Anonymous Referee #1 (authors’ responses in blue)

Review for 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
by Mikko Äijälä et al.

This paper describes the development of a new way to perform source apportionment, analysing eight different mass spectrometric datasets. The topic of this paper is interesting to the community and will help on improving future source apportionment studies. I recommend this paper for publication after the authors address the following comments.

We thank the Anonymous Referee for his/her time in reviewing the manuscript and appreciate the constructive comments.

Specific comments.

Introduction. The factorization tools used in this study are PMF and ME-2. However, the authors do not mention ME-2 in the introduction. It would be good to read how ME-2 helps on separating profiles when PMF struggles to do so.

To keep the introduction short and present an introductiory storyline for the reader, we have not listed all the statistical algorithms in this section (kmeans, ME-2, similarity metrics, etc.), but approach the topic from a broader perspective of why a chemometric approach is needed in this field. To accommodate the Referee’s valid point that constraining is sometimes needed, we added a mention of this to the discussion of analyst subjective choices: *“Specifically, while analyst imposed additional constraints in factorisation may sometimes be required to reduce rotational uncertainty and extract minor factors in data (e.g. Canonaco et al., 2013; Crippa et al. 2014) such procedures are especially prone to analyst subjective decisions.”*

Page 3 line 25. I think the authors want to stress the importance of local anthropogenic sources in the last paragraph. If that is the case, rephrase the last paragraph for something like: While previous studies have found biogenic SOA and long-range transport from industrial regions to be important, local anthropogenic aerosol sources are also present. At the moment that paragraph is confusing, please rephrase it.

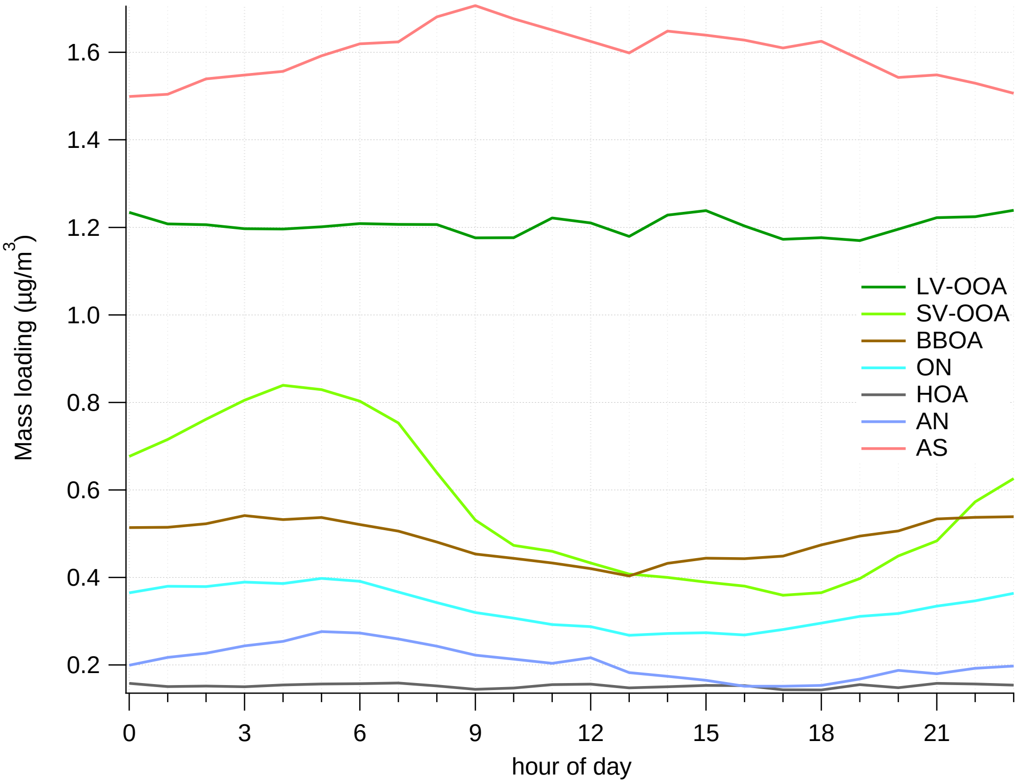
We rephrased and re-ordered the paragraph. Here we wanted to list the known anthropogenic aerosol sources, so we re-arranged them by distance (long-range -> regional -> local), to clarify the paragraph.

Section 2.3.1. When describing ME2, the method used to constrain solutions should be explained as well.

This is a valid point, which was probably not clarified enough. We now added a short explanation: *“In this study, when ME-2 constraints were applied to the factor profiles, we set upper and lower bounds for the allowed profile solutions. The bounds were based on variability estimates obtained from earlier analysis, as explained later, in Sect. 2.5.”* We also added an explanation of the ME-2 limits in 2.5: *"The allowed variabilities were constrained by setting upper and lower bounds (the estimated variability ranges from the previous phase) for factor profiles."* (See also response to Referee 2, comment #4).

Page 16 Second paragraph. When talking about BBOA and COA, one of the main differences between these factors is the diurnal profile, COA usually shows a small peak at lunch time and then increases in the evening. Do the authors had a look at diurnal profiles to differentiate between COA and BBOA? Diurnal profiles provide interesting information about the different profiles identified.

We added the diurnal analysis to the supplementary material to include this information. However, regarding the BBOA and COA, the temporal behaviour stated above only applies for urban environments or close-by sources. As described in Sect. 2.1.1, most anthropogenic aerosol (besides the plumes from the station itself, which is insignificant in terms of observed aerosol mass) is transported from 5 to 50 kilometres away. With common wind speeds of a few m/s (or 10-15 km/h) the distribution already becomes somewhat smeared, so we do not expect to see e.g. clear lunch-time peaks. We added a paragraph to 3.1.3: *"Diurnal cycles of the components for the entirety of data are available in S.I (Figure S.12). Due to the rural setting of the site and the generally long transport times of aerosol before reaching the site, diurnal cycles for the various aerosol types are not as characteristic as they would be for urban measurements (for e.g. temporal trends of HOA and BBOA). Also due to seasonal differences, the variability between data sets is considerable, resulting in high uncertainty in interpretation. The daily cycles are likely a mixed product of source emissions, boundary layer dynamics and aerosol temperature response. While of interest, disentangling these processes is beyond the topic of this study."*



Technical corrections

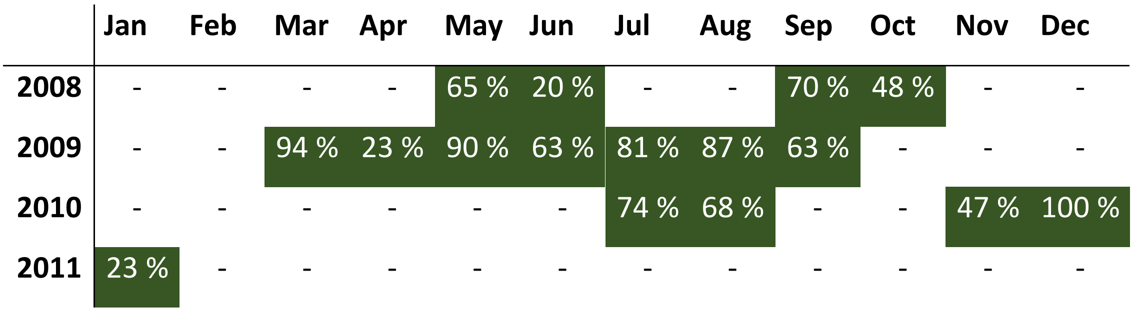
A number of typos were found in the manuscript. I suggest to go through the document again and correct the typos. These are a few minor comments I would like to provide.

Page 2 line 2. Change effects for properties
Changed

Page 2 line 3. Change almost for near.
Changed (p.2,l.5)

Page 3 line 2. Provide the references to the previous literature.
This is just an opening remark that we will not cover all the technical details in this article. The references are included in the relevant sections. We added:
"[Our instrumentation, data processing, measurement site and analysis algorithms have been conscientiously described in previous literature,] to which we refer in the corresponding sections."

Page 4 line 2. Delete the word "to" before 2008. Page 4 table 1. Perhaps add a column with the number of months for an easier comparison.
Corrected. Good suggestion. We added a graphical table indicating monthly data availability to Sect. 2.1.2.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2008	-	-	-	-	65 %	20 %	-	-	70 %	48 %	-	-
2009	-	-	94 %	23 %	90 %	63 %	81 %	87 %	63 %	-	-	-
2010	-	-	-	-	-	-	74 %	68 %	-	-	47 %	100 %
2011	23 %	-	-	-	-	-	-	-	-	-	-	-

Page 4 line 10. Please define if it was a compact or a high resolution AMS.
This information is already in the next sentence: "AMS instruments in general have been described by Canagaratna et al. (2007), and the compact ToF analyser version (CToF) used in this study by Drewnick et al. (2005)" (p.4.,l.12)

Page 6 line 30. Provide references where ME-2 has been used to partially constrain solutions.
We added here: *"[…] allowed to vary within narrow limits (derived from variability estimates; see Sect 2.5). Variability estimate of the final model is available in S.I (Figure S.13)."*

Page 9 line 4. Change: 'There exist' for 'There are'
We do not see any difference in meaning or clarity here, but rephrased the sentence for fluency:
"A variety of aerosol inorganic equilibrium models exist, and are typically used as modules…"

Page 21. "F57:f57 fractions", it should be f55:f57.
Corrected.

Anonymous Referee #2 **(authors' responses in blue)**

The manuscript "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" introduces a novel receptor model for organic and inorganic aerosol measured in Hyytiälä, Finland between 2008 and 2010. The measurements were performed with a CToF aerosol mass spectrometer and receptor model was applied to unit-mass resolution mass spectra. The benefit of this receptor model for organic and inorganic aerosol over traditional PMF/ME2 for organic MS is that it yields useful information for the modelling of submicron atmospheric aerosols physical and chemical properties, and the results illuminates the division between organic and inorganic aerosol types and dynamics of salt formation in aerosol.

General comments

This paper is written precisely, logically and clearly. It presents novel methodology and scientific results that are very useful for atmospheric scientists, both experimentalists and modelers. I think this paper should be published in ACP after minor revision.

We thank the Anonymous Referee for his/her time in reviewing the manuscript and appreciate the constructive comments.

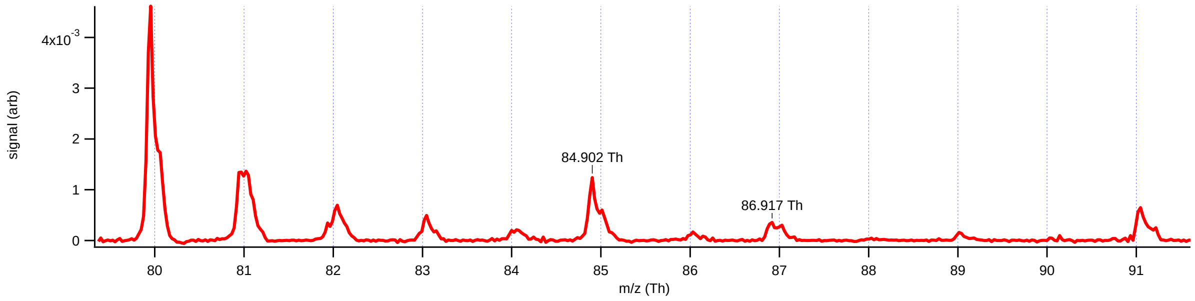
Specific comments

1. Page 1, Abstract; line 24; "simplistic inorganics apportionment methods" is unclear to me, do you mean PMF/ME2?

Specifically this refers to ion balance schemes described in 2.4 and the comparison discussed in 3.3.1. We changed this to: *"Compared to traditional, ion balance based inorganics apportionment schemes for aerosol mass spectrometer data, […]"*

2. Page 4, CToF-AMS; what is the mass resolving power of CToF? Is it possible to use high resolution mass spectra instead of UMR-MS? How reliably you can identify alkalimetals, especially rubidium, with CToF?

Mass resolution for our instrument was around 500 (*m/dm*), so it's not strictly high resolution capable. However, the resolution is enough to confirm presence of ions when distance between peaks is high enough, such as for some inorganics and metals with clearly negative mass defect. *m/z* calibrated mass spectrum below for *m/z* 80 to 91 Th is shown below. In addition to the natural isotopic ratio mentioned in 3.4.1, the exact masses for ^{85}Rb (84.912 a.m.u.) and ^{87}Rb (86.909 a.m.u.) correspond to the measured spectrum, supporting the Rubidium interpretation. We added the figure to the supplementary material (Fig. S.11) and a reference to Sect 3.4.2.

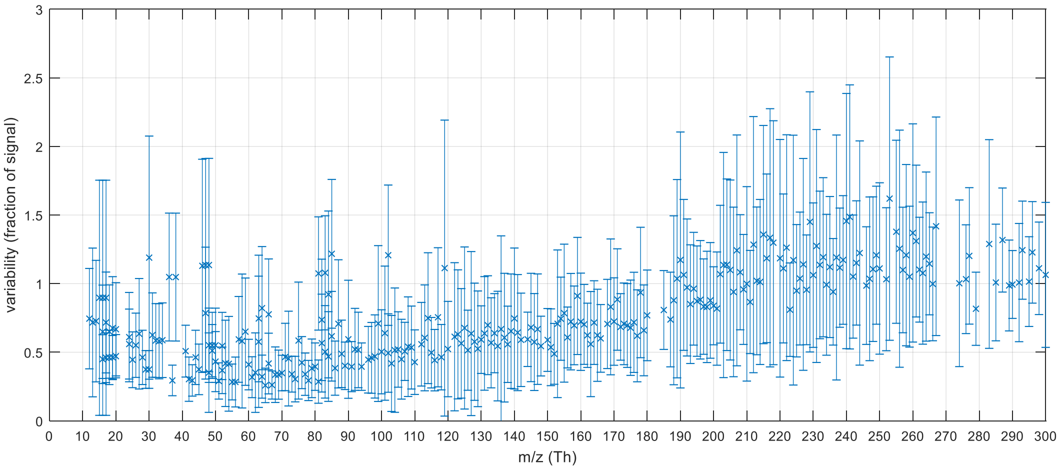


3. Page 5, Section 2.2.2 Data preparation and down-weighting; Could you explain here (shortly) how RIEs and fragmentation table were taken into account?

RIE and fragtable are (shortly) explained in the previous section (Sect 2.2.1; “The aerosol mass spectrometer (AMS) instrument and basic data processing”): *“The per-amu (atomic mass unit) analyser signal is subsequently quantified based on instrument response calibrations and corrections (among others the correction for relative ionisation efficiency between the species; RIE; Allan et al., 2004); supplementary information Sect S.4). Individual, unit-mass-resolution amu signals are then chemically speciated, based on chemical information on fragmentation and air composition (see Allan et al., 2003b), for details)”.* To maintain consistency of the level of technical specifics discussed in the article, we would prefer to avoid very detailed explanations on individual technical topics such as (R)IE calibration/correction or the inner workings of the fragtable. For an interested reader, a detailed description on these topics can be found in the Allan *et al.* papers, as referenced.

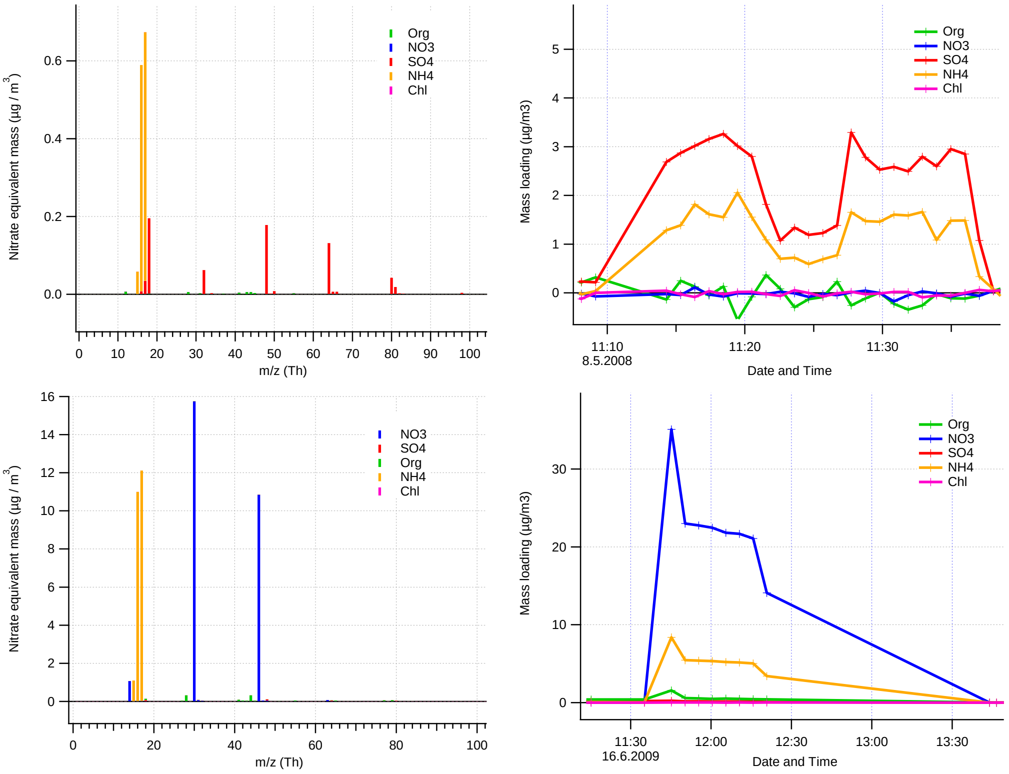
4. Page 6, line 28-29, “all the source profiles are constrained, but allowed to vary within narrow limits”, what alpha-value (constraning value)?

