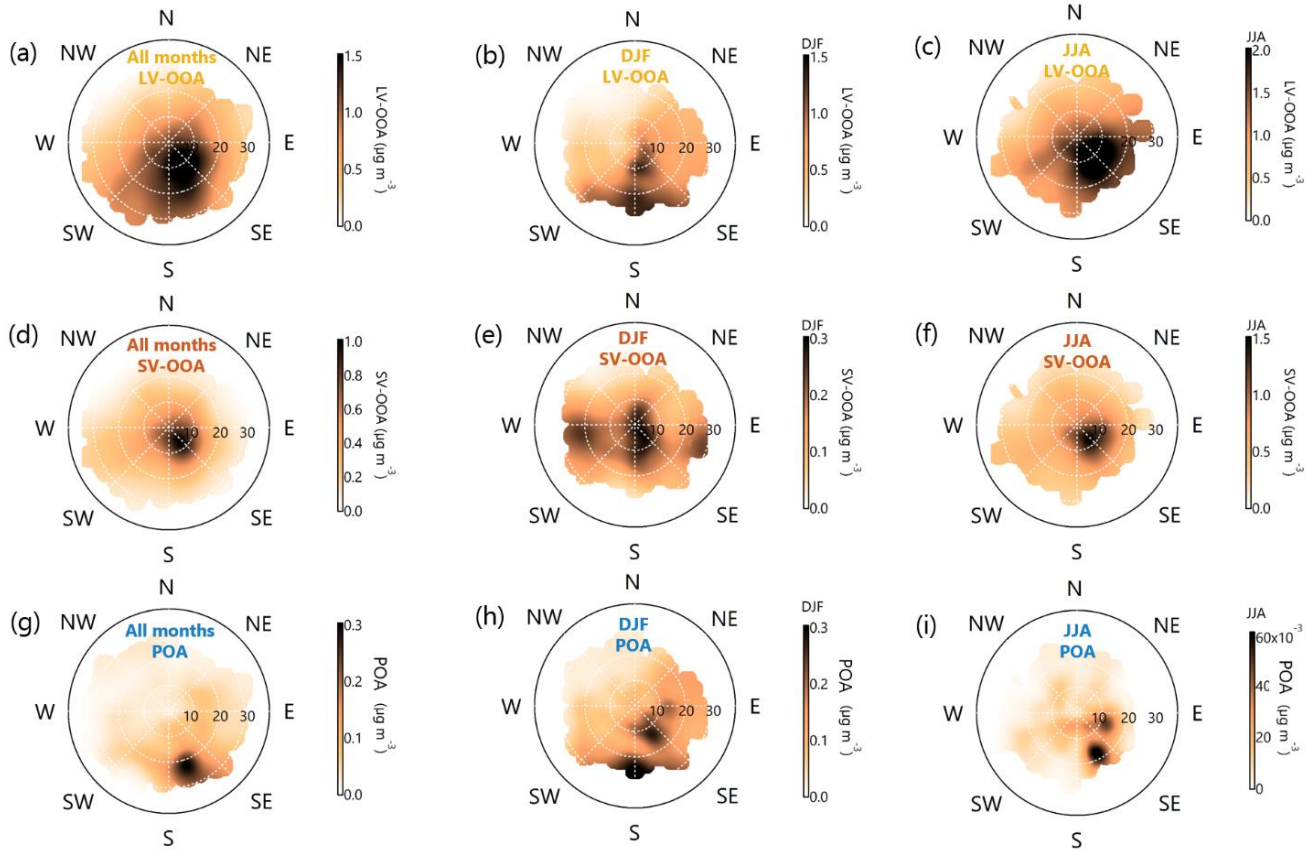

Figure 11 Openair polar plots (Carslaw and Ropkins, 2012) for LV-OOA (first row), SV-OOA (second row) and POA (third row) obtained via ZeFir pollution tracker Wavemetrics Igor Pro toolkit (Petit et al., 2017). The first column represents the median over all seasons, the second column the median over wintertime and third the median summertime. The distances from the circle origins indicate wind speeds (in km h^{-1}). Wind speed grid lines are presented with dark grey dashed lines. The colour scales represent the mass concentration of each OA type modelled via rolling rCMB during the specific wind direction and speed combinations. Note that the scales are different among the subplots. As these figures do not indicate any likelihood of these wind direction and speed combinations, Fig. S.1 is important to keep in mind while interpreting them. Briefly, N-NE-E is the least likely direction of wind, and S-SW-W is the most likely. Wind speeds rarely exhibit 20 km h^{-1}. The wind direction and speed data are collected above the boreal forest canopy.

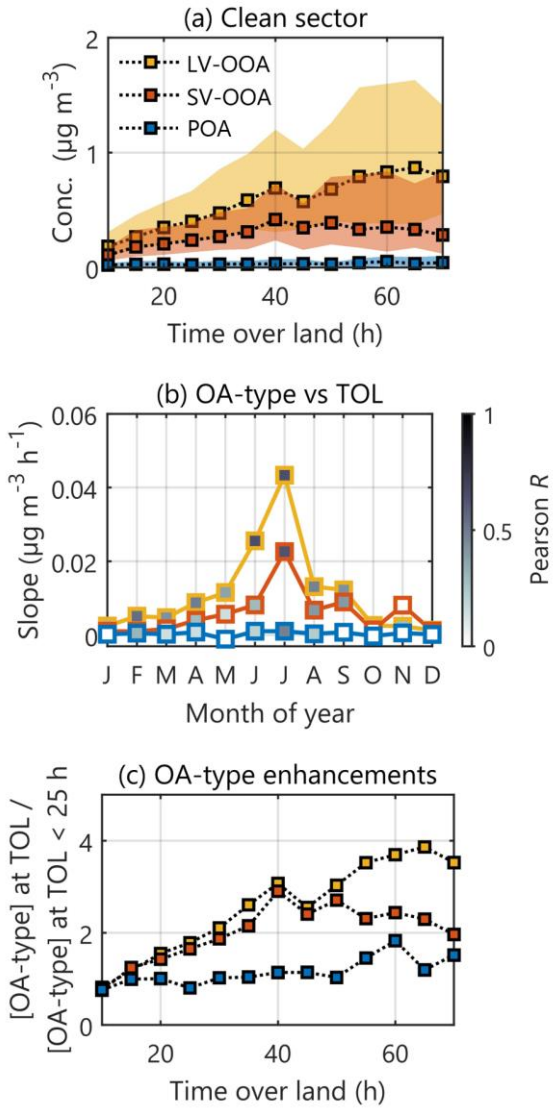

Figure 12 (a) The different rolling rCMB factors (*y* axes in µg m^{-3}) vs TOL (*x* axes in hours) for the clean sector (least polluted north-western sector as defined by Tunved et al., 2006; see Fig. S.2 for a more precise sector definition). The data are binned to 5-hourly TOL bins. The shaded areas represents the concentration interquartile ranges (25^{th} to 75^{th} percentile) and the square markers the median concentrations. (b) The slopes (in µg m^{-3} h^{-1}) are calculated for a linear fit between TOL ([20, 70] h) and the three different OA types. (c) The OA type concentration in the TOL bin divided by the median OA type concentration when TOL was < 25 h as a function of TOL. The plot visualizes how the SV-OOA and LV-OOA have similar behaviour until TOL = 40 h.