See similar comment by Anonymous Referee #1. We added here:
“[…] allowed to vary within narrow limits (derived from variability estimates; see Sect 2.5)”. We also added a sentence on this to 2.5, and added a figure to Supplementary material depicting the final variability estimate (P-III; Figure S.13).



5. Page 11, line 1; “we applied an ion ratio Rcalib = 0.42, taken as the average of mass spectrum based AN calibrations (S.I Sect S.6)”. Do you mean m/z 46/30 in Fig. S3? It seems to be much larger than 0.42. In general, Figure S3 is very difficult to read because it is unclear and has very small fonts. Please improve the quality of the figure.

The value 0.42 is from an older IE calibration analysis we did not have available in full for this work. As stated in the caption, the MS-mode calibration data in Figure S.3. is unprocessed (meaning incomplete in terms of fragtable corrections for *m/z* 30 Th etc., as it was omitted from the main analysis), so it was not used quantitatively. We agree with the Referee that the quality of Figure S.3. in supporting material was poor. The figure was re-drawn.



6. Page 15, line 16-18; similar comment, for cluster #8 m/z 46/30=0.44, according to Figure 3 the ratio is larger

The ratio mentioned in text (0.44) is taken from the final (P-III) result (Figure 4), as are all similar diagnostics values. We tried to clearly state this in the beginning of results section (Sect. 3., p.13): *“For easier comparability, all ratios and fractions of signals presented in the following sections are similarly calculated from the corresponding final spectra (P-III).”* We are aware that our decision to report all diagnostics as values from the final model (P-III) may be a source of some confusion. However, given the multi-phase methodology, it would likely be still more confusing to report all the values from different phases / solutions. The purpose of the values is also to provide values for reference for future studies, so we think reporting (only) P-III values is the clearest (or at least the least confusing) way. In order to further decrease the risk of confusion, we added a notice after the p-15 l.16-18 sentence: *“We note once more that these characteristic values for clusters are from the final model (P-III; Figure 4), as outlined before.”*

7. Page 30, line 12, what is the origin of KNO3 in Hyytiälä?

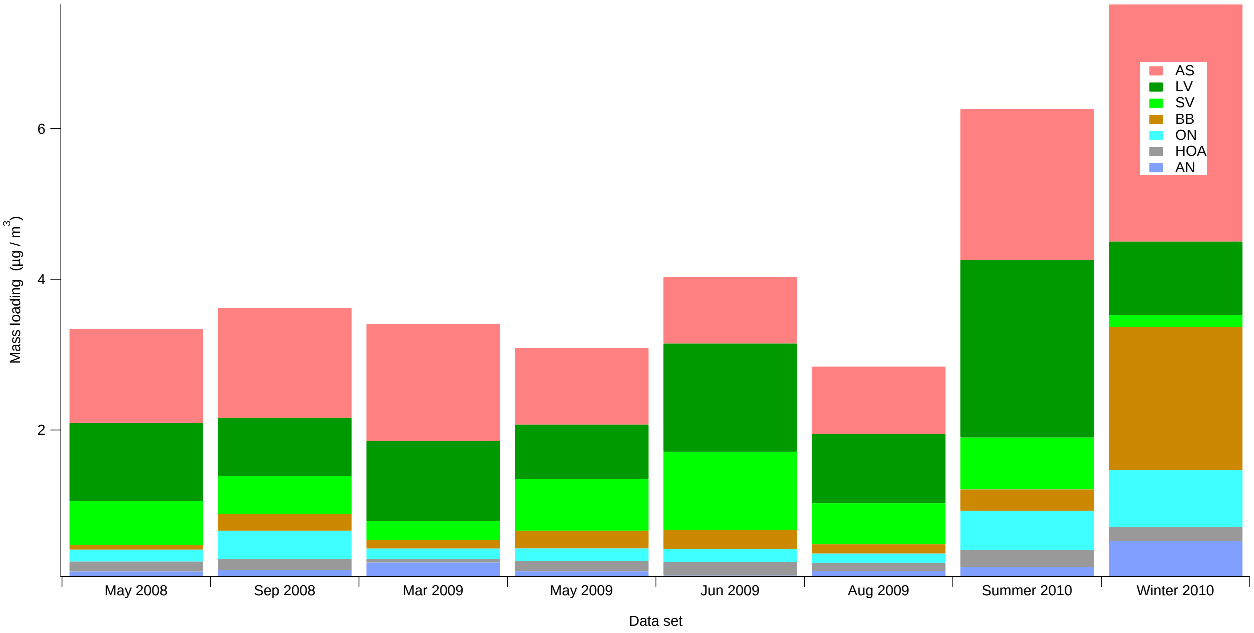
This has not been investigated for Hyytiälä specifically. We assume the main source of submicron K for Hyytiälä be to be biomass burning. e.g. (Li et al., 2003). We added this mention to Sect 3.2.2 (p.23, l.21).

8. Results and discussion in general: diurnal trends of the factors are not utilized at all when interpreting and identifying aerosol components, and similarly, any auxiliary gas (or particle) data is not exploited. Couldn't this additional data help interpreting the results?

We added a figure of the components' diurnal behaviour to supplementary material. However, due to the varying conditions of long-distance transport of aerosol to the site, the diurnal profiles are not as useful as for more urban sites. Thus, the uncertainties of the information in diurnal profiles are high. There is also plenty of auxiliary data available, and it is true this data is not fully exploited in this work. However, one of the main points of this work is to base most of the analysis on statistical diagnostics and machine learning methods, and focus less on auxiliary data and traditional PMF evaluation criteria, such as correlations with trace gases and examining diurnal cycles. We find that identifying the main components is rather unambiguous from the spectral similarities alone, and a detailed examination / interpretation of the outliers is somewhat outside of the scope of this work due to the already large volume of data and methods presented. We do agree there is a lot of room for a more detailed study of the auxiliary data, diurnal cycles and more in-depth interpretation of the chemical processes and source attribution etc., and would encourage taking this up as a topic in future studies. We added the following to Sect. 3.1.3: *"Diurnal cycles of the components for the entirety of data are available in S.I (Figure S.12). Due to the rural setting of the site and the generally long transport times of aerosol before reaching the site, diurnal cycles for the various aerosol types are not as characteristic as they would be for urban measurements (for e.g. temporal trends of HOA and BBOA). Also due to seasonal differences, the variability between data sets is considerable, resulting in high uncertainty in interpretation. The daily cycles are likely a mixed product of source emissions, boundary layer dynamics and aerosol temperature response. While of interest, disentangling these processes is beyond the topic of this study."*

9. Supplemental material, Figure S5; could you add total mass for each data set?

Good suggestion. As the campaign specific absolute mass loadings may be of interest, we added a panel to Figure S.5. with this information.



Technical corrections:

10. Page 11, line 18-19, second parenthesis is missing
Corrected.

11. Page 30, line 11; non-quantitative
Corrected.

References

Li, J., Pósfai, M., Hobbs, P. V., and Buseck, P. R. J. J. o. G. R. A.: Individual aerosol particles from biomass burning in southern Africa: 2, Compositions and aging of inorganic particles, 108, 2003.

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

Mikko Äijälä[1], Kaspar R. Daellenbach[1], Francesco Canonaco[2], Liine Heikkinen[1], Heikki Junninen[1,3], Tuukka Petäjä[1], Markku Kulmala[1], André S.H. Prévôt[2], and Mikael Ehn[1]

[1] Institute for Atmospheric and Earth System Research / Physics, University of Helsinki, Helsinki, Finland
[2] Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, Villigen, Switzerland
[3] Laboratory of Environmental Physics, University of Tartu, Tartu, Estonia

Correspondence to: Mikael Ehn (mikael.ehn@helsinki.fi)

Abstract. The interactions between organic and inorganic aerosol chemical components are integral to understanding and modelling climate and health-relevant aerosol physicochemical properties, such as volatility, hygroscopicity, light scattering and toxicity. This study presents a synthesis analysis for eight data sets, of non-refractory aerosol composition, measured at a boreal forest site. The measurements, performed with an aerosol mass spectrometer, cover in total around 9 months over the course of 3 years. In our statistical analysis, we use the complete organic and inorganic unit-resolution mass spectra, as opposed to the more common approach of only including the organic fraction. The analysis is based on iterative, combined use of (1) data reduction, (2) classification and (3) scaling tools, producing a data-driven chemical mass balance type of model capable of describing site-specific aerosol composition. The receptor model we constructed was able to explain 83 ± 8 % of variation in data, increased to 96 ± 3 % when signals from low signal-to-noise variables were not considered. The resulting interpretation of an extensive set of aerosol mass spectrometric data infers seven distinct aerosol chemical components for a rural boreal forest site: ammonium sulphate (35 ± 7 % of mass), low and semi-volatile oxidised organic aerosols (27 ± 8 % and 12 ± 7 %), biomass burning organic aerosol (11 ± 7 %), a nitrate containing organic aerosol type (7 ± 2 %), ammonium nitrate (5 ± 2 %), and hydrocarbon-like organic aerosol (3 ± 1%). Some of the additionally observed, rare outlier aerosol types likely emerge due to surface ionisation effects, and likely represent amine compounds from an unknown source and alkaline metals from emissions of a nearby district heating plant. Compared to traditional,~~, simplistic inorganics apportionment methods~~ ion balance based inorganics apportionment schemes for aerosol mass spectrometer data, our statistics-based method provides an improved, more robust approach, yielding readily useful information for the modelling of submicron atmospheric aerosols physical and chemical properties. The results also shed light on the division between organic and inorganic aerosol types and dynamics of salt formation in aerosol. Equally importantly, the combined methodology exemplifies an iterative analysis, using consequent analysis steps by a combination of statistical methods. Such an approach offers new ways to home in on physicochemically sensible solutions with minimal need for a priori information or analyst interference. We therefore suggest

that similar statistics-based approaches offer significant potential for un/semi supervised machine-learning applications in future analyses of aerosol mass spectrometric data.

1 Introduction

Along with particle size, aerosol chemical composition is fundamental in understanding aerosol physicochemical ~~effects~~ properties such as hygroscopicity, volatility, optics and toxicity (Bilde et al., 2015; Swietlicki et al., 2008; Zimmermann, 2015). In the past decade aerosol mass spectrometry has provided a way to quantitatively resolve basic chemical composition of aerosol in ~~almost~~near real-time. This not only enables basic chemical speciation into organic and common inorganic ion species, but also produces a wealth of complex mass spectrometric data. It has since become clear that these data sets, although superficially hard-to-interpret, are rich in chemical information and their statistical analysis yields considerable new knowledge. However, tapping into this information source requires use of advanced ~~statistical~~analysis tools and chemometric methods (i.e. “using mathematical and statistical methods to provide maximum chemical information by analyzing chemical data”; Kowalski, 1975). Consequently, advanced statistical methods for data reduction have quickly gained traction in aerosol mass spectrometry, and are presently widely used for deconvolution of complex, organic mass spectra into their underlying components. Specifically, the Positive Matrix Factorization algorithm (PMF;~~,~~ Paatero and Tapper, 1994) has achieved a predominant status as the state-of-the-art analysis tool for deconvolving aerosol mass spectrometric data. Factorisation methods such as PMF notably allow for the condensation of information found in high-dimension data matrices into a manageable number of factors, corresponding to e.g. aerosol chemical species, sources or processes. Data reduction often additionally allows for improved visualisation, aiding in interpretation of the underlying aerosol chemical phenomena.

In exploratory factor analysis, the principal difficulties often relate to deciding the optimal number of factors, choosing between multiple solutions of mathematically similar quality, and estimating the reliability and uncertainty of the results. Lacking robust but easy-to-use mathematical tools, the selection and interpretation of factorisation solutions remains prone to subjective bias by the analyst. Specifically, while analyst imposed additional constraints in factorisation may sometimes be required to reduce rotational uncertainty and extract minor factors in data (e.g. Canonaco et al., 2013; Crippa et al., 2014) such procedures are especially prone to analyst subjective decisions. Evaluation and verification of a factorisation solution thus generally requires meticulous study and understanding of e.g. correlations with auxiliary data, temporal changes and cycles and spectral references. While statistics-driven methods for spectra comparison and classification as of yet remain marginal in aerosol mass spectrometry, they do show promise in their capability to automatically group similar spectra based on their chemically relevant features, producing comparable classifications to those performed manually by expert analysts (Äijälä et al., 2017; Rebotier and Prather, 2007; Freutel et al., 2013).

The overwhelming majority of PMF analyses to date from AMS have been performed on the organic fraction alone (Zhang et al., 2011). Contrary to popular belief, there exists no tenable reasons to limiting chemometric analysis to organic signals, as exemplified by the analyses of Sun et al. (2012) and Hao et al. (2014). Although it requires some additional data preparation and processing, inclusion of inorganics provides additional insight into e.g. salt formation in aerosol. In this work, we apply data reduction and classification methods for analysing organic and inorganic aerosol mass spectral data from several measurement campaigns in the boreal forest. We then derive a comprehensive receptor model resolving the dominant aerosol categories at the site. In addition, by presenting an example of a semi-supervised, statistics-driven analysis of large mass spectral data sets, we hope to pave the way for machine learning based data analysis approaches, reducing the need for expert analyst input and subjective judgement at each step.

2 Methods

Our instrumentation, data processing, measurement site and analysis algorithms have been conscientiously described in previous literature~~.~~, to which we refer in the corresponding sections~~-~~. Thus, we focus on the new aspects of this work, showing how the individual methods can be connected to form an analysis chain, and to exemplify how chemometric information can be propagated through it. In short, we will first cover the measurement site, SMEAR II and the sets of data available to us (Section 2.1). We then describe our mass spectrometer instrument and preparation of data (Sect. 2.2). In Section 2.3, we will briefly go through the various statistical tools and algorithms, covering the basics of data factorisation, classification of spectra using a clustering algorithm, clustering solution evaluation, and detail the pre and post-weighting involved. Section 2.4 describes typical reference methods for inorganics and nitrate apportionment: an ion balance scheme and a separate parametrisation for estimating organonitrate loading, to provide a comparison for the inorganic speciation from our statistics-based receptor model. Finally, in Section 2.5, we present a summarised, step-by-step description of how the methods were combined to produce a receptor model for aerosol composition at the measurement site.

2.1 Measurement site and collection of data

2.1.1. The SMEAR II site

The AMS data of this study was collected at the SMEAR II site (Station for Measuring Ecosystem-Atmosphere-Relationships) in Hyytiälä, Southern Finland (61°50’40”N, 24°17’013”E). The site is a well-known and equipped atmospheric research station, representing rural, background atmosphere in the boreal forest biome. The site and earlier measurements therein have been extensively described and reported in literature (e.g. Hari and Kulmala, 2005; Williams et al., 2011; Äijälä et al., 2017). The environment consists mostly of Scots pine (*Pinus Sylvestris*) dominated forests - 90 % of land in the nearest 50 km, and 94 % in the nearest 5 km is forested (Williams et al., 2011).

A large part of~~much of~~ the aerosol loading at SMEAR II is attributable to regional biogenic secondary organic aerosol (SOA; Corrigan et al., 2013; Crippa et al., 2014; Allan et al., 2006) and long-range transport from industrial regions in Southern Finland, Western Russia and Central Europe (Kulmala et al., 2000; Patokoski et al., 2015; Niemi et al., 2009; Sogacheva et al., 2005).