7 Conclusions

Organic aerosol (OA) mass spectra are recorded continuously with an Aerosol Chemical Speciation Monitor (ACSM) since 2012 at SMEAR II station, located within the boreal forest in Southern Finland. The goal of the current paper was to yield understanding of the main OA components: their mass spectral features and temporal behaviours. The large extent of input data (eight years) and the relatively remote measurement location required us to develop a new framework for conducting OA chemical characterization, as to our knowledge there are no previous studies where equally long or longer time series of OA mass spectra from similar environments have been characterized. We approached the OA characterization via Positive Matrix Factorization (PMF; Paatero and Tapper, 1994). However, due to the length of the data set, we conducted the PMF with a 30-day rolling window approach, which enabled factor profile variability across the eight years (Canonaco et al., 2021;Parworth et al., 2015). The rolling PMF yielded an extremely large number of PMF solutions (20 900 solutions, 62 700 factor profiles). We explored the PMF profiles across the solution space using K-Means clustering to gain understanding of the dominant OA types at the station. We revealed/identified three significantly different OA clusters: low-volatility oxygenated OA (LV-OOA), semi-volatile oxygenated OA (SV-OOA) and primary OA (POA) from these data. To attain their temporal variabilities, we performed a rolling relaxed Chemical Mass Balance (rolling rCMB) run, anchored by the observed clusters and their intra-cluster variabilities as opposed to the more conventional methods introduced e.g. by Canonaco et al. (2021). The selection of K-Means and rolling rCMB combination instead of a conventional rolling PMF enabled us to quantify POA at SMEAR II. The rCMB run explained ca. 70% of the observed OA at SMEAR II and nearly two thirds of the unexplained variation was due to high noise level of the data leaving the real unexplained variation at only 11%. The analysis method utilized here turned out to be robust, and it required little analyst interference. Therefore, our framework presents a technique to effectively analyse long-term AMS or ACSM datasets while reducing subjective bias upon analysis. However, more work is potentially needed in the future to optimize the analysis stages proposed.

With equal importance to the tested data analysis framework, we also presented the OA composition and its variability at SMEAR II. The main conclusion to be drawn from the OA composition at SMEAR II is that this boreal OA is nearly exclusively oxidized organic aerosol, mostly highly oxidized LV-OOA. The result was well in line with previous studies from the Northern Hemisphere showing the ubiquity of OOA especially at rural measurement sites (Zhang et al., 2007). The LV-OOA seasonal cycle was bimodal culminating in February and summer. The wintertime LV-OOA was likely anthropogenic and the February peak coincided with NR-PM_1 inorganics (Heikkinen et al., 2020). The summertime LV-OOA had enhanced biogenic influence and it was linearly increasing the longer the air mass had spent over the boreal forest. We estimated natural LV-OOA production of several tens of ng m^{-3} per hour. These numbers were well in line with previous studies investigating the natural aerosol production in the boreal forest (Tunved et al., 2006;Liao et al., 2014). SV-OOA was the second most

abundant OA type and the maximum SV-OOA concentration was detected in early mornings during summer. Both biogenic processes and emissions from the nearby sawmill contribute to the SV-OOA mass as also exemplified in previous studies (e.g. Äijälä et al., 2017). Highest SV-OOA loadings were observed when sampling from shallow nocturnal surface layers, but it is possible that the production of SV-OOA was highest during daytime when most BVOC oxidation takes place. Finally, the POA, the mass spectrum of which resembled both hydrocarbon-like OA and biomass burning OA, attained significant OA mass fractions only in winter. Still, those OA mass fractions were significantly lower compared to earlier long-term descriptions of SMEAR II OA composition (Äijälä et al., 2019). This discrepancy could be for example linked to a decrease in POA emissions as hinted by decreasing BC trends at the site (Luoma et al., 2020), or the ACSM limited capability in detecting short-term (pollution) plumes, which average out even more due to the 3-hour averaging applied to the PMF input data, which was necessary to improve the SNR at this rural background site. More generally, due to OA composition sensitivity to meteorological conditions and anomalies, even longer time series need to be accumulated in order to reliably estimate trends of POA and other OA constituents at SMEAR II based on ACSM data.

Data availability

The ACSM NR-PM_1 OA concentration data are available at EBAS database under EMEP ACTRIS framework as well as upon request from the corresponding authors. The PMF matrices and OA classes' mass spectral profiles and time series are available upon request from the corresponding authors. The wind direction and speed data are available at the SmartSMEAR data repository (https://avaa.tdata.fi/web/smart) (Junninen et al., 2009). Contact of the original data contributors can be requested from atm-data@helsinki.fi.

Competing interests

The authors declare no conflict of interest.

Author contributions

LH, MÄ, ME, TP, MK, and DW designed the study. LH, MÄ and FG performed the ACSM measurements. PA provided size distribution data needed for 2019 ACSM data processing. MR performed time over land calculations with HYSPLIT trajectories. KL provided BC data for 2019 ACSM data processing. The ACSM data processing was performed by LH. KD, CG and MÄ assisted LH with (rolling) PMF. MÄ and DA assisted LH with the K-Means clustering. LH performed the overall analysis, data visualisation and wrote the paper with comments from the co-authors. ME supervised all the steps in this process.

Acknowledgements

We thank SMEAR II staff for keeping the measurements running, COST COLOSSAL for valuable guidance and discussions, and Francesco Canonaco (Datalystica Ltd) for PMF support. We thank Santtu Mikkonen and Jean-Eudes Petit for useful discussions. The European Research Council Horizon 2020 (grants 638703, 689443, 821205), the Academy of Finland (grants 317380, 320094, 307537, 324259, 333397 and 334792), and the project Source apportionment using long-term Aerosol Mass Spectrometry and Aethalometer Measurements (SAMSAM), funded by the Swiss National Science Foundation (SNF) in the framework of COST (IZCOZ0_177063), supported this research. Finally, we acknowledge the University of Helsinki and Academy of Finland support to ACTRIS infrastructure (grants 329274 and 328616).

References

Aiken, A. C., Decarlo, P. F., Kroll, J. H., Worsnop, D. R., Huffman, J. A., Docherty, K. S., Ulbrich, I. M., Mohr, C., Kimmel, J. R., Sueper, D., Sun, Y., Zhang, Q., Trimborn, A., Northway, M., Ziemann, P. J., Canagaratna, M. R., Onasch, T. B., Alfarra, R. M., Prevot, A. S. H., Dommen, J., Duplissy, J., Metzger, A., Baltensperger, U., and Jimenez, J. L.: O/C and OM/OC ratios of primary, secondary, and ambient organic aerosols with high-resolution time-of-flight aerosol mass spectrometry, Environmental Science & Technology, 42, 4478-4485, 2008.

Alfarra, M. R.: Insights into the atmospheric organic aerosols using an aerosol mass spectrometer, University of Manchester, UK, Thesis, 2004.

Arthur, D., and Vassilvitskii, S.: K-Means++: The Advantages of Careful Seeding, In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 1027-1035 pp., 2007.

Ball, G. H., and Hall, D. J.: ISODATA, a novel method of analysis and pattern classification, DTIC Document, Technical report, Stanford research institute, Menlo Park, CA, USA, 1965.

Barreira, F. M., Luis, Duporte, G., Parshintsev, J., Hartonen, K., Jussila, M., Aalto, J., Bäck, J., Kulmala, M., and Riekkola, M.-L.: Emissions of biogenic volatile organic compounds from the boreal forest floor and understory, Boreal Environment Research, 2017.

Bianchi, F., Kurtén, T., Riva, M., Mohr, C., Rissanen, M. P., Roldin, P., Berndt, T., Crounse, J. D., Wennberg, P. O., Mentel, T. F., Wildt, J., Junninen, H., Jokinen, T., Kulmala, M., Worsnop, D. R., Thornton, J. A., Donahue, N., Kjaergaard, H. G., and Ehn, M.: Highly Oxygenated Organic Molecules (HOM) from Gas-Phase Autoxidation Involving Peroxy Radicals: A Key Contributor to Atmospheric Aerosol, Chemical Reviews, 119, 3472-3509, 10.1021/acs.chemrev.8b00395, 2019.

Canagaratna, M. R., Jayne, J. T., Jimenez, J. L., Allan, J. D., Alfarra, M. R., Zhang, Q., Onasch, T. B., Drewnick, F., Coe, H., Middlebrook, A., Delia, A., Williams, L. R., Trimborn, A. M., Northway, M. J., DeCarlo, P. F., Kolb, C. E., Davidovits, P., and Worsnop, D. R.: Chemical and microphysical characterization of ambient aerosols with the aerodyne aerosol mass spectrometer, Mass Spectrometry Reviews, 26, 185-222, 10.1002/mas.20115, 2007.

Canagaratna, M. R., Jimenez, J. L., Kroll, J. H., Chen, Q., Kessler, S. H., Massoli, P., Hildebrandt Ruiz, L., Fortner, E., Williams, L. R., Wilson, K. R., Surratt, J. D., Donahue, N. M., Jayne, J. T., and Worsnop, D. R.: Elemental ratio measurements of organic compounds using aerosol mass spectrometry: characterization, improved calibration, and implications, Atmospheric Chemistry and Physics, 15, 253-272, 10.5194/acp-15-253-2015, 2015.

Canonaco, F., Crippa, M., Slowik, J. G., Baltensperger, U., and Prévôt, A. S. H.: SoFi, an IGOR-based interface for the efficient use of the generalized multilinear engine (ME-2) for the source apportionment: ME-2 application to aerosol mass spectrometer data, Atmospheric Measurement Techniques, 6, 3649-3661, 10.5194/amt-6-3649-2013, 2013.

Canonaco, F., Tobler, A., Chen, G., Sosedova, Y., Slowik, J. G., Bozzetti, C., Daellenbach, K. R., El Haddad, I., Crippa, M., Huang, R. J., Furger, M., Baltensperger, U., and Prévôt, A. S. H.: A new method for long-term source apportionment with time-dependent factor profiles and uncertainty assessment using SoFi Pro: application to 1 year of organic aerosol data, Atmos. Meas. Tech., 14, 923-943, 10.5194/amt-14-923-2021, 2021.

Carslaw, D. C., and Ropkins, K.: Openair—an R package for air quality data analysis, Environmental Modelling & Software, 27, 52-61, 2012.

Crenn, V., Sciare, J., Croteau, P. L., Verlhac, S., Fröhlich, R., Belis, C. A., Aas, W., Äijälä, M., Alastuey, A., Artiñano, B., Baisnée, D., Bonnaire, N., Bressi, M., Canagaratna, M., Canonaco, F., Carbone, C., Cavalli, F., Coz, E., Cubison, M. J., Esser-Gietl, J. K., Green, D. C., Gros, V., Heikkinen, L., Herrmann, H., Lunder, C., Minguillón, M. C., Močnik, G., O'Dowd, C. D., Ovadnevaite, J., Petit, J. E., Petralia, E., Poulain, L., Priestman, M., Riffault, V., Ripoll, A., Sarda-Estève, R., Slowik, J. G., Setyan, A., Wiedensohler, A., Baltensperger, U., Prévôt, A. S. H., Jayne, J. T., and Favez, O.: ACTRIS ACSM intercomparison – Part 1: Reproducibility of concentration and fragment results from 13 individual Quadrupole Aerosol Chemical Speciation Monitors (Q-ACSM) and consistency with co-located instruments, Atmospheric Measurement Techniques, 8, 5063-5087, 10.5194/amt-8-5063-2015, 2015.

Crippa, M., Canonaco, F., Lanz, V., Äijälä, M., Allan, J., Carbone, S., Capes, G., Ceburnis, D., Dall'Osto, M., Day, D., DeCarlo, P. F., Ehn, M., Eriksson, A., Freney, E., Hildebrandt Ruiz, L., Hillamo, R., Jimenez, J. L., Junninen, H., Kiendler-Scharr, A., Kortelainen, A.-M., Kulmala, M., Laaksonen, A., Mensah, A. A., Mohr, C., Nemitz, E., O'Dowd, C., Ovadnevaite, J., Pandis, S. N., Petäjä, T., Poulain, L., Saarikoski, S., Sellegri, K., Swietlicki, E., Tiitta, P., Worsnop, D. R., Baltensperger, U., and Prévôt, A. S. H.: Organic aerosol components derived from 25 AMS data sets across Europe using a consistent ME-2 based source apportionment approach, Atmospheric Chemistry and Physics, 14, 6159-6176, 2014.

Cubison, M., Ortega, A., Hayes, P., Farmer, D., Day, D., Lechner, M., Brune, W. H., Apel, E., Diskin, G., Fisher, J., Hecobian, A., Knapp, D., Mikoviny, T., Riemer, D., Sache, G., Sessions, W., Weber, R., Weinheimer, A., Wisthaler, A., and Jimenez, J. L.: Effects of aging on organic aerosol from open biomass burning smoke in aircraft and laboratory studies, Atmospheric Chemistry and Physics, 11, 12049-12064, 2011.

Daellenbach, K. R., Stefenelli, G., Bozzetti, C., Vlachou, A., Fermo, P., Gonzalez, R., Piazzalunga, A., Colombi, C., Canonaco, F., Hueglin, C., Kasper-Giebl, A., Jaffrezo, J.-L., Bianchi, F., Slowik, J. G., Baltensperger, U., El-Haddad, I., and Prévôt, A. S. H.: Long-term chemical analysis and organic aerosol source apportionment at nine sites in central Europe: source identification and uncertainty assessment, Atmospheric Chemistry and Physics, 17, 13265-13282, 2017.