2.1.2 Data sets

In this study, the aerosol composition was monitored by an AMS between ~~to~~ 2008 and 2011, during several short measurement campaigns. Notable larger, intensive campaigns at the time were the EUCAARI project (2008-2009; Kulmala et al., 2009; Kulmala et al., 2011) and HUMPPA-COPEC (2010; Williams et al., 2011; Corrigan et al., 2013). The sets of data used along with their timeframes are shown in Table 1. Data availability by year and month is presented in Table 2. ~~.~~

Table 1. Data sets used in this study and their timeframes.

Data set number	Data set name	Campaign	Start time	End time
I	"May 2008"	EUCAARI	29.4.2008	8.6.2008
II	"Sep 2008"	EUCAARI	10.9.2008	15.10.2008
III	"Mar 2009"	EUCAARI	4.3.2009	6.4.2009
IV	"May 2009"		29.4.2009	28.5.2009
V	"Jun 2009"		12.6.2009	8.8.2009
VI	"Aug 2009"		13.8.2009	19.9.2009
VII	"Summer 2010"	HUMPPA-COPEC	9.7.2010	7.8.2010
VIII	"Winter 2010"		10.11.2010	7.1.2011

Table 2. Data availability. Months when AMS data was available are shown in green. Percentages indicate the fraction of days with at least one data point.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2008	-	-	-	-	65 %	20 %	-	-	70 %	48 %	-	-
2009	-	-	94 %	23 %	90 %	63 %	81 %	87 %	63 %	-	-	-
2010	-	-	-	-	-	-	74 %	68 %	-	-	47 %	100 %
2011	23 %	-	-	-	-	-	-	-	-	-	-	-

2.2 Instrumentation, data processing and preparation

2.2.1 The aerosol mass spectrometer (AMS) instrument and basic data processing

The mass spectrometric data for this study was acquired with a Time-of-Flight Aerosol Mass Spectrometer (ToF-AMS), developed by Aerodyne Research Inc. (Billerica, MA, U.S.). AMS instruments in general have been described by Canagaratna et al. (2007), and the compact ToF analyser version (CToF) used in this study by Drewnick et al. (2005). Additional, more specific details related to the specific instrument we used are available in our previous study (Äijälä et al., 2017).

In brief, the AMS instrument sucks sample aerosol from atmospheric pressure to vacuum conditions through an inlet system consisting of a critical orifice and a particle concentrating aerodynamic lens (Liu et al., 2007). The sample aerosol beam is

Formatted: Font: 10 pt, Bold, English (United States)

Formatted: Font: 10 pt, Bold, English (United States)

Formatted: Justified

Formatted: English (United States)

Formatted Table

Formatted: Font: 10 pt, Bold, English (United States)

Formatted: Font: 10 pt

Formatted: Centered

Formatted: Font: 10 pt, Bold, English (United States)

Formatted: Centered

Formatted: Font: 10 pt

Formatted: Font: 10 pt, Bold, English (United States)

Formatted: Centered

Formatted: Font: 10 pt

Formatted: Font: 10 pt, Bold, English (United States)

Formatted: Centered

Formatted: Font: 10 pt

directed at a vaporizer operated at 600 degrees Celsius, whereby flash vapourisation of non-refractory aerosol components occurs. The resulting vapour is ionized using 70 eV electron impact ionisation – a well-characterised hard ionisation technique, resulting in rather universal and predictable but highly fragmenting ionisation. Finally, the ions are led to an orthogonal extraction reflectron time-of-flight mass analyser, where the ions' mass-to-charge (m/z) ratios are measured.

The per-amu (atomic mass unit) analyser signal is subsequently quantified based on instrument response calibrations and corrections (among others the correction for relative ionisation efficiency between the species; RIE; Allan et al., 2004); supplementary information Sect S.4). Individual, unit-mass-resolution amu signals are then chemically speciated, based on chemical information on fragmentation and air composition (see Allan et al., 2003b), for details). Additional, specific minor modifications to our instrument have been discussed in our previous work (Äijälä et al., 2017).

2.2.2 Data preparation and down-weighting

After basic processing, the data was further prepared, to serve as input for factorisation (described in following Section, 2.3). The organic and inorganic data and related uncertainties were extracted, and down-weighting of signals performed. The procedure for extraction and preparation of AMS organic signal and related error matrices has been described by Allan et al. (2003b) and Ulbrich and co-workers (2009).

In short, measurement points or variables with missing data were omitted and error matrices calculated, based on a function accounting for both counting statistics induced uncertainty as well as background noise from the detector and electronics. The signals were then down-weighted by multiplying the error matrix conveyed uncertainty values for low signal-to-noise ratio (SNR) variables with a scalar: "weak" variables (SNR < 3) were down-weighted by a factor of 2 and "bad" variables (SNR < 1) by 10. The procedure for inorganics (SO_4, NO_3, NH_4, Chl; i.e. sulphates, nitrates, ammonia and chloride species) was similar to that used for the organics ("org"), including for the down-weighting of signals derived from fragmentation calculations. Analogous to the basic procedure of down-weighting "duplicate information" organic signals, e.g. those derived from *m/z* 44 Th (mainly CO_2^+), similarly derived inorganic signal weights were normalized so that their weight of the original plus "duplicate" signals equalled that of the original signal. Finally, the matrices for all the ion species (org, SO_4, NO_3, NH_4, Chl; in nitrate equivalent mass), were combined to form the final input matrices for factorisation, while retaining speciation information in the ion indexing.

2.3 Statistical methods and algorithms

2.3.1 Positive matrix factorixation

For factorisation, we used the Positive Matrix Factorisation (PMF) model developed by P. Paatero and colleagues (Paatero, 1997, 1999; Paatero and Tapper, 1994), and widely used for analysis of AMS data since 2007 (Lanz et al., 2007b; Zhang et al., 2011). In brief, PMF is a statistical model, typically resolving a bilinear linear combination of factor profiles (*G*) and time series (*F*) best describing the measured data matrix (*X*; Equation 1). The residual matrix *E* then denotes the portion of data left unexplained by the model (i.e. residual). PMF model is thus formulated:

$$\boldsymbol{X}_{(t\times v)} = \boldsymbol{G}_{(t\times f)} \times \boldsymbol{F}_{(f\times v)} + \boldsymbol{E}_{(t\times v)}. \qquad \textbf{(1)}$$

The brackets indicate matrix dimensions, with *v* denoting number of variables, *t* the number of time points, and *f* the number of factors. As shown in Equation 1, the model can be solved for any *f* (< *v*, *t*), requiring it to be selected by the analyst.

The main features setting PMF apart of other similar factorisation models, and making it particularly suitable for atmospheric aerosol models, are on one hand the limitation of factor profiles and time series to positive values and hence drastically reducing the amount of rotational ambiguity. On the other hand, the improved error model where the quantity to minimise is the weighted (typically the measurement uncertainty) residual, resulting in higher weight for the variables with better SNR. In PMF, the minimum weighted residual is solved using one of the related algorithms, i.e. PMF2 or Multilinear Engine 2 (ME2; Paatero, 1999). Of the two algorithms, ME2 can take in additional equations defined by the user, i.e. constraints the solutions need to adhere to. In this study, when ME-2 constraints were applied the factor profiles, we set upper and lower bounds for the allowed profile solutions. The bounds were based on variability estimates obtained from earlier analysis, as explained later, in Sect. 2.5. Variability estimate of the final model is available in S.I (Figure S.13). For running the PMF and ME2 algorithms, we used the Igor Pro (Wavemetrics Inc.) based SoFi (v. 4.8) user interface developed by F. Canonaco and co-workers at Paul Scherrer Institute (PSI). The interface allows input of the pre-processed data and user selected parameters, and calls on the solver algorithms (PMF2 or ME2, depending on assignment) to return a solution to be displayed and analysed in SoFi (Canonaco et al., 2013).

When PMF is used as a standalone method for source attribution, the selection of solution needs to be carefully validated. Sensitivities towards different number of factors, rotations and initialisation seeds are meticulously analysed, and correlations with auxiliary data computed. A case is then made for why the selection is the best possible. Contrarily, in our analysis approach, we *do not* claim to arrive at optimum solutions for *individual* PMF/ME-2 runs. Instead, we rely on multitude of data de-convolution runs to uncover the main structures in the ensemble of all data sets, and use statistical classification methods to evaluate the general outlook and commonalities between PMF/ME-2 factors at each analysis phase. As discussed in Section 2.5, this trade-off instead enables us to concentrate on best modelling the entirety *of all data sets*.

2.3.2 Relaxed chemical mass balance model

To harmonise the description of aerosol components, we constructed a constrained receptor model, where all the profile components were constrained. For this purpose we applied a ME-2 based, chemical mass balance (CMB) type of model. CMB models are typically used as receptor models for cases where source profiles are known, and only the mass loading information needs resolving (Friedlander, 1973; Gordon, 1988; Hopke, 1991; Miller et al., 1972;Hopke, 2016). In such mass conservation-based models, the observed loadings are modelled as a sum of multiple individual sources. Although CMB is mathematically often presented as sum of loadings (supplementary information, hereafter also S.I; Sect S.1, Eq. S.1), it can also be thought of as a special case of the bi-linear model described in Equation 1. Only now the profile matrix (*F*) is assumed fixed, simplifying the problem to resolving the loading matrix (*G*) which minimises the residual (*E*). CMB can be run using the SoFi interface, using the same ME-2 solver as for PMF and ME-2 applications (Canonaco et al., 2013).

In this work, we use a relaxed CMB–like bilinear model (henceforth abbreviated as r-CMB), where all the source profiles are constrained, but allowed to vary within narrow limits (derived from variability estimates; see Sect 2.5; S.I. Figure S.13). In strict technical terms this approach could be labelled “an extremely constrained ME-2 model”, but we choose to use the term “relaxed CMB” to differentiate between the typical use of ME-2 or constraining only part of the profiles, which allows the model considerably more freedom. We regard our use of the model is much closer to the idea of constraining all profiles than (semi-)exploratory factorisation typical for ME-2. The naming also serves to better highlight the conceptual differences between models in the different analysis phases.

Generally, the biggest problems of the CMB models relate to the selection of source profiles, typically from spectral libraries, and handling of their uncertainty. In our use, the anchor spectra as well as the limits for their allowed variabilities are experimentally derived from data, alleviating some of these typical concerns.

2.3.3 k-means clustering

For spectra classification, we selected the k-means algorithm, specifically because in our previous tests it was successful in classifying similar spectral data. The earlier tests additionally yielded of useful information on selection of the dissimilarity metric, as well as algorithm initialisation types and data weighting (Äijälä et al., 2017). K-means (e.g. Ball and Hall, 1965; MacQueen, 1967; Steinhaus, 1956; Jain, 2010)) is a rather simple, iterative algorithm that partitions a group of objects to a predesignated number of groups or ‘clusters’ based on their relative distances (i.e. dissimilarities). For each iteration, the algorithm assigns all objects to their closest centroids, which are then re-calculated from the mean variable values of the objects in the updated cluster. The aim is to minimize the within-cluster sum of distance (variance) (*J*) between the objects’ (C_n) locations (x_i) and the cluster centroid μ_n they are assigned to (Eq. 2):

$$\boldsymbol{J(C_n) = \sum_{x \in C_n} \|x_i - \mu_n\|^2}. \qquad \textbf{(2)}$$

The k-means algorithm iteratively converges on (any) minimum of total *J (C)* obtained by summing over all objects $C_{n.}$ To increase chances of finding a global minimum, repetitions using different initialisations are used. Specifically, we used the improved stepwise initialisation ‘kmeans++’ (Arthur and Vassilvitskii, 2007); available in e.g. Matlab v. 2017a; Math Works Inc., Natick, MA. U.S.).

2.3.4 Spectral similarity and mass scaling

Based on our earlier metric comparison (Äijälä et al., 2017), we used (Pearson) correlation as a metric for spectral dissimilarity (or “distance”, d; Fortier and Solomon, 1966;Mcquitty, 1966):

$$\boldsymbol{d(u,v) = 1 - \frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{\sqrt{(\sum_{i=1}^{n}(u_i-\bar{u})^2}\sqrt{(\sum_{i=1}^{n}(v_i-\bar{v})^2}}}, \quad \textbf{(3)}$$

where u and v are the spectra in vector form, with *m/z* variables as vector components. $\bar{u}$ and $\bar{v}$ are the arithmetic mean values of *u* and *v*.

In clustering mass spectra, data weighting is often applied. Based on previous tests (Äijälä et al., 2017), we applied mass scaling of variables, advocated by Stein and Scott and others (Stein and Scott, 1994; Kim et al., 2012; Horai et al., 2010), giving additional emphasis to higher mass signals. This common practice is based on the idea that higher mass fragment ions are more indicative of their parent ions, and thus the original characteristic composition, while smaller fragments can be produced from a wider variety of molecular fragmentation events. In mass scaling the weighted variables ($\hat{x}$) are calculated by multiplying the original variables (*x*) by mass-to-charge-specific weights (*w*), as presented in Equation 4.

$$\hat{x}_{m/z} = x_{m/z} \times w_{m/z} \; ; \; w_{m/z} = (m/z)^{\mathbf{s_m}}, \quad \textbf{(4)}$$

where the scaling factor s_m was optimised for each classification separately (SI; Sect. S.2).

2.3.5 Silhouette metric and post-weighting

The optimisation of mass scaling was based on silhouette metric (later also abbreviated as “silh”; (Rousseeuw, 1987), ranging between -1 to 1 and providing a straightforward, quantitative way to evaluate performance of the classification algorithm. The object-specific silhouette value s_i, defined as

$$\mathrm{s_i} = \begin{cases} 1 - \frac{\mathrm{a_i}}{\mathrm{b_i}};\ \text{for } \mathrm{a_i} < \mathrm{b_i} \\ 0;\ \text{for } \mathrm{a_i} = \mathrm{b_i} \\ \frac{\mathrm{b(i)}}{\mathrm{a(i)}} - 1 \text{ for } \mathrm{a_i} > \mathrm{b_i} \end{cases}, \quad \textbf{(5)}$$

where a_i corresponds to the mean distance to other objects in the same cluster, and b_i similarly to the mean distance to objects in the nearest neighbouring cluster. A silhouette value close to unity indicates the object is well clustered, while a value close to zero indicates the classification is uncertain, and the point is likely situated in-between two possible centroids. A negative cluster value is indicatory of possible misclassification. Silhouette values can be calculated for any single cluster as the arithmetic mean of the cluster members' silhouettes, or similarly as a mean over all objects, to evaluate the quality of the clustering solution as a whole.

In order to mitigate the k-means algorithm's known sensitivity to outliers, and to improve handling of between-cluster samples, we applied a simple post-processing to all cluster centroids and variability calculations: the centroid spectra and variabilities were calculated as *weighted* averages μ, and *weighted* standard deviations ($\hat{\sigma}^2$; Eq. 6) respectively, instead of the normal unweighted values (similar to Äijälä et al., 2017). As weights, we used the object specific silhouette values $s_i > 0$ (Eq. 5):

$$\hat{\mu} = \frac{\sum_{i=1}^{N} s_i v_i}{\sum_{i=1}^{N} s_i}; \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{N} s(v-\mu)^2}{\sum_{i=1}^{N} s_i}; \quad s_i = \max\,(s_i, 0), \qquad \textbf{(6)}$$

where v_i are the cluster member objects (spectra) This procedure down-weights likely misclassified objects (silhouette < 0) to zero, and penalises the more uncertain or questionable assignations (low silhouette) compared to the decidedly well-clustered objects (silhouette close to unity). Singleton clusters were omitted from this calculation, and their variability thus left undefined.

2.4 Standard approximations for aerosol inorganic speciation and organonitrate

2.4.1 Ion balance model for inorganics

Aerosol inorganic chemical speciation is better understood than the organic speciation, due to much lower diversity of the chemical compounds involved. A variety of aerosol inorganic equilibrium models exist, and are typically used as modules~~There exist a variety of aerosol inorganic equilibrium models, typically used as modules~~ in atmospheric meteorological and air quality models. However, performing thermodynamic equilibrium calculations is computationally demanding (e.g Fountoukis and Nenes, 2007), and requires a good deal of auxiliary data on thermodynamic conditions and chemical activities. Due to the complexity of the models and increased data needs, simpler approximations are often used in connection to AMS inorganic speciation. In the following ion-balance-scheme description, we denote the respective AMS ion species molar concentrations in square brackets (e.g. $[NH_4^+]$, $[NO_3^-]$, $[SO_4^{2-}]$)

A typical salt formation approximation used for AMS results is the Hong et al. (2017) ion pairing scheme, used in e.g. aerosol volatility and light scattering models (Hong et al., 2017; Zieger et al., 2015). The Hong et al (2017) scheme is based on similar approximation of Gysel et al. (2007), which in turn is a simplification of the more extensive model by Reilly and Wood (1969). We modified the Hong et al. scheme to additionally allow organonitrate ($orgNO_3$) and speciate any leftover $[NH_4^+]$ as its own class ("excess NH_4^+"). The full scheme is available in supplementary material (Section S.3), and a schematic description is presented in Figure 1.

Briefly, in the scheme we apply, NH_4^+ is first combined with SO_4^{2-} to form ammonium bisulphate and/or ammonium sulphate depending on the relative concentrations of $[NH_4^+]$ and $[SO_4^{2-}]$. Any leftover $[NH_4^+]$ then combines with $[NO_3^-]$, until all of $[SO_4^{2-}]$ and $[NO_3^-]$ is fully consumed in forming $(NH_4)_2SO_4$ and NH_4NO_3. After this point, any leftover $[NH_4^+]$ is considered “excess” and assigned to a separate class. For comparability with other models, any nitrate not in NH_4NO_3 is labelled organic. Despite the label, we note this class not only encompasses organonitrates, but also any NO^+ fragment signal from amines, N-containing organics and may even contain influences of other inorganic nitrate species, such as KNO_3, which are not considered separately in this simple model. Finally, since chloride loadings at the measurement site are generally negligible, neutralisation of hydrochloric acid (H_2O:HCl) was not included to keep this scheme rather simple. We note ion balance schemes depending on relative ion abundances, such as the one described here, can be sensitive to measurement uncertainties (e.g. errors in RIE values) of these ratios. The topic is further discussed in supplementary information (Sect S.4)

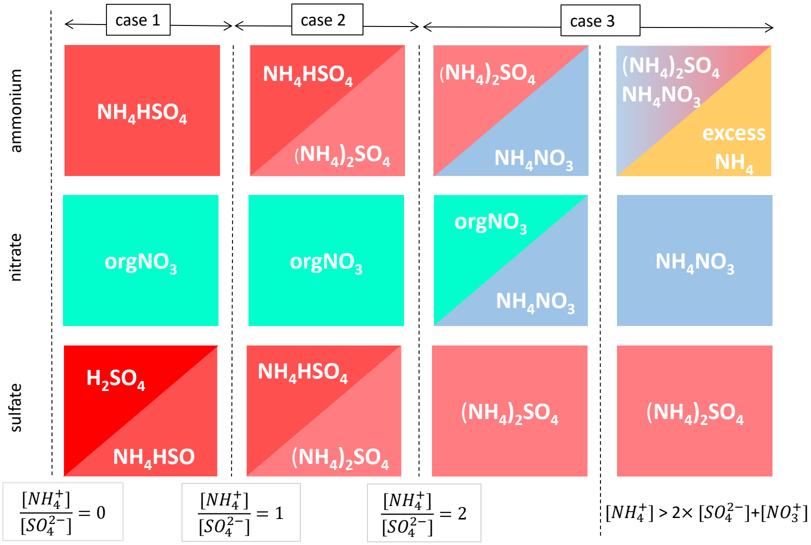


Figure 1. Schematic representation of the inorganic apportionment scheme. The scheme is divided into three cases according to the ratio of $[NH_4^+]$ to $[SO_4^{2-}]$. $[NH_4^+]$ first combines with$[SO_4]$ to form NH_4HSO_4 (Case 1), then further to $(NH_4)_2SO_4$ (Case 2). In these cases, any nitrate observed is considered organic. In Case 3 leftover $[NH_4^+]$ then associates with $[NO_3^-]$ until all the norganic anions are neutralised. Any leftover $[NH_4^+]$ is labelled as “excess NH_4^+”. Full description of the scheme is given in supplementary material (Sect S.3).

2.4.2 Kiendler-Scharr parameterisation for organonitrate

Organic nitrate estimate in the above model is very sensitive to calibration parameters (see S.I Sect S.4). Therefore, in addition to the ion-balance based scheme above, we additionally calculated a particulate organonitrate mass estimate (orgNO3$_{mass}$), based on the nitrate fragmentation ratio-based parameterisation of Kiendler-Scharr et al. (Kiendler-Scharr et al., 2016; Farmer et al., 2010):

$$orgNO3_{mass} = NO3_{total}\frac{(1+R_{orgNO3})x\ (R_{measured}-R_{calib})}{(1+R_{measured})x\ (R_{orgNO3}-R_{calib})}, \qquad \textbf{(6)}$$

where R refers to the ratio of nitrate signals at 46 and 30 Th, i.e. R = NO_3 (m/z 46 Th) : NO_3 (m/z 30 Th), for organonitrate (“orgNO_3”), NH_4NO_3 calibration (“calib”), and ambient measurement (“measured”), respectively. For the parameterisation, we applied an ion ratio R_{calib} = 0.42, taken as the average of mass spectrum based AN calibrations (S.I Sect S.6). R_{OrgNO3} value of 0.1 was used, based on the estimate by Kiendler-Scharr and co-workers for their observations on organonitrate spectral properties (Kiendler-Scharr et al., 2016).

2.5 Constructing a data-driven r-CMB receptor model

As stated in the Introduction, one of the aims of our work was to derive a robust, harmonised receptor model for the measurement site via explorative analysis. Considering the large amount of campaigns during different seasons, resulting in changing aerosol source contributions and mass spectral profiles, factorisation needed to be performed on a per-campaign (data set) basis. However, instead of performing traditional PMF complete with correlation analysis, source validation and the various sensitivity analyses separately, which would be an arduous task even for a single measurement set, we used the large amount of data sets to our advantage. Instead of optimising individual factorisations, we constructed an r-CMB model applicable to all data sets. A similar task of constructing a semi-exploratory synthesis aerosol model, albeit applying a different methodology, was undertaken and reported by Sofowote et al. (2015).

To derive the anchors and constraints for a synthesis r-CMB model, we analysed the data in three phases (P-I to P-III; Figure 2), each consisting of factorisation, classification and silhouette-based post-weighting of anchor spectra and their allowed variabilities. The allowed variabilities were constrained by setting upper and lower bounds (the estimated variability ranges from the previous phase) for factor profiles. In Phases I and II, a fixed number of 10 factors were resolved. This amount of factors was semi arbitrarily chosen, and in our case likely to be somewhat above the optimal amount for most data sets, leading to over-resolved factor solutions. However, unlike in traditional PMF analysis, we can use additional statistical diagnostics and post-processing options available to deal with potential fallout of unrealistic factor splitting (i.e. classification for identifying outliers and post-processing down-weighting or nullifying their influence). Sensitivity to initialisation seed was examined by performing all runs using 10 initialisation seeds, and generally selecting the solution with lowest normalised residual. In rare cases of a physically unrealistic solutions as the one with lowest residual (e.g. only NH_4 species in a factor), a higher residual solution was chosen instead. We conclude the solutions were generally insensitive to seed selection, especially for the factors with non-negligible mass contribution.

2.5.1 Phase I: anthropogenic aerosols

In phase I, we performed unconstrained factorisation for all the 8 data sets. With 10 factors this resulted in a total of 80 factor mass spectra. We then determined the dominant spectra classes using k-means clustering. To that purpose, we applied optimised mass scaling for improved data structure, and used silhouette diagnostics to evaluate the optimal number of clusters. We identified the known, common anthropogenic aerosol classes from the silhouette-weighted cluster centroids. This is also an approach advocated by Crippa et al. (2014) in their similar work on a synthesis analyses of several data sets.

For a cluster centroid to qualify as an anchor for further phases of our analysis, we applied the following two criteria: (1) The spectra forming the cluster were present in multiple (≥ 3) the data sets, (2) The spectra were interpretable chemically, and had adequate support from previous studies in form of literature and/or calibrations. We note that defining what constitutes as ”interpretable” or “adequate support” is inevitably an analyst (subjective) decision, so we endeavour to make our reasoning transparent in the respective discussion sections. Adhering to criterion (1) also means that factors showing up only for 1 to 2 campaigns, due to special conditions (emission, meteorology etc.), are omitted from the final r-CMB model. We will briefly cover some of the more interesting “outlier observations” in Section 3.4. At the end of Phase I, a number of constrained anchor spectra and within-cluster-variabilities were obtained. In this case, these corresponded to four anthropogenic classes, which will be discussed in more details in the results section.

2.5.2 Phase II: biogenic, secondary organic aerosols

Using the anchors and within-cluster-variabilities, we re-ran factorisation as in P-I, except now partly constrained (ME2; 4 of 10 factors constrained using anchors from P-I). In phase II, we focused on analysing the remaining free factors, likely corresponding to the biogenic, and assumedly more variable factors (Canonaco et al., 2015;Crippa et al., 2014). The procedure for classification, and the selection criteria for the (assumedly) biogenic SOA in this phase was same as in phase I.

Due to the data-driven analysis approach, specifically the constrained factors being selected from phase I, we do not expect major changes between phase I and phase II results. While arguably the methodology could be further developed to constrain the r-CMB components directly from phase I result, phase II of our analysis currently serves several purposes: 1) it should narrow down the solution space for improved description of the various SOA types, by constraining the anthropogenic, assumedly primary aerosols. 2) Compared to P-I, the allowed solutions are more similar for all data sets in P-II, which reduces the scatter of the factorisation solutions. This reduces the spectral variability (uncertainty) arising from the analysis process itself, allowing us to iteratively converge on more realistic limit values for the constraints. Ultimately, the limits should reflect the actual, natural chemical variabilities within the aerosol types. 3) Similarity of results between successive, un/semi-constrained phases allows evaluation of stability, reliability and repeatability of the method, so that it is not e.g. overly sensitive to rotational ambiguity or initialisation parameters of algorithms. This is important since the method described here is new, and its robustness needs to be demonstrated, but less so in potential later use.

2.5.3 Phase III: final, constrained receptor model

In phase III, we constructed the r-CMB receptor model. In this phase, all the factors were constrained using anchors and variabilities from the previous phase result. The number of components in the final r-CMB model, in our case 7, was equal to the total number of selected aerosol types in phase II. With these model constraints, we performed runs for each of the 8 data sets separately. Using the resulting 8×7 factor profiles, we determined the likely range of variability for the aerosol types, and calculated final, silhouette weighted reference spectra for the components by performing a final round of clustering.

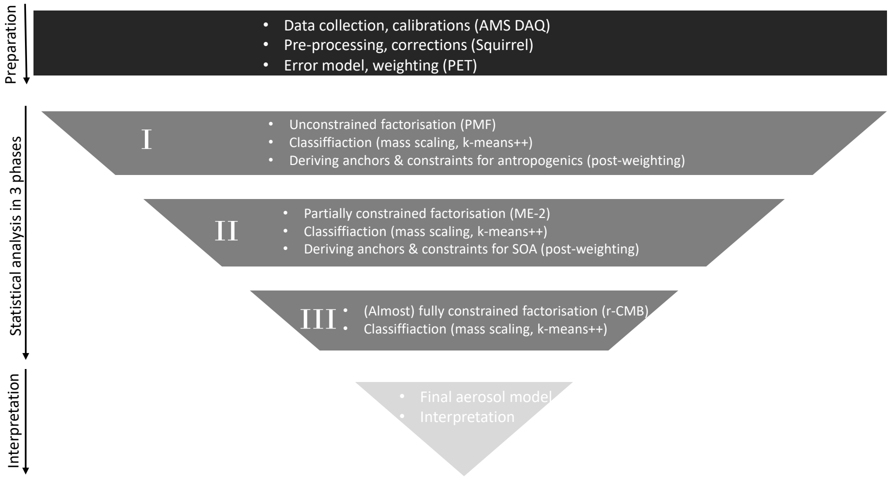


Figure 2. A flowchart illustrating the analysis using combined methodology. After initial data collection and preparation, statistical analysis is performed in three phases (P-I to P-III). Each phase limits the freedom given to factorisation from completely free (PMF) to nearly fully constrained (r-CMB). Finally, we evaluate and interpret the r-CMB model from an aerosol chemical perspective.

3. Results and discussion

In Section 3.1, we briefly describe the results from analysis phases I to III (P-I to P-III; corresponding to Sections 3.1.1 to 3.1.3), but concentrate more on the receptor model results and their interpretation (Sections 3.2). Finally we will compare our results with reference methods (Section 3.3). Comparison results are available in literature for organic aerosol components (Sect 3.3.1), and in Sect. 3.2 we will compare inorganic speciation with the alternative inorganic attribution methods, described in Methods (Sect 2.4). Finally, we briefly describe some of the outlier observations which contain potentially interesting chemical information (Section 3.4).

When interpreting and identifying aerosol components, we evaluate spectral similarity using the same similarity metric (mass scaled correlation) as for the clustering (Equations 3 and 4). We thus report mass scaled squared correlation coefficients (r_s^2) between reference spectra and our corresponding final spectrum for the class (P-III silhouette-weighted centroids; s_m=1.81). For easier comparability, all ratios and fractions of signals presented in the following sections are similarly calculated from the corresponding final spectra (P-III).

3.1 Receptor model construction steps

3.1.1 Phase I: identification of anthropogenic aerosol components

In phase I, we performed unconstrained PMF runs using 10 factors for all 8 datasets separately. The resulting 80 factor spectra were subsequently clustered. Maximal data structure (silhouette 0.56) was achieved at mass scaling s_m = 2.12 for 17 clusters (for details on silhouette analysis, see supplementary material, Sect. S.2). The eight clusters with largest population for the phase I solution are shown in Figure 3, and the rest in Section 3.4, where outlier observations are further discussed. Generally, the solutions agreed closely on the largest clusters, lending credibility to the robustness of the approach. The solutions differed mainly regarding outlier classification, which is of secondary importance for our r-CMB model, since outliers are discarded from the model.

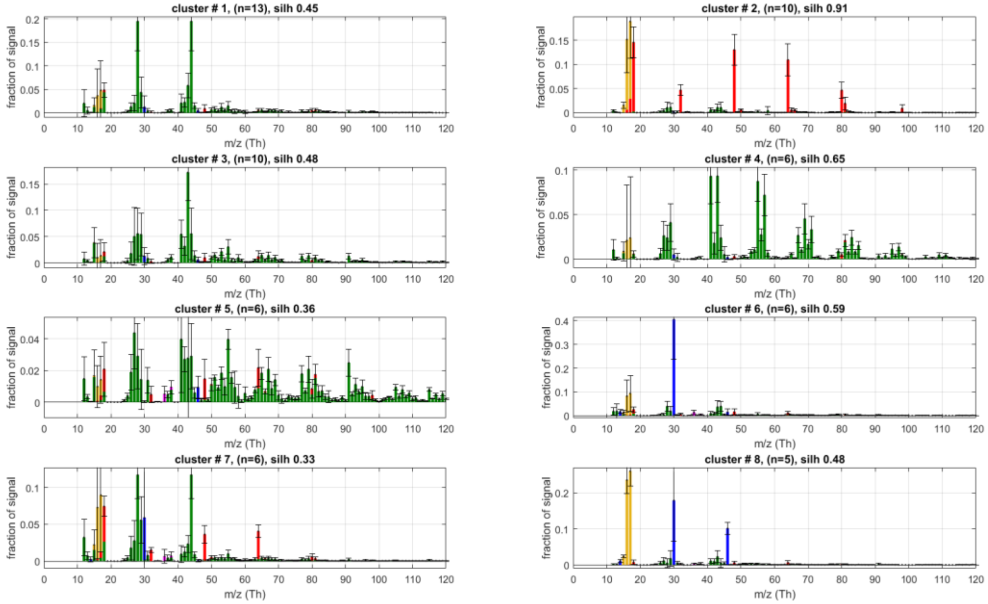


Figure 3. The 8 largest clusters for P-I classification of factorisation results. Cluster centroids (coloured bars; red: SO_4: blue NO_3; orange: NH_4; magenta: chlorides; green: org) and variabilities (error bars) are silhouette-weighted averages and standard deviations for the cluster members. The main anthopogenic aerosol types were identified as Clusters #2 ('Ammonium sulfate', AS), #4 ('Hydrocarbon-like organic aerosol', HOA), #5 ('Biomass burning organic aerosol', BBOA) and #8 ('Ammonium nitrate', AN). Cluster number, silhouette and population (n) are shown in panel titles.