De Gouw, J. A., Middlebrook, A. M., Warneke, C., Goldan, P. D., Kuster, W. C., Roberts, J. M., Fehsenfeld, F. C., Worsnop, D. R., Canagaratna, M. R., Pszenny, A. A. P., Keene, W. C., Marchewka, M., Bertman, S. B., and Bates, T. S.: Budget of organic carbon in a polluted atmosphere: Results from the New England Air Quality Study in 2002, Journal of Geophysical Research: Atmospheres, 110, 10.1029/2004jd005623, 2005.

Donahue, N. M., Trump, E. R., Pierce, J. R., and Riipinen, I.: Theoretical constraints on pure vapor-pressure driven condensation of organics to ultrafine particles, Geophysical Research Letters, 38, 10.1029/2011gl048115, 2011.

Duplissy, J., DeCarlo, P. F., Dommen, J., Alfarra, M. R., Metzger, A., Barmpadimos, I., Prevot, A. S. H., Weingartner, E., Tritscher, T., Gysel, M., Aiken, A. C., Jimenez, J. L., Canagaratna, M. R., Worsnop, D. R., Collins, D. R., Tomlinson, J., and Baltensperger, U.: Relating hygroscopicity and composition of organic aerosol particulate matter, Atmospheric Chemistry and Physics, 11, 1155-1165, 10.5194/acp-11-1155-2011, 2011.

Eerdekens, G., Yassaa, N., Sinha, V., Aalto, P., Aufmhoff, H., Arnold, F., Fiedler, V., Kulmala, M., and Williams, J.: VOC measurements within a boreal forest during spring 2005: on the occurrence of elevated monoterpene concentrations during night time intense particle concentration events, Atmospheric Chemistry and Physics, 9, 8331-8350, 2009.

Efron, B.: Bootstrap methods: another look at the jackknife, Annual Statistics, 20, 393-403, 1979.

Ehn, M., Kleist, E., Junninen, H., Petäjä, T., Lönn, G., Schobesberger, S., Dal Maso, M., Trimborn, A., Kulmala, M., Worsnop, D. R., Wahner, A., Wildt, J., and Mentel, T. F.: Gas phase formation of extremely oxidized pinene reaction products in chamber and ambient air, Atmospheric Chemistry and Physics, 12, 5113-5127, 10.5194/acp-12-5113-2012, 2012.

Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., Springer, M., Rubach, F., Tillmann, R., Lee, B., Lopez-Hilfiker, F., Andres, S., Acir, I.-H., Rissanen, M., Jokinen, T., Schobesberger, S., Kangasluoma, J., Kontkanen, J., Nieminen, T., Kurtén, T., Nielsen, L. B., Jørgensen, S., Kjaergaard, H. G., Canagaratna, M., Maso, M. D., Berndt, T., Petäjä, T., Wahner, A., Kerminen, V.-M., Kulmala, M., Worsnop, D. R., Wildt, J., and Mentel, T. F.: A large source of low-volatility secondary organic aerosol, Nature, 506, 476-479, 10.1038/nature13032, 2014.

Freney, E., Zhang, Y., Croteau, P., Amodeo, T., Williams, L., Truong, F., Petit, J.-E., Sciare, J., Sarda-Esteve, R., Bonnaire, N., Arumae, T., Aurela, M., Bougiatioti, A., Mihalopoulos, N., Coz, E., Artinano, B., Crenn, V., Elste, T., Heikkinen, L., Poulain, L., Wiedensohler, A., Herrmann, H., Priestman, M., Alastuey, A., Stavroulas, I., Tobler, A., Vasilescu, J., Zanca, N., Canagaratna, M., Carbone, C., Flentje, H., Green, D., Maasikmets, M., Marmureanu, L., Minguillon, M. C., Prevot, A. S. H., Gros, V., Jayne, J., and Favez, O.: The second ACTRIS inter-comparison (2016) for Aerosol Chemical Speciation Monitors (ACSM): Calibration protocols and instrument performance evaluations, Aerosol Science and Technology, 53, 830-842, 10.1080/02786826.2019.1608901, 2019.

Fröhlich, R., Crenn, V., Setyan, A., Belis, C. A., Canonaco, F., Favez, O., Riffault, V., Slowik, J. G., Aas, W., Äijälä, M., Alastuey, A., Artiñano, B., Bonnaire, N., Bozzetti, C., Bressi, M., Carbone, C., Coz, E., Croteau, P. L., Cubison, M. J., Esser-Gietl, J. K., Green, D. C., Gros, V., Heikkinen, L., Herrmann, H., Jayne, J. T., Lunder, C. R., Minguillón, M. C., Močnik, G., O'Dowd, C. D., Ovadnevaite, J., Petralia, E., Poulain, L., Priestman, M., Ripoll, A., Sarda-Estève, R., Wiedensohler, A., Baltensperger, U., Sciare, J., and Prévôt, A. S. H.: ACTRIS ACSM intercomparison – Part 2: Intercomparison of ME-2 organic source apportionment results from 15 individual, co-located aerosol mass spectrometers, Atmospheric Measurement Techniques, 8, 2555-2576, 10.5194/amt-8-2555-2015, 2015.

Goldstein, A. H., and Galbally, I. E.: Known and Unexplored Organic Constituents in the Earth's Atmosphere, Environmental Science & Technology, 41, 1514-1521, 10.1021/es072476p, 2007.

Guenther, A. B., Zimmerman, P. R., Harley, P. C., Monson, R. K., and Fall, R.: Isoprene and monoterpene emission rate variability: model evaluations and sensitivity analyses, Journal of Geophysical Research: Atmospheres, 98, 12609-12617, 1993.

Guenther, A. B., Jiang, X., Heald, C. L., Sakulyanontvittaya, T., Duhl, T., Emmons, L. K., and Wang, X.: The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions, Geoscientific Model Development, 5, 1471-1492, 10.5194/gmd-5-1471-2012, 2012.

Hakola, H., Hellén, H., Hemmilä, M., Rinne, J., and Kulmala, M.: In situ measurements of volatile organic compounds in a boreal forest, Atmospheric Chemistry and Physics, 12, 11665-11678, 2012.

Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, T. F., Monod, A., Prévôt, A. S. H., Seinfeld, J. H., Surratt, J. D., Szmigielski,

R., and Wildt, J.: The formation, properties and impact of secondary organic aerosol: current and emerging issues, Atmospheric Chemistry and Physics, 9, 5155-5236, 10.5194/acp-9-5155-2009, 2009.

Hao, L., Garmash, O., Ehn, M., Miettinen, P., Massoli, P., Mikkonen, S., Jokinen, T., Roldin, P., Aalto, P., Yli-Juuti, T., Joutsensaari, J., Petäjä, T., Kulmala, M., Lehtinen, K. E. J., Worsnop, D. R., and Virtanen, A.: Combined effects of boundary layer dynamics and atmospheric chemistry on aerosol composition during new particle formation periods, Atmospheric Chemistry and Physics, 18, 17705-17716, 10.5194/acp-18-17705-2018, 2018.

Hari, P., and Kulmala, M.: Station for measuring ecosystem-atmosphere relations (SMEAR II), Boreal Environment Research, 10, 315-322, 2005.