Unsurprisingly, the classification returns two large clusters of organic resembling the ubiquitous low-volatile oxidised organic aerosols (#1; "LV-OOA) and semi-volatile oxidized organic aerosol ("SV-OOA"; e.g. Aiken et al., 2007; Jimenez et al., 2009; Zhang et al., 2011). Comparing to library spectra, the *m/z* 44 Th (CO_2^+) dominated aerosol type (#1) best matches with LV-OOA and OOA-I (Oxidised Organic Aerosol; a historical label corresponding to LV-OOA; (Aiken et al., 2008; Zhang et al., 2011) spectra from Paris (r_s^2 = 0.97; Crippa et al., 2013), Zurich (0.96; Lanz et al., 2007a; Crippa et al., 2013) and Borneo rainforest (0.99; Robinson et al., 2011) as well as the average LV-OOA calculated from 15 Northern Hemisphere datasets (0.94; Ng et al., 2010). Cluster #3 is characterized by high *m/z* 43 Th signal ($C_2H_3O^+$; Aiken et al., 2008), and correlates with SV-OOA and OOA-II (Aiken et al., 2008) spectra from Pasadena (0.74; Hersey et al., 2011), Borneo (0.86; Robinson et al., 2011) and the 15 data set average (0.76; Ng et al., 2010) as well as the laboratory generated SOA spectra generated from typical pine forest emitted volatile organic compounds (e.g. a-pinene, 0.81; a-terpinene, 0.83; terpinolene, 0.84; (Bahreini et al., 2005)). Abiding by the typical naming convention of AMS derived aerosol types, we label these species LV-OOA (cluster #1) and SV-OOA (#3).

The solution also contains a large cluster (#2) with spectra dominated by ammonium and sulphate ion species. This is in agreement with ammonium sulphate being a main component of ambient aerosols. Although it contains also trace amounts of other species, we name the $(NH_4)_2SO_4$ –dominated aerosol class (#2) ammonium sulphate ("AS") for brevity.

The main nitrate-containing spectra are divided into two clusters (#6 and #8). The divisive feature seems to be the ratio of *m/z* 46 to 30 Th signals (i.e. $R_{measured}$ in Equation 7), which is much higher in cluster type #8 (0.44 ± 0.11) versus for #6 (0.08 ± 0.07; P-III ~~values~~; see S.I. Sect S.5 for error estimate). We note once more that these characteristic values for clusters are from the final model (P-III; Figure 4), as outlined before. Based on literature we interpret the split to correspond to the division between nitrogen in form of inorganic (ammonium) nitrate (AN), and organic nitrogen, matching with previous AMS observations (Hao et al., 2014; Farmer et al., 2010; Kiendler-Scharr et al., 2016). The interpretation of cluster #8 as AN is additionally corroborated by its similarity to spectra from pure ammonium nitrate calibration for the instrument, available in supplementary material (Sect S.6). On average, the (Brute-Force Single Particle, BFSP; Drewnick et al., 2015) AN calibrations performed for the instrument yielded an R_{calib} (Equation 7) ratio of 0.49 ± 0.05 (mean ± standard deviation), while an MS mode calibration returned an R_{calib} of 0.42. Similarly to naming of the AS class, we use labels organic nitrogen ("ON"; cluster #6) and ammonium nitrate (AN; cluster #8) for the nitrate-dominated aerosol types. The ON cluster is further discussed in Section 3.3.2. The label ON was chosen to differentiate between the (presumably) organic nitrogen dominated aerosol class (ON), and the part of NO_3 ion species deemed likely to be organonitrate ($orgNO_3$).

A fraction of the organic signal observed at *m/z* 44 Th for inorganic salt classes (AS and AN) may be explained by an CO_2^+ artefact induced by thermal decomposition of inorganic salts (Pieber et al., 2016). For ammonium nitrate, the ratio of organic signal at *m/z* 44 Th to total nitrate signal is 2.9% (P-III). Pieber et al. (2016) estimate a contribution of 3.4 %, suggesting most of the organic signal observed in AN may arise from this artefact. This proposition is further discussed in supplementary information (Sect. S.6)

Two of the clusters (#4 and #5) seem related to anthropogenic (primary) organic aerosol types. Cluster #4 has a similar spectrum as the hydrocarbon-like-organic aerosol ("HOA") spectra from the AMS spectral database (Ulbrich et al., 2009), and closely matches, among others, HOA reported by Zhang et al. (Zhang et al., 2005) for Pittsburgh (r_s^2 = 0.91) and the average of de-convolved 15 HOA spectra reported by Ng et al (2010; r_s^2 = 0.89). The spectra also exhibits high similarity with traffic emission spectra of diesel bus exhaust (0.86), lubricating oil (0.82) and fuel (0.75), reported by Canagaratna et al. (2004).

Cluster (#5) features high signals for ions typical of biomass burning organic aerosol ("BBOA"; e.g. Alfarra et al., 2007) and cooking organic aerosol ("COA"; e.g. Mohr et al., 2012). The spectra features the marker signals of levoglucosan (Cubison et al., 2011; Schneider et al., 2006) at *m/z* 60 ($C_2H_4O_2^+$) and 73 Th ($C_3H_5O_2^+$) along with chloride ions (at *m/z* 35 and 36 Th) and high fraction of signal at *m/z* 55 Th ($C_3H_3O^+$; Mohr et al., 2012), pointing to cooking and/or biomass burning emissions. Highest similarities to library spectra (de-convolved via PMF) are found with COA (Mohr et al., 2012; Barcelona; r_s^2 = 0.70 and Crippa et al., 2013; Paris; r_s^2 = 0.59) and BBOA (e.g. 15 dataset average reported by Ng et al. (2010; r_s^2 = 0.51) and BBOA de-convolved by Crippa et al. (2013; for Paris; r_s^2 = 0.50). Similarity to SV-OOA library samples are also moderately high (e.g. Ng et al., 2010; 15 dataset average; r_s^2 = 0.59).

The differentiation between HOA versus BBOA or COA can often be resolved even from unit resolution spectra, using the *f55* to *f57* ratio (Mohr et al., 2012), and the differences in mass spectral fingerprints higher up on the *m/z* axis (resolvable using mass scaling; Äijälä et al., 2017). However, the distinction between COA and BBOA aerosol types is much more delicate due to very high UMR spectral similarity also for higher *m/z* variables, (e.g. $r_s^2 = 0.79$ for COA and BBOA reported by Mohr et al., (2012). The main difference between the COA and BBOA aerosol types is the absolute level signals from levoglocosan fragments, the quantitative interpretation of which is difficult due to (1) levoglucosan production being determined by combustion temperature (Shafizadeh, 1984), (2) levoglucosan originating both from BBOA and COA (Mohr et al., 2012) and (3) levoglucosan sinks may be considerable in the atmosphere (Hoffmann et al., 2009), which affects especially transported aerosol. Due to the remote location of the measurement site and general prevalence of BBOA over COA in urban aerosol loadings (e.g. Daellenbach et al., 2017) we conclude BBOA is more likely the dominant component for this aerosol type, so we will use the class label "BBOA" for brevity. Due to high spectral similarity, we find it extremely likely any COA contribution would be apportioned to this class, but without the benefit of high mass resolution data, the convolution seems insolvable at this time.

Cluster 7 spectrum offers little in terms of unique spectral features, and it appears as it could be represented as a combination of the more distinct AS (#2), LV-OOA #1) and ON (#6) aerosol types. It is unclear if this class represents an actual aerosol chemical type, or if it is due to incomplete resolving of the aforementioned species in the PMF model. We note the organics part of AS, LV-OOA and ON are all highly oxidised, which may imply similar level of aging, thus similar origins for these species. Organic spectral component are further analysed and discussed in Sect. 3.2.2.

Based on this interpretation and evaluation of criteria outlined in Sect. 2.5, we decided to select as the main representative anthropogenic aerosol types the following: ammonium sulfate (AS, cluster #2, $n = 10$, silhouette = 0.91) ammonium nitrate (AN, #8, $n = 5$, silh = 0.48), hydrocarbon-like organic aerosol, (HOA, #2, n = 6, silh = 0.65) and biomass burning organic aerosol (BBOA, #5, $n = 6$, silh = 0.36). The silhouette values can be taken to represent separation distance from neighbouring aerosol types. For comparison, silhouette values for some of the anthropogenic organic aerosol types are available in Äijälä et al. (2017), but to our knowledge no precedent exists for mixed or inorganic aerosols. Generally, the more 'unique' the spectra of a group and the higher within-cluster-cohesion, the higher the silhouette.

3.1.2 Phase II: classification of biogenic, secondary organic aerosols

In the second phase of our analysis, ME-2 factorisations were run for ten factors for all the data sets. We constrained 4 out of the 10 factors with the anchors and variabilities for anthropogenic aerosol types, derived from previous phase (AS, AN, HOA, BBOA). The resulting 80 factor profiles were again extracted and classified. The classification solutions featured generally higher silhouette value than in the first phase, which is at least partly explained by constrained spectra being forced to conform to their set limits. The highest total silhouette (0.66) was obtained for 15 clusters (at $s_m = 2.41$). Again, the inter-solution variability for the solutions inspected was low for the main classes. The phase II solution is available in S.I (Figure S.4). Overall, the solution very closely resembles the result from phase I (Figure 3).

The expected LV-OOA (#1; *n* = 14; silh 0.64) and SV-OOA (#3; *n* = 9; silh 0.44) aerosol types again rank among the most typical classifications. Their moderate silhouettes reflect higher variability within these classes, corresponding to results from earlier studies (e.g. Canonaco et al., 2015), and/or closer proximity to neighbouring aerosol types, than for the AN, AS or HOA types. The result may suggest seasonal or other dataset-specific variability for SOA, which supports partitioning the data on a per-campaign basis. In accordance with typical AMS organic aerosol classification conventions laid out by e.g. Aiken et al. (2008), we opt for two classes of oxidised aerosols. We thus select clusters #1 and #3 (P-II) to represent LV-OOA and SV-OOA (Aiken et al., 2008; Jimenez et al., 2009) respectively.

For P-III of our analysis, we additionally fix the organic nitrogen class, (ON, P-II cluster #8). Irrespective of the exact chemical composition and label of this aerosol component, we assess there is enough literature support (among others Kiendler-Scharr et al., 2016; Farmer et al., 2010; Drewnick et al., 2015; Murphy et al., 2007; Hao et al., 2014) for inclusion of nitrogen containing aerosol types other than AN, to warrant the inclusion of this class. In any case, the classification of nitrate signal at *m/z* 30 Th to a distinct class seems statistically robust, as exhibited by its emergence as a free factor in both P-I and P-II solutions. Due to the importance of nitrogen containing species in SOA composition and formation (e.g. (Kiendler-Scharr et al., 2016; Berkemeier et al., 2016) we find it an important aerosol class to include, examine and further interpret. The mixed cluster #7 also emerges for 4 data sets, but with notably low silhouette (0.18), suggestive of low within-cluster cohesion. As we still lack a distinct chemical interpretation for this class, beyond the hypothesis of incomplete resolution of aged aerosol species in factorisation, we will not include the mixed class (#7) in our final receptor model.

3.1.3 Phase III: Final r-CMB receptor model

In the final phase (P-III) of constructing our r-CMB receptor model, we used 7 factors which were all constrained with the profiles and allowed variabilities from the previous phase (P-II, AS, LV-OOA, SV-OOA, BBOA, ON, HOA, AN). The ME-2 algorithm was tasked to resolve the factors' temporal behaviour.

To derive final characteristic spectra for the model components, as well as to study the variability of spectra in the solutions, we once more applied the same clustering procedure and silhouette analysis as for previous phases. The maximal structure (silh 0.85) was achieved for the seven cluster solution (s_m = 1.81), which was to be expected considering ME-2 was run with 7 rather strictly constrained factors in this phase. With silhouette weighting applied, we obtain the final spectra and variabilities (Figure 4). We note this final clustering and weighting step mainly serves to provide an estimate of variability within each aerosol type, but also yields final spectra to be used as library references for the outcome of this work. Details of the solution of the r-CMB model are discussed in following sections, from the perspective of mass attribution (Section 3.2.1) and spectral characteristics (Section 3.2.2). Diurnal cycles of the components for the entirety of data are available in S.I (Figure S.12). Due to the rural setting of the site and the generally long transport times of aerosol before reaching the site, diurnal cycles for the various aerosol types are not as characteristic as they would be for urban measurements (for e.g. temporal trends of HOA and BBOA). Also due to seasonal differences, the variability between data sets is considerable, resulting in high uncertainty in interpretation. The daily cycles are likely a mixed product of source emissions, boundary layer dynamics and aerosol temperature response. While of interest, disentangling these processes is beyond the topic of this study.

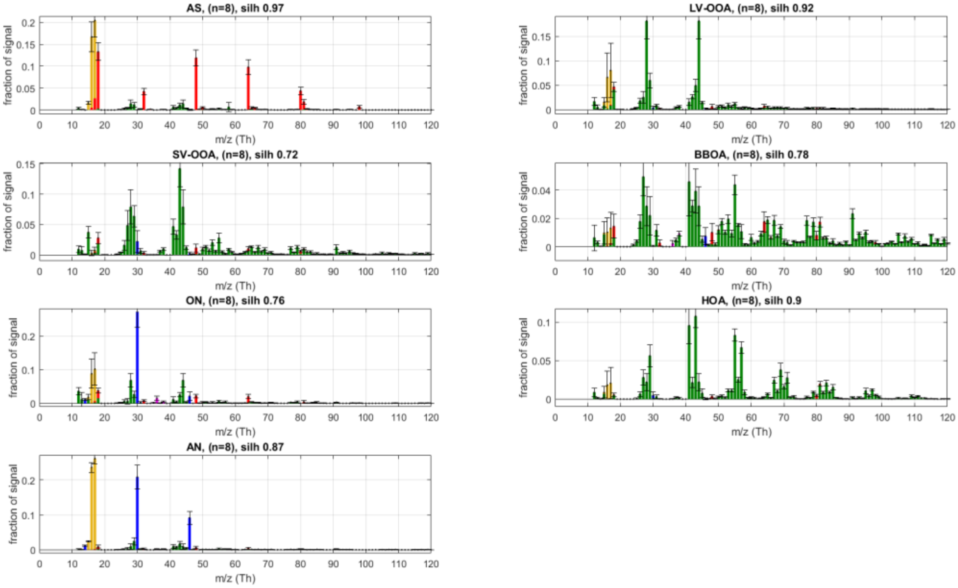


Figure 4. Final silhouette-weighted reference spectra and variabilities for the r-CMB model components. (red: SO_4: blue NO_3; orange: NH_4; magenta: chlorides; green: org)

3.2 Overview of r-CMB model results

3.2.1 Mass attribution and "default" AMS chemical speciation for r-CMB components

Tabulation of final explained variations (EV; Paatero, 2000; Canonaco et al., 2013) for the r-CMB model are shown in Table 3~~2~~. The seven-component r-CMB model, explains 83 ± 8 % of the variation in loadings, when variation from low SNR variables is included, and 97 ± 3 % when only residuals of variables with SNR > 2 are considered. The components with lowest loadings (ON, HOA, AN) explain around 4 to 5 % of variation, which seems to roughly match the general rule of thumb, of PMF / ME2 being able to extract components of around 5 % of contribution (Ulbrich et al., 2009).

Model results for campaign VIII, especially regarding BBOA, are very different from other data sets, including the other cold season results available in e.g. data set III (Figure S.5.). Upon closer examination, we attribute the VIII anomaly at least partly to pronounced surface ionisation effects, discussed more in Section 3.4. While we consider the r-CMB results for campaign VIII too unreliable for use in models or further studies, we decided not to omit data set VIII, since other AMS data is likely also affected by the same processes, albeit to a lesser degree. The attribution of anomalies to exact processes is very difficult, and surface ionisation effects remain hard to quantify. We hope that reporting our results in full also furthers the discussion on surface ionisation in the AMS, and potentially helps other AMS users observing similar observations.

The composition of our r-CMB components is shown in Figure 5, panel (b), and the same in absolute in mass units in panel (a). The opposite visualisation, i.e. attribution of default species into r-CMB components, is similarly given for absolute mass concentration and relative units in Figure 5, panels (c) and (d). Unlike mass spectral variables and estimated EV, where signals at m/z are in units "nitrate equivalent mass" (RIE not applied), all mass concentrations reported are corrected for relative ionisation efficiency (see S.I., Sect S.4).

Table ~~2~~3. Explained variations (EV, in percent) for the r-CMB-like model.

Data set		**r-CMB component**							**Residual**		**Rate of explanation**	
nb	**name**	**AS**	**LV**	**SV**	**BB**	**ON**	**HOA**	**AN**	**low SNR**	**high SNR**	**all**	**high SNR**
I	**May 2008**	29 %	15 %	15 %	3 %	5 %	5 %	3 %	21 %	4 %	75 %	94 %
II	**Sep 2008**	34 %	12 %	14 %	11 %	5 %	5 %	3 %	14 %	3 %	84 %	97 %
III	**Mar 2009**	40 %	19 %	8 %	8 %	5 %	2 %	8 %	9 %	1 %	90 %	99 %
IV	**May 2009**	25 %	16 %	19 %	14 %	6 %	5 %	3 %	10 %	2 %	88 %	98 %
V	**Jun 2009**	18 %	28 %	23 %	11 %	4 %	6 %	2 %	7 %	1 %	92 %	99 %
VI	**Aug 209**	22 %	26 %	19 %	9 %	4 %	5 %	4 %	10 %	1 %	89 %	99 %
VII	**Summer 2010**	21 %	25 %	10 %	7 %	4 %	4 %	2 %	21 %	6 %	73 %	93 %
VIII	**Winter 2010**	25 %	5 %	1 %	29 %	7 %	2 %	4 %	17 %	9 %	74 %	89 %
mean		27 %	18 %	14 %	12 %	5 %	4 %	4 %	14 %	3 %	83 %	96 %
st.dev.		7 %	8 %	7 %	8 %	1 %	1 %	2 %	5 %	3 %	8 %	3 %

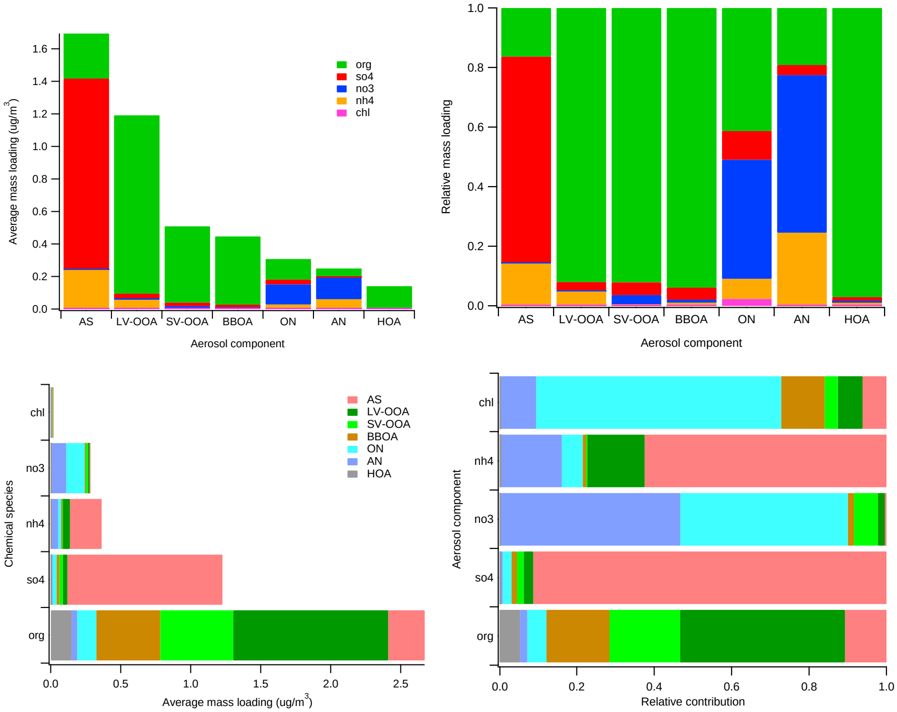


Figure 5. "Default" chemical speciation for r-CMB components; mass loadings (upper left) and relative contributions (upper right) of default species in components. Apportionment of default species to r-CMB components by mass (lower left) and relative contribution (lower right).

Generally, the separation between the inorganics r-CMB components (AS, AN) and organics (LV-OOA, SV-OOA, BBOA, HOA) seems clear (Figure 5). Ammonium nitrate and sulphate components consist primarily of inorganic ion species (81 to 84 %), while for organic components the inorganic ion species contribution is small (LV-OOA: 8%, SV-OOA: 8%, BBOA: 6%, HOA: 3%). However, extensive oxidation of organics in aerosol typically results in formation of organic acids (Yatavelli et al., 2015; Vogel et al., 2013; Duplissy et al., 2011) and we hypothesise organic salt formation with $[\mathrm{NH4}^+]$ could account for the notable 5 % mass contribution of ammonium to this aerosol type.

Explanations for the observed mixing of ion species can include (1) mixed emission profiles at sources, variabilities within a source type, as well as collocation of sources. (2) Atmospheric processes, such as mass transfer between the species by evaporation, condensation (e.g. Ye et al., 2016) as well as coagulation. (3) PMF / r-CMB modelling uncertainties. We will discuss the relative ratios and neutralisation balances of inorganic ion species in Section 3.3.2, in relation to inorganic salt formation scheme. The interesting exception to the rather clear-cut ion species separation is the ON component, which contains 40 % of NO_3 species ions, and 41 % ions defined as organic. The possible interpretations for this distribution are further discussed in Section 3.3.2

As for the organics-inorganics division, the two speciations (default vs r-CMB) give similar results (Figure 6). For all the data sets combined, the default organic ion species ("org") explains an average 57 % of total aerosol mass at the site. Similarly, combining the mass of all organic-dominated components (LV-OOA, SV-OOA, BBOA, HOA and ON) results in 60 % mass fraction versus 40 % explained by ammonium nitrate (5 %) and ammonium sulphate (35 %) salts. The per-data set mass apportionment is presented in supplementary information (Figure S.9).

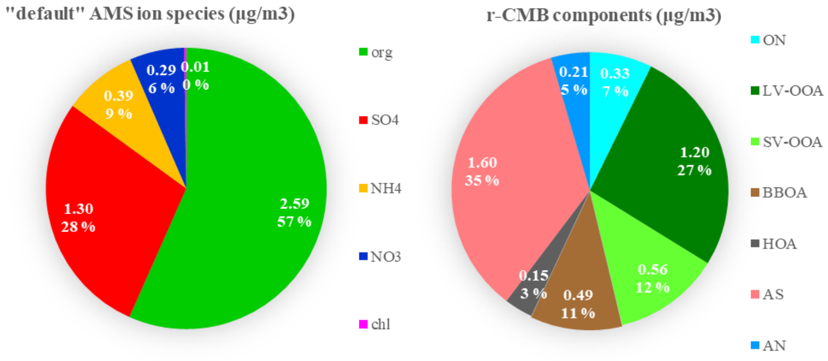


Figure 6. Mass attribution in the default AMS speciation scheme (left) and by r-CMB components (right) for all the 8 data sets combined. Values are (data set length-weighted) averages for all data combined. Absolute mass concentrations are in units (µg/m^3).

3.2.2 Spectral characteristics of organic components

As discussed above, despite the mixing observed, the inorganic aerosol classes generally seem separate from organic aerosols. The scaled correlation values between inorganic and organic spectra are extremely low (S.I Sect S.8, Tables S.1 and S.2), indicating near-zero similarity and clear-cut separation between the inorganic and organic aerosol types by the clustering algorithm. For inter-correlations between the organics-dominated aerosol classes, the picture is somewhat more complex.

To understand the drivers for the separation of the organic aerosol types, we visualised the phase I (unconstrained PMF) and phase III (r-CMB) classification results with a projection of the clustering solutions onto a plane defined by an axis corresponding to estimated oxidation level and another connected to source type (P-III in Figue 7; P-I available in S.I, Fig. S.6.). Similar to Äijälä et al. (2017), we describe the oxidation level of the organic fraction of each component using the oxygen-to-carbon ratio (O:C) parametrisation of Aiken et al. (2008), and use the ratio of~~between~~ *f57:f57* ~~fractions~~ to imply source type. The O:C generally separates LV-OOA and SV-OOA species from each other and from the fresher aerosol classes.

The *f55:f57* ratio is typically used for differentiation between HOA and COA/BBOA (Mohr et al., 2012), but equally seems to set apart the biogenic SOA types from the anthropogenic aerosols (Äijälä et al., 2017). This is due to the low signal of *m/z* 57 Th, a typical anthropogenic spectral marker, originating from $C_4H_9^+$ and $C_3H_5O^+$ compounds (Mohr et al., 2012; Zhang et al., 2005).

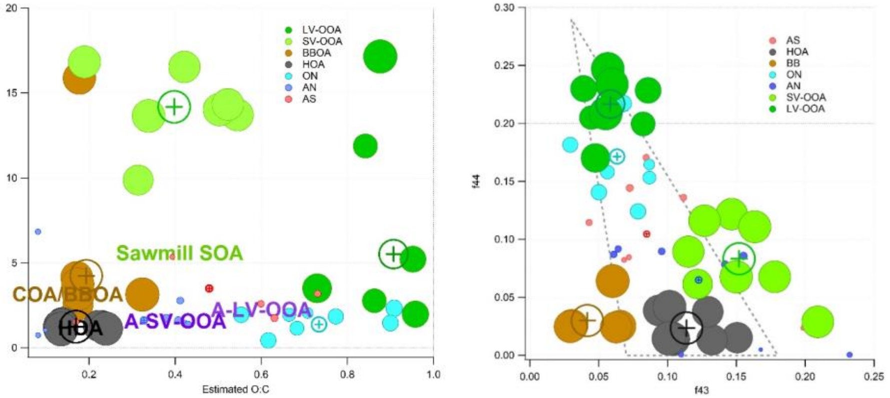


Figure 7. (Left panel:) P-III (r-CMB) solution - cluster projections onto a *f55/f57* (Mohr et al., 2012*), O:C* (estimated, Aiken et al., 2008) plane. Circles correspond to the members of the cluster and the cross markers to cluster centroids. The text markers indicate respective positions of anthropogenic organic aerosol types from Äijälä et al. (2017). Marker size indicates organic mass fraction in spectra. Axes are truncated. (Right panel:) P-III solution, projected onto *f44*, *f43* plane (i.e. the 'Sally's triangle' plot; Ng et al., 2011). Circles correspond to objects in cluster and the cross markers to cluster centroids. Marker size indicates organic mass fraction in spectra. Dotted line marks the area where most laboratory data for organic aerosol falls (Ng et al., 2010).

The LV-OOA aerosol type, characterised by the dominant *m/z* 44 and 28 Th signals is usually considered a highly oxidised aerosol type that results from oxidation of SV-OOA and various fresh emissions (among others Canonaco et al., 2015)). The *f55:f57* ratio of LV-OOA is considerably lower than for SV-OOA in both solutions, indicating inclusion of other sources beyond the *f57*-poor biogenic SOA contribution. SV-OOA, on the other hand, has the highest *f55:f57* ratio of the classes, hinting to the predominantly biogenic origin of the SV-OOA at the site. The difference is further amplified for phase II and III solutions compared to the unconstrained PMF. We hypothesise this change can result from improved differentiation between SV-OOA and the BBOA species (in P-II), as these aerosol types may be difficult to separate initially due to similar oxidation level and features of the spectra (r_s^2 = 0.34; Table S.3). The SV-OOA is characterised by the non-oxygen-containing ions at *m/z* 29, 43 and 55 Th (Mohr et al., 2009), as well as mass-to-charge *m/z* 53 Th signal ($C_4H_5^+$) typical of boreal forest biogenic background (e.g. Corrigan et al., 2013). The NO_2^+/NO^+ ratio of 0.10 for nitrate-containing SV-OOA reported by Hao et al (2014) matches our observations for the nitrates in SV-OOA (NO_2^+/NO^+ of 0.11 ± 0.15; Equations 7 and S.5). This may indicate presence of organonitrate species in the SV-OOA factor.

We also projected the P-I and P-III solutions to the (*f44*, *f43*) plane (P-III in Figure 7; P-I in S.I, Fig S.6), to produce a result comparable to the triangle plot by Ng et al. (2010). The result indicates a clear separation between the low and semi volatile aerosol types, as well as the primary combustion aerosols (HOA, BBOA), and the spectral shifts from phase 1 "bulk PMF" results to the final r-CMB model.

As stated in Section 3.1., the spectra of BBOA and HOA aerosol types match the previously published observations. The HOA spectrum is characterised by the ion series C_nH_{2n+1} (*m/z* 29, 43, 57, 71, 85, 99 Th etc.) and C_nH_{2n-1} (*m/z* 41, 55, 69, 83, 97 Th etc.) resulting from alkanes and aromatics from traffic emissions (diesel exhaust, lubricating oil; Chirico et al., 2010; Mohr et al., 2009; Canagaratna et al., 2004). The biomass burning organic aerosol levoglucosan marker signals at *m/z* 60 ($C_2H_4O_2^+$) and 73 Th ($C_3H_5O_2^+$) (Cubison et al., 2011; Schneider et al., 2006; Elsasser et al., 2012) are clearly identifiable in the BBOA spectra (Figure 3; Figure 4), and set this class apart from HOA and SV-OOA with some similar features. The contribution of often biogenic signal at *m/z* 53 Th is also lower for BBOA than for the biogenic, semi-volatile SOA. The pronounced signal from aromatic ring (tropyllium cation $C_7H_7^+$) at *m/z* 91 Th is a typical result of fragmentation of aromatic hydrocarbon compounds (Lindon et al., 2016). As stated previously, we presume the BBOA class also encompasses any COA contributions, which are likely unresolvable as a separate class due to high spectral similarity (0.79; Sect 3.1.1).

In terms of spectral characteristics, the organic contributions of AS and AN classes fall somewhere between the distinct organic classes and offer little in terms of significant organic markers. Notably, the organics in the ON class exhibit some of the characteristics of LV-OOA and feature generally high *f44*. This may indicate high degree of oxidation of the organics for this aerosol type (Aiken et al., 2008). However, alternative plausible interpretations exist: AMS response from oxidation products of amine compounds and amine-nitrate salts feature similarly high f*44* (Murphy et al., 2007) as does a typical amine fragment ion $C_2H_6N^+$ (McLafferty and Turecek, 1993). Furthermore, as discussed in Section 3.3.2, an equally plausible explanation would be inorganic nitrate salts such as KNO_3 (from e.g. biomass burning; Li et al., 2003) contributing to this class in form of the Pieber et al (2016) thermal decomposition artefact. The contribution of *m/z* 55 and 57 Th signals to the ON species are both low and the ratio 1.37 of *f55:f57* is much lower than for the biogenic aerosol species. Without more detailed analysis, and due to the uncertainties surrounding the origins of this aerosol type (Section 3.3.2), it is difficult to say with any certainty if this is due to anthropogenic nature of this aerosol, or e.g. due to fragmentation pattern of characteristic organic compounds in this aerosol type.