Heald, C. L., Kroll, J. H., Jimenez, J. L., Docherty, K. S., DeCarlo, P. F., Aiken, A. C., Chen, Q., Martin, S. T., Farmer, D. K., and Artaxo, P.: A simplified description of the evolution of organic aerosol composition in the atmosphere, Geophysical Research Letters, 37, 10.1029/2010gl042737, 2010.

Heikkinen, L., Äijälä, M., Riva, M., Luoma, K., Daellenbach, K., Aalto, J., Aalto, P., Aliaga, D., Aurela, M., Keskinen, H., Makkonen, U., Rantala, P., Kulmala, M., Petäjä, T., Worsnop, D., and Ehn, M.: Long-term sub-micrometer aerosol chemical composition in the boreal forest: inter- and intra-annual variability, Atmospheric Chemistry and Physics, 20, 3151-3180, 10.5194/acp-20-3151-2020, 2020.

Jain, A. K.: Data clustering: 50 years beyond K-means, Pattern Recognition Letters, 31, 651-666, 2010.

Jiang, J., Aksoyoglu, S., El-Haddad, I., Ciarelli, G., Denier van der Gon, H. A. C., Canonaco, F., Gilardoni, S., Paglione, M., Minguillón, M. C., Favez, O., Zhang, Y., Marchand, N., Hao, L., Virtanen, A., Florou, K., O'Dowd, C., Ovadnevaite, J., Baltensperger, U., and Prévôt, A. S. H.: Sources of organic aerosols in Europe: a modeling study using CAMx with modified volatility basis set scheme, Atmospheric Chemistry and Physics, 19, 15247-15270, 10.5194/acp-19-15247-2019, 2019.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson, J. M., Collins, D. R., Cubison, M. J., Dunlea, E. J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper, D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., and Worsnop, D. R.: Evolution of Organic Aerosols in the Atmosphere, Science, 326, 1525-1529, 10.1126/science.1180353, 2009.

Junninen, H., Lauri, A., Keronen, P., Aalto, P., Hiltunen, V., Hari, P., and Kulmala, M.: Smart-SMEAR: on-line data exploration and visualization tool for SMEAR stations, Boreal Environment Research, 2009.

Kanamitsu, M.: Description of the NMC Global Data Assimilation and Forecast System %J Weather and Forecasting, 4, 335-342, 10.1175/1520-0434(1989)004<0335:Dotngd>2.0.Co;2, 1989.

Kaufman, L., and Rousseeuw, P. J.: Finding groups in data: an introduction to cluster analysis, John Wiley & Sons, 2009.

Kleinman, L. I., Springston, S. R., Daum, P. H., Lee, Y. N., Nunnermacker, L. J., Senum, G. I., Wang, J., Weinstein-Lloyd, J., Alexander, M. L., Hubbe, J., Ortega, J., Canagaratna, M. R., and Jayne, J.: The time evolution of aerosol composition over the Mexico City plateau, Atmospheric Chemistry and Physics, 8, 1559-1575, 10.5194/acp-8-1559-2008, 2008.

Lanz, V. A., Alfarra, M. R., Baltensperger, U., Buchmann, B., Hueglin, C., and Prévôt, A. S. H.: Source apportionment of submicron organic aerosols at an urban site by factor analytical modelling of aerosol mass spectra, Atmospheric Chemistry and Physics, 7, 1503-1522, 10.5194/acp-7-1503-2007, 2007.

Liao, L., Dal Maso, M., Taipale, R., Rinne, J., Ehn, M., Junninen, H., Äijälä, M., Nieminen, T., Alekseychik, P., Hulkkonen, M., Worsnop, D., Kerminen, V.-M., and Kulmala, M.: Monoterpene pollution episodes in a forest environment: indication of anthropogenic origin and association with aerosol particles, Boreal Environment Research, 16, 288-303, 2011.

Liao, L., Kerminen, V. M., Boy, M., Kulmala, M., and Dal Maso, M.: Temperature influence on the natural aerosol budget over boreal forests, Atmospheric Chemistry and Physics, 14, 8295-8308, 10.5194/acp-14-8295-2014, 2014.

Liu, P. S., Deng, R., Smith, K. A., Williams, L. R., Jayne, J. T., Canagaratna, M. R., Moore, K., Onasch, T. B., Worsnop, D. R., and Deshler, T.: Transmission efficiency of an aerodynamic focusing lens system: Comparison of model calculations and laboratory measurements for the Aerodyne Aerosol Mass Spectrometer, Aerosol Science and Technology, 41, 721-733, 2007.

Luoma, K., Niemi, J. V., Helin, A., Aurela, M., Timonen, H., Virkkula, A., Rönkkö, T., Kousa, A., Fung, P. L., Hussein, T., and Petäjä, T.: Spatiotemporal variation and trends of equivalent black carbon in the Helsinki metropolitan area in Finland, Atmospheric Chemistry and Physics Discussions, 2020, 1-30, 10.5194/acp-2020-201, 2020.

MacQueen, J.: Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 21 June–18 July 1965 and 27 December 1965–7 January 1966, Statistical Laboratory of the University of California, Berkeley, USA, 281–297, 1967.

Middlebrook, A. M., Bahreini, R., Jimenez, J. L., and Canagaratna, M. R.: Evaluation of composition-dependent collection efficiencies for the aerodyne aerosol mass spectrometer using field data, Aerosol Science and Technology, 46, 258-271, 2012.

Ng, N. L., Canagaratna, M. R., Zhang, Q., Jimenez, J. L., Tian, J., Ulbrich, I. M., Kroll, J. H., Docherty, K. S., Chhabra, P. S., Bahreini, R., Murphy, S. M., Seinfeld, J. H., Hildebrandt, L., Donahue, N. M., DeCarlo, P. F., Lanz, V. A., Prévôt, A. S. H., Dinar, E., Rudich, Y., and Worsnop, D. R.: Organic aerosol components observed in Northern Hemispheric datasets from Aerosol Mass Spectrometry, Atmospheric Chemistry and Physics, 10, 4625-4641, 10.5194/acp-10-4625-2010, 2010.

Ng, N. L., Canagaratna, M. R., Jimenez, J. L., Chhabra, P. S., Seinfeld, J. H., and Worsnop, D. R.: Changes in organic aerosol composition with aging inferred from aerosol mass spectra, Atmospheric Chemistry and Physics, 11, 6465-6474, 10.5194/acp-11-6465-2011, 2011a.

Ng, N. L., Canagaratna, M. R., Jimenez, J. L., Zhang, Q., Ulbrich, I. M., and Worsnop, D. R.: Real-Time Methods for Estimating Organic Component Mass Concentrations from Aerosol Mass Spectrometer Data, Environmental Science & Technology, 45, 910-916, 10.1021/es102951k, 2011b.

Ng, N. L., Herndon, S. C., Trimborn, A., Canagaratna, M. R., Croteau, P. L., Onasch, T. B., Sueper, D., Worsnop, D. R., Zhang, Q., Sun, Y. L., and Jayne, J. T.: An Aerosol Chemical Speciation Monitor (ACSM) for Routine Monitoring of the Composition and Mass Concentrations of Ambient Aerosol, Aerosol Science and Technology, 45, 780-794, 2011c.

Norris, G., Vedantham, R., Wade, K., Brown, S., Prouty, J., and Foley, C.: EPA Positive Matrix Factorization (PMF) 3.0 Fundamentals & user guide, 2008.

Paatero, P., and Tapper, U.: Analysis of different modes of factor analysis as least squares fit problems, Chemometrics and Intelligent Laboratory Systems, 18, 183-194, https://doi.org/10.1016/0169-7439(93)80055-M, 1993.

Paatero, P., and Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5, 111-126, 10.1002/env.3170050203, 1994.

Paatero, P.: Least squares formulation of robust non-negative factor analysis, Chemometrics and Intelligent Laboratory Systems, 37, 23-35, https://doi.org/10.1016/S0169-7439(96)00044-5, 1997.

Paatero, P., and Hopke, P. K.: Discarding or downweighting high-noise variables in factor analytic models, Analytica Chimica Acta, 490, 277-289, https://doi.org/10.1016/S0003-2670(02)01643-4, 2003.

Paatero, P.: User's guide for positive matrix factorization programs PMF2 and PMF, University of Helsinki, Helsinki, Finland, 2004.

Paatero, P., and Hopke, P. K.: Rotational tools for factor analytic models, Journal of Chemometrics, 23, 91-100, 10.1002/cem.1197, 2009.

Paatero, P., Eberly, S., Brown, S. G., and Norris, G. A.: Methods for estimating uncertainty in factor analytic solutions, Atmospheric Measurement Techniques, 7, 781-797, 10.5194/amt-7-781-2014, 2014.

Parworth, C., Fast, J., Mei, F., Shippert, T., Sivaraman, C., Tilp, A., Watson, T., and Zhang, Q.: Long-term measurements of submicrometer aerosol chemistry at the Southern Great Plains (SGP) using an Aerosol Chemical Speciation Monitor (ACSM), Atmospheric Environment, 106, 43-55, https://doi.org/10.1016/j.atmosenv.2015.01.060, 2015.

Patokoski, J., Ruuskanen, T. M., Kajos, M. K., Taipale, R., Rantala, P., Aalto, J., Ryyppö, T., Nieminen, T., Hakola, H., and Rinne, J.: Sources of long-lived atmospheric VOCs at the rural boreal forest site, SMEAR II, Atmospheric Chemistry and Physics, 15, 13413-13432, 2015.

Peräkylä, O., Vogt, M., Tikkanen, O.-P., Laurila, T., Kajos, M. K., Rantala, P. A., Patokoski, J., Aalto, J., Yli-Juuti, T., Ehn, M., Sipilä, M., Paasonen, P., Rissanen, M., Nieminen, T., Taipale, R., Keronen, P., Lappalainen, H. K., Ruuskanen, T. M., Rinne, J., Kerminen, V.-M., Kulmala, M., Bäck, J., and Petäjä, T.: Monoterpenes' oxidation capacity and rate over a boreal forest, Boreal Environment Research, 19, 293-310, 2014.

Petit, J.-E., Favez, O., Albinet, A., and Canonaco, F.: A user-friendly tool for comprehensive evaluation of the geographical origins of atmospheric pollution: Wind and trajectory analyses, Environmental modelling and software, 88, 183-187, 2017.

Pieber, S. M., El Haddad, I., Slowik, J. G., Canagaratna, M. R., Jayne, J. T., Platt, S. M., Bozzetti, C., Daellenbach, K. R., Fröhlich, R., Vlachou, A., Klein, F., Dommen, J., Miljevic, B., Jiménez, J. L., Worsnop, D. R., Baltensperger, U., and Prévôt, A. S. H.: Inorganic Salt Interference on CO2+ in Aerodyne AMS and ACSM Organic Aerosol Composition Studies, Environmental Science & Technology, 50, 10494-10503, 10.1021/acs.est.6b01035, 2016.

Prăvălie, R.: Major perturbations in the Earth's forest ecosystems. Possible implications for global warming, Earth-Science Reviews, 185, 544-571, 2018.

Rinne, J., Bäck, J., and Hakola, H.: Biogenic volatile organic compound emissions from the Eurasian taiga: current knowledge and future directions, Boreal Environment Research, 14, 807-826, 2009.

Riuttanen, L., Hulkkonen, M., Maso, M. D., Junninen, H., and Kulmala, M.: Trajectory analysis of atmospheric transport of fine particles, SO2, NOx and O3 to the SMEAR II station in Finland in 1996–2008, Atmospheric Chemistry and Physics, 13, 2153-2164, 2013.

Rose, C., Zha, Q., Dada, L., Yan, C., Lehtipalo, K., Junninen, H., Mazon, S. B., Jokinen, T., Sarnela, N., Sipilä, M., Petäjä, T., Kerminen, V.-M., Bianchi, F., and Kulmala, M.: Observations of biogenic ion-induced cluster formation in the atmosphere, 4, eaar5218, 10.1126/sciadv.aar5218 2018.

Rousseeuw, P. J.: Silhouettes – a Graphical Aid to the Interpretation and Validation of Cluster-Analysis, Journal of Computational and Applied Mathematics, 20, 53-65, https://doi.org/10.1016/0377-0427(87)90125-7, 1987.

Shah, V., Jaeglé, L., Jimenez, J. L., Schroder, J. C., Campuzano-Jost, P., Campos, T. L., Reeves, J. M., Stell, M., Brown, S. S., Lee, B. H., Lopez-Hilfiker, F. D., and Thornton, J. A.: Widespread Pollution From Secondary Sources of Organic Aerosols During Winter in the Northeastern United States, Geophysical Research Letters, 46, 2974-2983, 10.1029/2018gl081530, 2019.

Shrivastava, M., Cappa, C. D., Fan, J., Goldstein, A. H., Guenther, A. B., Jimenez, J. L., Kuang, C., Laskin, A., Martin, S. T., Ng, N. L., Petaja, T., Pierce, J. R., Rasch, P. J., Roldin, P., Seinfeld, J. H., Shilling, J., Smith, J. N., Thornton, J. A., Volkamer, R., Wang, J., Worsnop, D. R., Zaveri, R. A., Zelenyuk, A., and Zhang, Q.: Recent advances in understanding secondary organic aerosol: Implications for global climate forcing, Reviews of Geophysics, 55, 509-559, 10.1002/2016rg000540, 2017.

Sokal, R. R., and Sneath, P. H.: Principles of numerical taxonomy,
Principles of numerical taxonomy, Taxon, 12, 190-199, 1963.

Stefenelli, G., Pospisilova, V., Lopez-Hilfiker, F. D., Daellenbach, K. R., Hüglin, C., Tong, Y., Baltensperger, U., Prevot, A. S. H., and Slowik, J. G.: Organic aerosol source apportionment in Zurich using an extractive electrospray ionization time-of-flight mass spectrometer (EESI-TOF-MS) – Part 1: Biogenic influences and day–night chemistry in summer, Atmospheric Chemistry and Physics, 2019, 14825-14848, 2019.