3.3 Comparisons with reference methods

3.3.1 Comparison with "traditional" ME-2 analysis for aerosol organic component

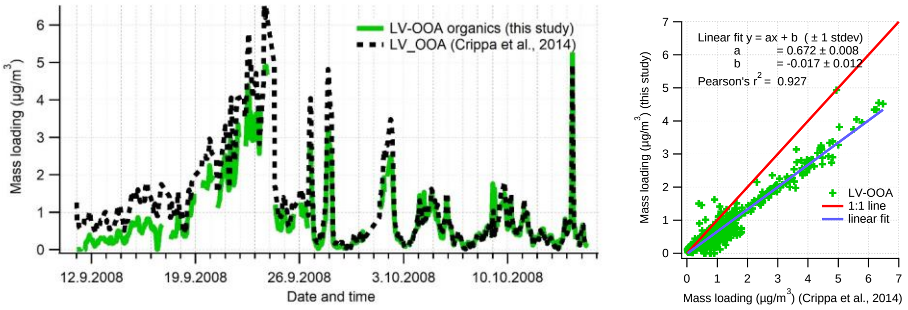
In order to evaluate the performance of the source apportionment approach presented in this study for organic aerosol, we compare our results to results only relying on the organic mass spectral fingerprints. Specifically, two data sets covered in this study (data sets II and III; -Table 1~~Error! Reference source not found.~~) were also included in the Crippa et al. (2014) analysis, which allows us to compared factorisation results directly. We chose to compare the Crippa et al. results to ours from data set

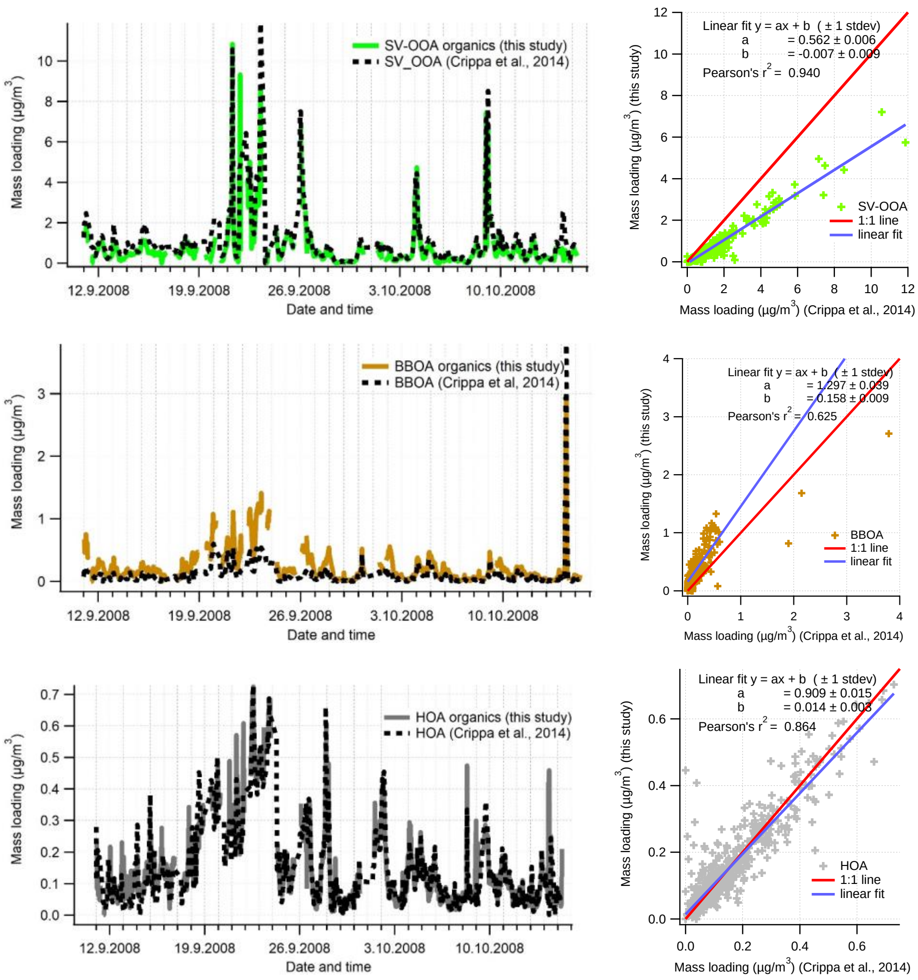
Formatted: Font: Not Bold, English (United Kingdom)

II. We note that while there are minor differences in the pre-processing and corrections for data covered in Crippa et al (2014), the factorisation input is very similar in both cases. The ME2 model used by Crippa and co-workers included only the organic spectra and apportioned its mass to four factors: LV-OOA, SV-OOA, BBOA and HOA. The latter two components were constrained using a HOA profile from an urban aerosol study in Paris (Crippa et al., 2013) and an average BBOA of those extracted for Mexico City, Mexico, and Houston, U.S (Ng et al., 2011). The allowed variability around these anchors for all variables (*m/z*), were 5 % (HOA) and 30 % (BBOA).

We compared the solutions for Crippa et al. factorisation to our r-CMB model solution data set II, both for loadings (Figure 8) and profiles (Figure 9). Generally the solutions correlated highly – the loadings (F) and profiles (G) for LV-OOA (F: r^2=0.92; G: r_s^2= 0.96) and SV-OOA (*F*: 0.94; *G*: 0.99) agreed the closest, whilst the HOA also had high similarities (*F*: 0.85; G: 88). The BBOA factor / component correlated markedly less (*F*: 0.63; *G*: 0.42), which we hypothesise to be due to differences in the anchors used, COA likely attributed to this class, high spectral similarity between SV-OOA and BBOA and the generally low loadings of BBOA observed at SMEAR II.

The discrepancy in distribution of absolute mass for the LV-OOA and SV-OOA components, indicated by the sub-unity slope, suggests the r-CMB model attributes a part of the organic mass from the SOA factors into BBOA, AS, AN and ON components, while HOA is represented rather identically in both models. A difference in mass distribution between the results is to be expected, considering the r-CMB model allows for organics in 7 components, while the model of Crippa et al. model only comprises 4 components. Generally, we take the similar results of the methods, as shown by the high correlation values, to indicate that inclusion of inorganics in the model does not significantly perturb modelling of the organics. We also note the r-CMB components included (HOA BBOA, LV-OOA, SV-OOA) are predominantly composed of organics (92 to 97 %; Figure 5), and the 4 components presented comprise 82 % of total organics.




SV-OOA organics (this study)
SV_OOA (Crippa et al., 2014)
Mass loading (µg/m^3)
12.9.2008
19.9.2008
26.9.2008
3.10.2008
10.10.2008
Date and time
Linear fit y = ax + b (± 1 stdev)
a = 0.562 ± 0.006
b = -0.007 ± 0.009
Pearson's r^2 = 0.940
SV-OOA
1:1 line
linear fit
Mass loading (µg/m^3) (this study)
Mass loading (µg/m^3) (Crippa et al., 2014)
BBOA organics (this study)
BBOA (Crippa et al, 2014)
Linear fit y = ax + b (± 1 stdev)
a = 1.297 ± 0.039
b = 0.158 ± 0.009
Pearson's r^2 = 0.625
BBOA
HOA organics (this study)
HOA (Crippa et al., 2014)
Linear fit y = ax + b (± 1 stdev)
a = 0.909 ± 0.015
b = 0.014 ± 0.003
Pearson's r^2 = 0.864
HOA

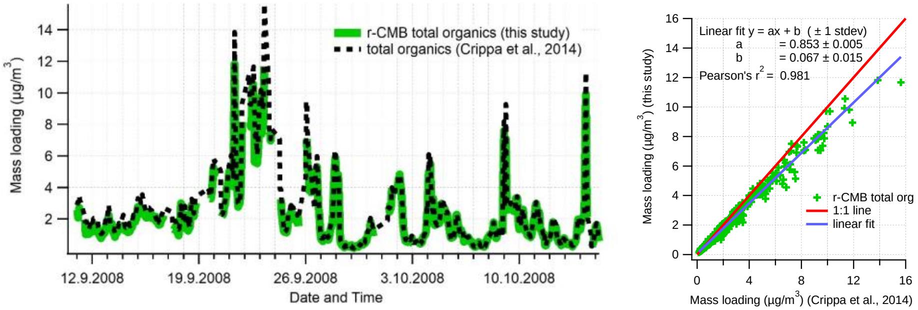
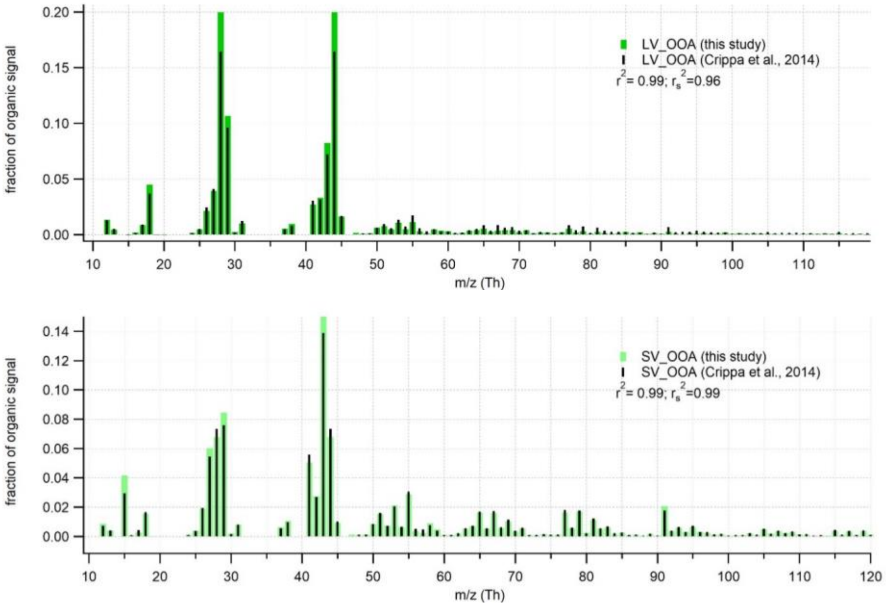


Figure 8. Time series comparison of aerosol organic component with Crippa et al. (2014) for the September 2008 campaign (data set II). For comparability, only the organic part of r-CMB model components are considered. Data from this work has been averaged to 1 hr resolution. Organics in other r-CMB components (AS, AN, ON) are taken into account for the total amount but not shown separately. Discrepancy in total organics loading is due to differences in pre-processing values (e.g. ionisation efficiency, collection efficiency)



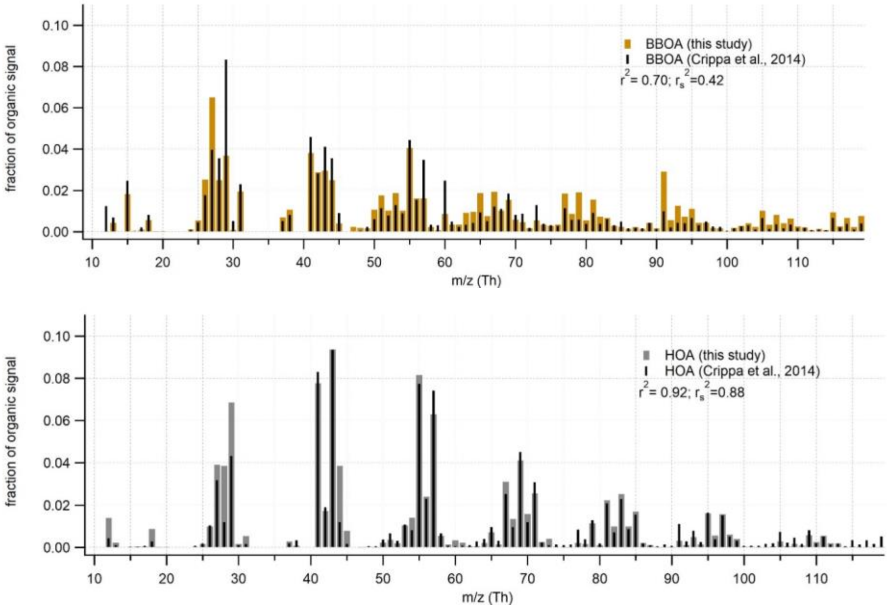


Figure 9. Comparison of organic part of spectra with Crippa et al. (2014) for data set II. The r-CMB model results from this study are shown in colour, and the Crippa et al. spectra in black. For comparability, the Crippa et al. spectra were corrected for a difference in fragmentation tables used (included *m/z* 28 Th, updated to modern calculation of *m/z* 16, 17 and 18 Th organic signals) and total signal subsequently re-normalised to unity. Spectra similarity is evaluated using Pearson's squared correlation coefficients; unscaled (r^2) and with mass scaling (r_s^2).

3.3.2 Comparison of inorganic salt and organic nitrogen results with reference methods

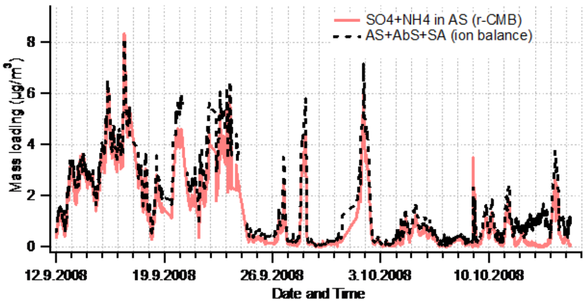
To evaluate the inorganic mass apportionment result, we compared the loadings from the r-CMB solution against the result from the inorganics apportionment scheme (Section 2.4.1). The comparison, again performed for data set II, is presented in Figure 10. We additionally compared the r-CMB ON component loadings with orgNO$_3$ mass estimate from the Kiendler-Scharr parameterisation (Equation 7; Section 2.4.2).

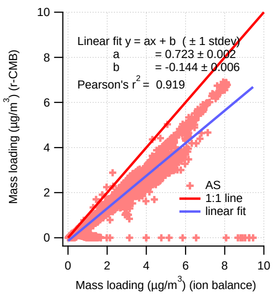
The loadings for the (r-CMB) AS component compare well with the combined NH_4HSO_4 + $(NH_4)_2SO_4$ + H_2SO_4 loading, indicating ammonium(bi)sulphate is described similarly by both models (r^2=0.92). We assume the r-CMB AS component to comprise of both NH_4HSO_4 and $(NH_4)_2SO_4$, which would very likely be classified together in PMF / clustering due to their high spectral similarity. For ammonium nitrate the correlation between loadings is very low (r^2 =0.16). Looking at the time series, the reason seems to be that the speciation scheme–based model often predicts total absence of AN, due to high amount of sulphate in aerosol. While the r-CMB model also generally estimates loadings to be low, they are clearly non-zero in the r-

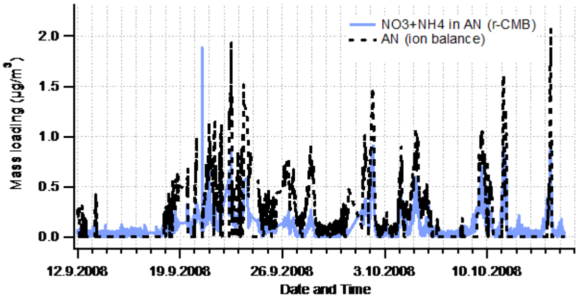
CMB model. We take the result to reflect the assumption of complete and instantaneous, internal and external mixing of aerosol in the speciation scheme (Section 2.4.1).

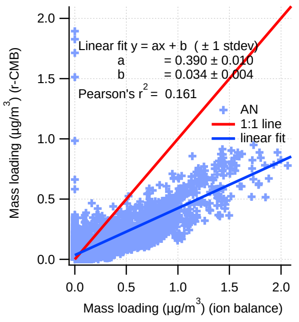
The loading prediction for organic nitrogen by the speciation scheme model is similarly event-driven and the model results do not correlate. This is caused by the nitrate assignment to organonitrate class when not explained by NH_4NO_3. Same can be said for the excess NH_4 class, which corresponds to the NH_4 species in the other, mostly organic r-CMB factors, principally the LV-OOA; the ion balance scheme predicts zero concentration for many of the data points, an estimate not matching with the r-CMB-based result.

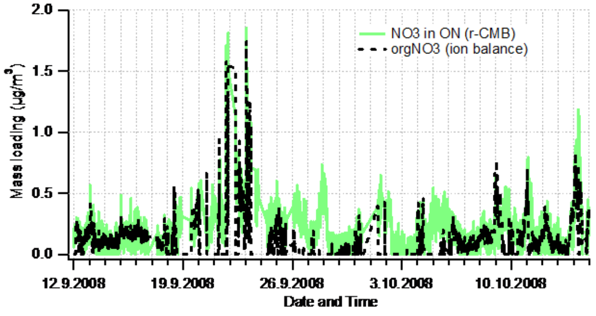
On these differences between the models, we note that the ion balance-based apportionment scheme is sensitive towards small changes in NH_4 concentrations, especially for data with generally low NH_4 concentrations, such as ours. A simple sensitivity estimate, available in supporting material (Sect. S.4) was performed for data set III. The result indicates that a 33 % change in RIE_{NH4} changes the component mass concentrations on average 5% for AS; 56 % for AN, 66 % for orgNO3 and 164 % for excess_NH4 components. On the other hand, the r-CMB model is rather insensitive to error in RIE estimate, since (1) the spectra in factorisation and clustering have the variables' signals in "NO_3 equivalent mass concentration" units, which is not (yet) corrected for RIE of different species, (2) mass scaling causes low mass signals such as NH_4 fragments (*m/z* 15 to 17 Th) to weight less (relative to higher *m/z* variables) for determining the solution, and (3) NH_4 seem not to be an unique marker of any of the classes. We therefore suggest a factorisation-based model such as the r-CMB model presented here is much more robust for resolving speciation of inorganic aerosol components. The sensitivity test (S.I, Sect S.4) also indicates that the temporal differences between the ion balance scheme and r-CMB are not explained by a difference in RIE_{NH4}. Thus, the reasons for the discrepancies are more likely related to the unrealistic assumptions of the inorganics apportionment model.