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D., and Ngan, F.: NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System, Bulletin of the American Meteorological Society, 96, 2059-2077, 10.1175/BAMS-D-14-00110.1, 2016.

Stein, S. E., and Scott, D. R.: Optimization and testing of mass spectral library search algorithms for compound identification, Journal of the American Society for Mass Spectrometry, 5, 859-866, 10.1016/1044-0305(94)87009-8, 1994.

Steinhaus, H.: Sur la division des corp materiels en parties, Bulletin de l'académie polonaise des sciences, 1, 801-804, 1956.

Stolzenburg, D., Fischer, L., Vogel, A. L., Heinritzi, M., Schervish, M., Simon, M., Wagner, A. C., Dada, L., Ahonen, L. R., Amorim, A., Baccarini, A., Bauer, P. S., Baumgartner, B., Bergen, A., Bianchi, F., Breitenlechner, M., Brilke, S., Buenrostro Mazon, S., Chen, D., Dias, A., Draper, D. C., Duplissy, J., El Haddad, I., Finkenzeller, H., Frege, C., Fuchs, C., Garmash, O., Gordon, H., He, X., Helm, J., Hofbauer, V., Hoyle, C. R., Kim, C., Kirkby, J., Kontkanen, J., Kürten, A., Lampilahti, J., Lawler, M., Lehtipalo, K., Leiminger, M., Mai, H., Mathot, S., Mentler, B., Molteni, U., Nie, W., Nieminen, T., Nowak, J. B., Ojdanic, A., Onnela, A., Passananti, M., Petäjä, T., Quéléver, L. L. J., Rissanen, M. P., Sarnela, N., Schallhart, S., Tauber, C., Tomé, A., Wagner, R., Wang, M., Weitz, L., Wimmer, D., Xiao, M., Yan, C., Ye, P., Zha, Q., Baltensperger, U., Curtius, J., Dommen, J., Flagan, R. C., Kulmala, M., Smith, J. N., Worsnop, D. R., Hansel, A., Donahue, N. M., and Winkler, P. M.: Rapid growth of organic aerosol nanoparticles over a wide tropospheric temperature range, 115, 9122-9127, 10.1073/pnas.1807604115 %J Proceedings of the National Academy of Sciences, 2018.

Surratt, J. D., Chan, A. W. H., Eddingsaas, N. C., Chan, M., Loza, C. L., Kwan, A. J., Hersey, S. P., Flagan, R. C., Wennberg, P. O., and Seinfeld, J. H.: Reactive intermediates revealed in secondary organic aerosol formation from isoprene, Proceedings of the National Academy of Sciences, 107, 6640-6645, 10.1073/pnas.0911114107, 2010.

Tunved, P., Hansson, H.-C., Kerminen, V.-M., Ström, J., Maso, M. D., Lihavainen, H., Viisanen, Y., Aalto, P. P., Komppula, M., and Kulmala, M.: High Natural Aerosol Loading over Boreal Forests, Science, 312, 261-263, 10.1126/science.1123052 2006.

Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., and Jimenez, J. L.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmospheric Chemistry and Physics, 9, 2891-2918, 10.5194/acp-9-2891-2009, 2009.

Van Krevelen, D. W.: Graphical-statistical method for the study of structure and reaction processes of coal, Fuel, 24, 269-284, 1950.

Watson, J. G., Cooper, J. A., and Huntzicker, J. J.: The effective variance weighting for least squares calculations applied to the mass balance receptor model, Atmospheric Environment (1967), 18, 1347-1355, https://doi.org/10.1016/0004-6981(84)90043-X, 1984.

Williams, J., Crowley, J., Fischer, H., Harder, H., Martinez, M., Petaja, T., Rinne, J., Back, J., Boy, M., Hakala, J., Kajos, M., Keronen, P., Rantala, P., Aalto, J., Aaltonen, H., Paatero, J., Vesala, T., Hakola, H., Levula, J., Pohja, T., Herrmann, F., Auld, J., Mesarchaki, E., Song, W., Yassaa, N., Nolscher, A. C., Johnson, A. M., Custer, T., Sinha, V., Thieser, J., Pouvesle, N., Taraborrelli, D., Tang, M. J., Bozem, H., Hosaynali-Beygi, Z., Axinte, R., Oswald, R., Novelli, A., Kubistin, D., Hens, K., Javed, U., Trawny, K., Breitenberger, C., Hidalgo, P. J., Ebben, C. J., Geiger, F. M., Corrigan, A. L., Russell, L. M., Ouwersloot, H. G., Vila-Guerau De Arellano, J., Ganzeveld, L., Vogel, A., Beck, M., Bayerle, A., Kampf, C. J., Bertelmann, M., Kollner, F., Hoffmann, T., Valverde, J., Gonzalez, D., Riekkola, M.-L., Kulmala, M., and Lelieveld, J.: The summertime Boreal forest field measurement intensive (HUMPPA-COPEC-2010): an overview of meteorological and chemical influences, Atmospheric Chemistry and Physics, 2011.

Visser, S., Slowik, J. G., Furger, M., Zotter, P., Bukowiecki, N., Canonaco, F., Flechsig, U., Appel, K., Green, D. C., Tremper, A. H., Young, D. E., Williams, P. I., Allan, J. D., Coe, H., Williams, L. R., Mohr, C., Xu, L., Ng, N. L., Nemitz, E., Barlow, J. F., Halios, C. H., Fleming, Z. L., Baltensperger, U., and Prévôt, A. S. H.: Advanced source apportionment of size-resolved trace elements at multiple sites in London during winter, Atmospheric Chemistry and Physics, 15, 11291-11309, 10.5194/acp-15-11291-2015, 2015.

Yan, C., Nie, W., Äijälä, M., Rissanen, M. P., Canagaratna, M. R., Massoli, P., Junninen, H., Jokinen, T., Sarnela, N., Häme, S. A. K., Schobesberger, S., Canonaco, F., Yao, L., Prévôt, A. S. H., Petäjä, T., Kulmala, M., Sipilä, M., Worsnop, D. R., and Ehn, M.: Source characterization of highly oxidized multifunctional compounds in a boreal forest environment using positive matrix factorization, Atmospheric Chemistry and Physics, 16, 12715-12731, 10.5194/acp-16-12715-2016, 2016.

Yttri, K. E., Simpson, D., Nøjgaard, J. K., Kristensen, K., Genberg, J., Stenström, K., Swietlicki, E., Hillamo, R., Aurela, M., Bauer, H., Offenberg, J. H., Jaoui, M., Dye, C., Eckhardt, S., Burkhart, J. F., Stohl, A., and Glasius, M.: Source apportionment of the summer time carbonaceous aerosol at Nordic rural background sites, Atmospheric Chemistry and Physics, 11, 13339-13357, 10.5194/acp-11-13339-2011, 2011.

Zhang, Q., Worsnop, D. R., Canagaratna, M. R., and Jimenez, J. L.: Hydrocarbon-like and oxygenated organic aerosols in Pittsburgh: insights into sources and processes of organic aerosols, Atmospheric Chemistry and Physics, 5, 3289-3311, 10.5194/acp-5-3289-2005, 2005.

Zhang, Q., Jimenez, J. L., Canagaratna, M., Allan, J., Coe, H., Ulbrich, I., Alfarra, M., Takami, A., Middlebrook, A., Sun, Y., Dzepina, K., Dunlea, E. J., Docherty, K. S., DeCarlo, P. F., Salcedo, D., Onasch, T., Borrmann, S., Weimer, S., Demerjian, K., Williams, P., Bower, K., Bahreini, R., Cottrell, L., Griffin, R., Rautiainen, J., Sun, J. Y., Zhang, Y. M., and Worsnop, D.: Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes, Geophysical Research Letters, 34, 2007.

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Ulbrich, I. M., Ng, N. L., Worsnop, D. R., and Sun, Y.: Understanding atmospheric organic aerosols via factor analysis of aerosol mass spectrometry: a review, Analytical and bioanalytical chemistry, 401, 3045-3067, 10.1007/s00216-011-5355-y, 2011.

Zhang, Y., Favez, O., Petit, J. E., Canonaco, F., Truong, F., Bonnaire, N., Crenn, V., Amodeo, T., Prévôt, A. S. H., Sciare, J., Gros, V., and Albinet, A.: Six-year source apportionment of submicron organic aerosols from near-continuous highly time-resolved measurements at SIRTA (Paris area, France), Atmospheric Chemistry and Physics, 19, 14755-14776, 10.5194/acp-19-14755-2019, 2019.

Äijälä, M., Heikkinen, L., Fröhlich, R., Canonaco, F., Prévôt, A. S., Junninen, H., Petäjä, T., Kulmala, M., Worsnop, D., and Ehn, M.: Resolving anthropogenic aerosol pollution types–deconvolution and exploratory classification of pollution events, Atmospheric Chemistry and Physics, 17, 3165-3197, 2017.

Äijälä, M., Daellenbach, K. R., Canonaco, F., Heikkinen, L., Junninen, H., Petäjä, T., Kulmala, M., Prévôt, A. S., and Ehn, M.: Constructing a data-driven receptor model for organic and inorganic aerosol–a synthesis analysis of eight mass spectrometric data sets from a boreal forest site, Atmospheric Chemistry and Physics, 19, 3645-3672, 2019.

Appendix A: Statistical analysis workflow

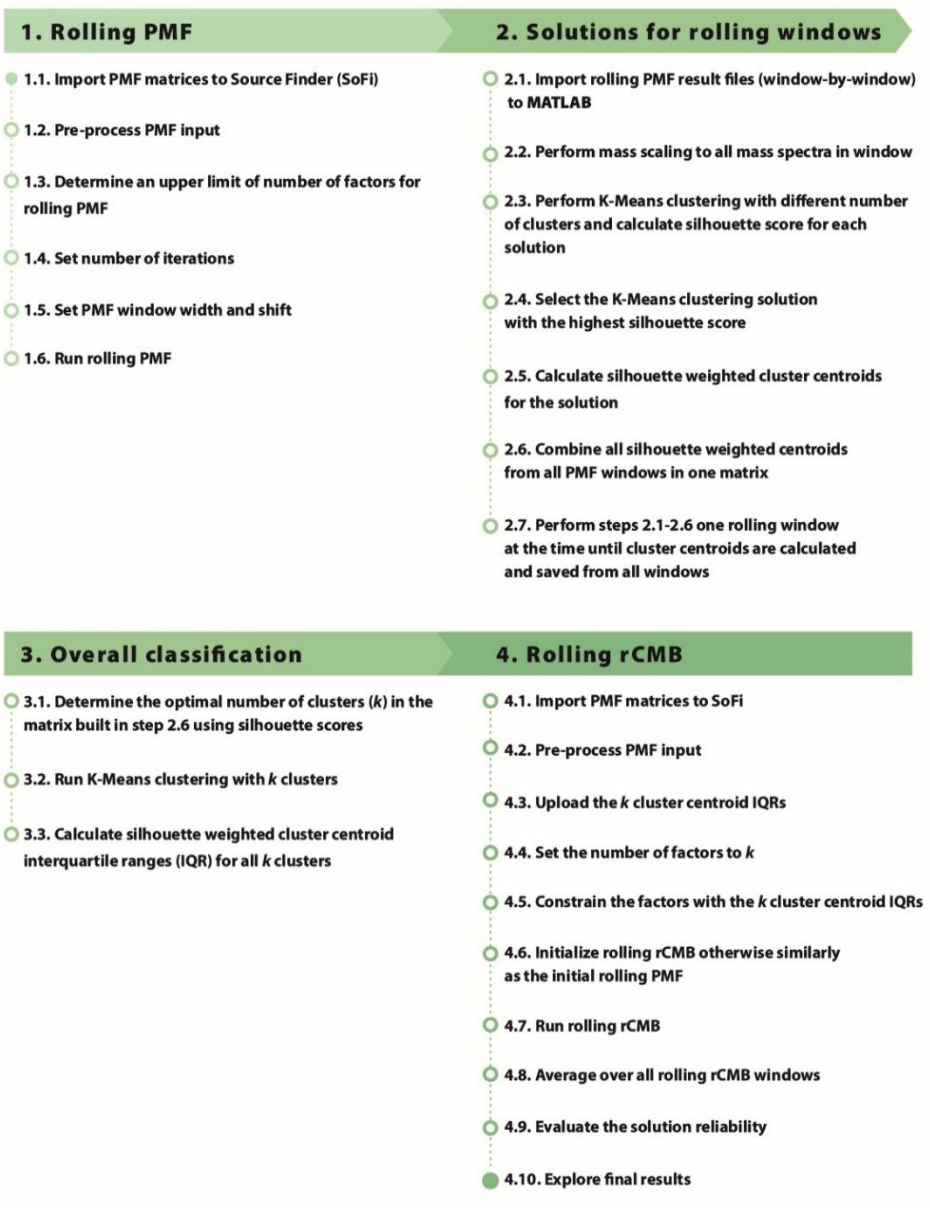

Figure A.1 Work flow describing the machine learning analysis approach utilized in the current study. In a nutshell, this method describes how K-Means clustering can be used to classify OA mass spectral profiles from a large number of unconstrained rolling PMF runs and how this information can be further utilised in a relaxed CMB run to gain insight into the OA classes' temporal behaviours. The method comprises four main phases: 1. Performing rolling PMF (Sect. 5.1), 2. Performing window-by-window (file-by-file) clustering of rolling window iterations (Phase I clustering; Sect. 5.2.1), 3. Conducting overall classification of the centroids calculated for all PMF windows (Phase II clustering; Sect. 5.2.2), and finally 4. Performing rolling relaxed- chemical mass balancing using the centroids retrieved in the previous step as CMB anchors (Sect. 5.3). Sections 4 and 5 in the paper introduce all the vocabulary needed for understanding this figure. These sections also contain detailed descriptions of each step in the method.