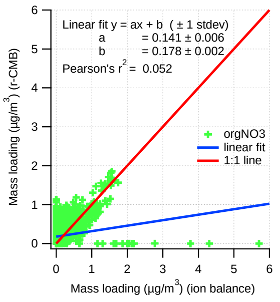

SO4+NH4 in AS (r-CMB)
AS+AbS+SA (ion balance)
Mass loading (µg/m^3)
0
2
4
6
8
12.9.2008
19.9.2008
26.9.2008
3.10.2008
10.10.2008
Date and Time


Linear fit y = ax + b (± 1 stdev)
a = 0.723 ± 0.002
b = -0.144 ± 0.006
Pearson's r^2 = 0.919
AS
1:1 line
linear fit
Mass loading (µg/m^3) (r-CMB)
Mass loading (µg/m^3) (ion balance)
0
2
4
6
8
10


NO3+NH4 in AN (r-CMB)
AN (ion balance)
Mass loading (µg/m^3)
0.0
0.5
1.0
1.5
2.0
12.9.2008
19.9.2008
26.9.2008
3.10.2008
10.10.2008
Date and Time


Linear fit y = ax + b (± 1 stdev)
a = 0.390 ± 0.010
b = 0.034 ± 0.004
Pearson's r^2 = 0.161
AN
1:1 line
linear fit
Mass loading (µg/m^3) (r-CMB)
Mass loading (µg/m^3) (ion balance)
0.0
0.5
1.0
1.5
2.0


NO3 in ON (r-CMB)
orgNO3 (ion balance)
Mass loading (µg/m^3)
0.0
0.5
1.0
1.5
2.0
12.9.2008
19.9.2008
26.9.2008
3.10.2008
10.10.2008
Date and Time


Linear fit y = ax + b (± 1 stdev)
a = 0.141 ± 0.006
b = 0.178 ± 0.002
Pearson's r^2 = 0.052
orgNO3
linear fit
1:1 line
Mass loading (µg/m^3) (r-CMB)
Mass loading (µg/m^3) (ion balance)
0
1
2
3
4
5
6

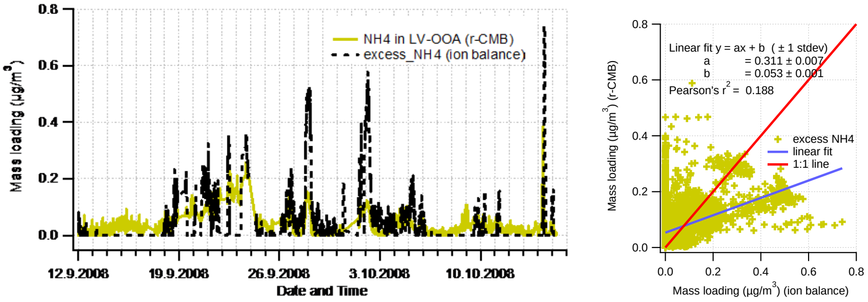


Figure 10. Comparison of Inorganics apportionment methods (r-CMB and ion balance scheme. The estimates from the ion balance scheme (Section 2.4.1) are shown in black, and the r-CMB model results in colour. The linear fits (right panels) represent the data poorly due to high amount of zero-value points and outliers.

In addition to deriving organic nitrogen mass from the ion balance scheme, we compared the r-CMB-derived ON loading with the Kiendler-Scharr method for estimating the orgNO3 mass loading (Equation 6). The comparison, shown in Figure 11, indicates that the two methods produce a very similar result for organic nitrogen mass ($r^2 = 0.94$). The discrepancy in absolute mass is likely explained by the difference in the ratio values (R) used for Equation 6 parameterisation, and those featured in the r-CMB AN and ON components (R_{AN} = 0.~~56~~44 ± 0.11; R_{ON} = 0.08 ± 0.07; P-III, Equation S.5~~5~~). The similarity to Kiendler-Scharr parameterisation result does seem to support the interpretation of nitrogen component in ON as organonitrate (orgNO3). Some similarities in temporal behaviour between the ON component and (non-quantita~~t~~tive) K^+ ions were observed, potentially suggesting thermal ionisation of Potassium salts (e.g. KNO^3) might contribute an unknown fraction to ON (S.I, Sect S.11). Also, 63% of chloride ions species associate with the ON component. The reason is unclear, and although chloride signals were very low in general, we cannot rule out that some of the ON component could still be explained by other chemical compositions than organonitrate.

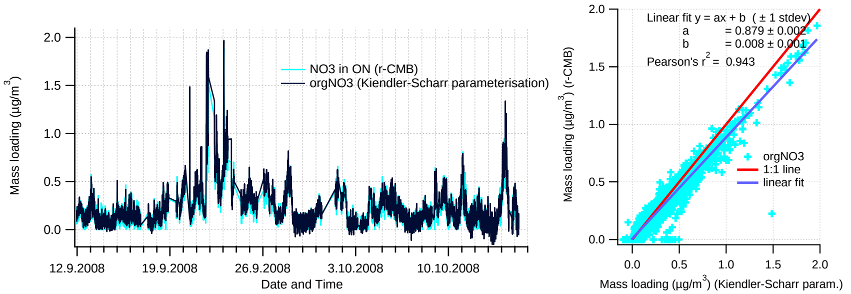


Figure 11. Comparison of Kiendler-Scharr parametrisation (Kiendler-Scharr et al., 2016; black line; moving median filter for 11 points window applied; R_{calib} = 0.42, R_{orgNO3} = 0.1) for organonitrate with NO_3 ion species in ON factor from our r-CMB model (in colour).

The NO_3 : org ratio of our ON factor is close to unity (Figure 5), while e.g. Farmer et al. (2010) report a nitrogen-to-carbon ratio of 0.04, and oxygen-to-carbon of 0.25 for AMS spectra of organonitrate standards. However, several factors are likely to affect the NO_3 : org ratio observable in atmospheric ON factorisations. Firstly, two different pathways for organonitrates exist: i) the primarily daytime reactions of organic peroxy radicals with NO (Orlando and Tyndall, 2012), and ii) the NO_3-radical initiated oxidation of unsaturated compounds during night-time (Peräkylä et al., 2014). While the nitrate functionality in all these reactions are identical, the organic part can be vastly different, as peroxy radicals are formed in almost all atmospheric oxidation reactions, irrespective of oxidant (e.g. OH or ozone) or VOC (biogenic or anthropogenic). Therefore, it is not to be expected that a specific organic spectrum should be linked to the organic nitrate functionality. Secondly, as described by e.g. Lee et al. (2016), the particle phase lifetime of organonitrates is of the order of hours with respect to hydrolysis. This reaction will convert the nitrate functionality to nitric acid, while the organic part remains intact, except for the conversion of the -ONO_2 group to -OH. This conversion will only have a small impact on the volatility of the organic molecule (e.g. Kroll and Seinfeld, 2008), while the nitric acid may well evaporate in the fairly low-ammonia boreal forest environment. Taken together, the diverse formation pathways as well as the atmospheric processing are likely to cause ON spectra retrieved from ambient air factorisations to look different from e.g. freshly formed organic aerosol from organonitrate standards, such as those used by Farmer et al. (2010). We therefore avoid putting too much emphasis on the organic parts observed in our ON factor.

3.4 Outlier observations

During the course of our analysis we encountered some anomalous observations likely stemming from surface ionisation effects, *i.e.* molecules being thermally ionised at the heater surface rather than at the ionisation region by electron impact. A thorough review and discussion on AMS related surface ionisation effects was recently published by Drewnick and colleagues (2015). Drewnick et al. emphasise that the division between refractory and non-refractory aerosol is not binary, and there exist a number of semi-refractory compounds, that the AMS can measure, albeit non-quantitatively.

Our observations on extracted “outlier” PMF factors from the different phases of analysis match well with the finding and calculations of Drewnick et al (2015), as well as other similar AMS observations published. In Figure 12, we present the outlier clusters from phase I classification solution that were excluded from further analysis due to low number of occurrences or/and questionable interpretability. The emergence of most of these spectra are likely attributable to over-resolution or questionable separation of main PMF factors, due to setting the number of PMF factors to 10. Despite their questionable value for the main analysis, we find they contain many potentially interesting mass spectral features, and seem not to emerge by chance. Below we will present some hypotheses on their possible interpretation.

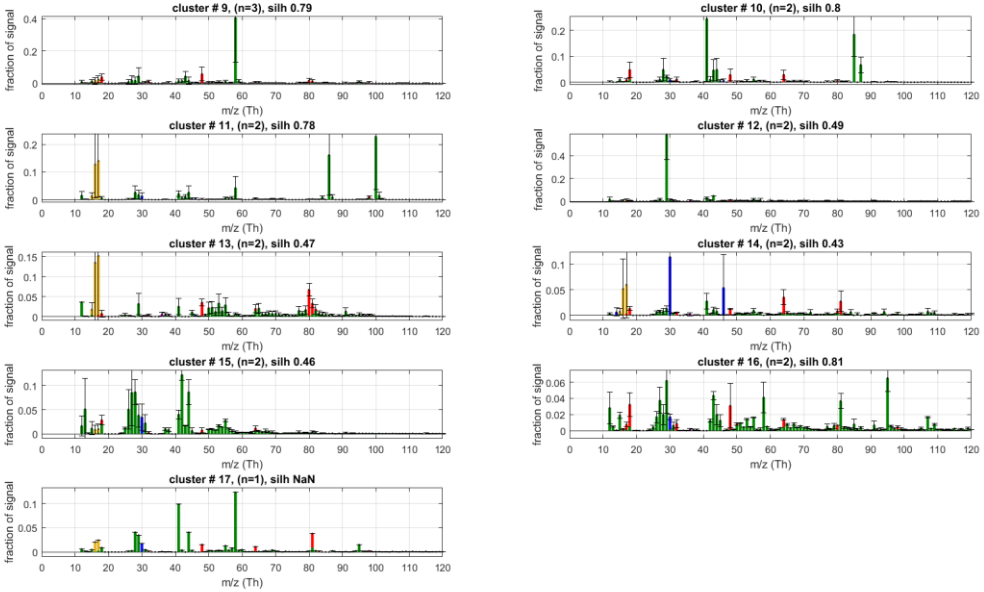


Figure 12. Spectra of outlier clusters (#9 to #17) for P-I. The spectra for these outlier classes were omitted from our analysis due to not meeting the criteria of (1) occurrence or/and (2) interpretability (on an acceptable level). Despite their mostly speculative value, many of them feature some chemically interesting characteristics, potentially pointing to presence of amines (signals at *m/z* 58, 86 and 100 Th; clusters #9, #11 and #17), alkali metals (^{85}Rb, ^{87}Rb; #10), cycloalkanes (signals at series *m/z* 69, 79, 81, 95, 107 and 109 Th; #16), and organic sulphate (signal at *m/z* 80, 81 Th; #13, #17), as well as effects of surface ionisation ($^{41}K^+$; $^{39}K^{+++}$; #10, #17) and a likely artefact from poor airbeam correction (signal at *m/z* 29 Th; #12). (red: SO_4: blue NO_3; orange: NH_4; magenta: chlorides; green: org)

3.4.1 Surface ionisation and data correction artefacts

Drewnick et al. (2015) note that the main semi-refractory elements eligible for ionisation in the AMS are Cd (*m/z* 112 Th), Cs (132 Th), Hg (200 Th), K (39 Th), Na (23 Th), Rb (85 Th) and Se (79 Th). The proneness of potassium (K) and sodium (Na) for non-quantitative thermal ionisation effects in the AMS is well known (e.g. Allan et al., 2003a), which is also why they are excluded from AMS (quantitative) data analysis. Although the main potassium isotope omitted, the ^{41}K isotope (with 6,7 % relative abundance; Haynes, 2014), is not, and a correction applied in fragmentation table instead. The K-derived signals were especially prominent in data set VIII (see S.I Fig S.7), with contributions of 1 to 2 order of magnitudes higher than highest well-behaving signals such as *m/z* 44 Th or 48 Th. We hypothesise the strong signals at *m/z* 41 Th observable in many of the outlier spectra (cluster #10, #15, and #17) may be due to insufficient accuracy of the ^{41}K isotope correction.

A similar data processing/correction artefact is likely seen in cluster #12 with a lone, dominant signal at *m/z* 29 Th. This mass-to-charge is a problematic one for lower-resolution AMS data due to contribution of $^{29}N_2$ isotopic peak, and location on the slope of the enormous N_2 peak at *m/z* 28 Th. Although the signal at *m/z* 29 Th is corrected for the (measured) isotope contribution, even a slight mismatch in the correction results in notable error in the estimation of the organic signal fraction at *m/z* 29 Th. We attribute this problem specifically to the scarce availability of filters for the earliest sets of data.

3.4.2 Alkali metals

The prominent signals at *m/z* 85 and 87 Th for cluster #10 correspond to Rubidium alkali metal ions, and their respective ratios (*m/z* 85 Th signal : *m/z* 87 Th signal = 73.2 : 26.8) to what we would expect based on isotopic distribution of Rb observed in nature (^{85}Rb : ^{87}Rb 72.2 % : 27.8 %; Haynes, 2014)). Examination of the raw mass spectrum, available in Sect S.12, also supports Rubidium as a likely candidate. Unlike for the potassium signal, the temporal behaviour of the factors corresponding to cluster #10 is highly plume-like. Preliminary analysis of wind direction shows the plume direction to correspond to the arrival direction from the district heating plant (co-located with a sawmill and a pellet factory) at Juupajoki, 5 km due south-east (Supplementary information, Sect S.12). Similar observations of Rubidium from coal burning are previously published by Irei et al. (2014). It seems likely that this aerosol class would originate from the heating plant.

3.4.3 Organic nitrogen and sulphur

As for the signals often attributed to amines at 86 and 100 Th, (Mclafferty, 1959), featured in cluster #11, in absence of alternative explanation for the 100 and 86 signals, we are inclined to believe they actually represent atmospheric amines. The cluster spectrum corresponds also to the spectra of pollution plumes, extracted for data sets I to III in our previous study on pollution events (Äijälä et al., 2017). We note amines are also reported to be prone to surface ionisation, and e.g. trimethylamine is thermally ionised above temperatures 300°C, with high thermal ionisation efficiency at 600°C (50% of the maximum efficiency observed at around 350°C; Rasulev and Zandberg, 1988)). It thus seems plausible surface ionisation effects could contribute to the amine observations as well. In our earlier work (Äijälä et al., 2017), we also attributed similar spectral signal at *m/z* 58 Th to amines ($C_3H_8N^+$). However, in light of the recent results of Drewnick et al (2015) on surface ionisation of NaCl, and the detachment of *m/z* 58 Th signal from the class of other amine-attributed signals at 86 and 100 Th, another plausible explanation for the *m/z* 58 Th signal observed in clusters #9, #11, #16 and #17 exists. Namely, we find it plausible such a spectrum would arise from surface ionisation of sodium chloride and thus represent atmospheric $NaCl^+$.
Clusters #13, #15 and 16 are interesting from the viewpoint of organonitrates and sulphates. Nitrate signal in clusters #15 and #16 is composed mostly of *m/z* 30 Th signal, with negligible *m/z* 46 Th contribution. With the high organic contribution, this would make these classes potential candidates for containing organonitrates. However, an equally plausible explanation is the surface ionisation of KNO_3, discussed previously. The pronounced signals at *m/z* 80 and/or 81 Th featured in cluster #13, #14 and #17 are likely explained by humidity-induced fragmentation changes in ionisation of sulphate species, (particularly H_2SO_4 and SO_3; Drewnick et al., 2015). We do note organosulphur-containing samples characterised by Farmer at al. (2010) also feature increased ratio of *m/z* 80 and 81 Th signals compared to $(NH_4)_2SO_4$, so we can not rule out organic sulphate contribution.

3.4.4 Cycloalkanes

Finally, we wish to draw attention to the ion series of cluster #16, with prominent organic signals at (69), 79, 81, 95, 107 and 109 Th, which have been connected to cycloalkanes (McLafferty and Turecek, 1993; Alfarra et al., 2004). Cycloalkanes are common in e.g. lubricating oils (Liang et al., 2018), which are an important, even dominant, component in traffic emissions (Worton et al., 2014). The closest literature match on ambient observations we found was the study of Takami and colleagues (2007), where they observed similar high concentrations of mass-to-charge 95, 107, 109 Th, as well as 58 and 85 Th, but were unable to attribute the observation to a specific source.

4. Conclusions

We performed a synthesis analysis on eight AMS data sets from a boreal forest site, and constructed a data-driven chemical mass balance type of receptor model, with relaxed constraints on the component profiles (r-CMB). Notably, the data comprised both inorganic and organic aerosol components. The resulting 7-component model explained 83 ± 8 % of variability in data (96 ± 3 % with low-SNR variables excluded). The model components for the SMEAR II boreal forest site were, in order of average aerosol mass contribution: Ammonium sulphate (35 ± 7 %; mean mass fraction ± standard deviation over data sets), LV-OOA (27 ± 8 %), SV-OOA (12 ± 7 %), BBOA (11 ± 7 %) Organic nitrogen (7 ± 2 %) Ammonium nitrate (5 ± 2 %) and HOA (3 ± 1 %).

Remarkably, organic nitrogen seems a larger component than ammonium nitrate for the site. However, ambiguity remains on the interpretation of the organic nitrogen class as organonitrate, prompting caution against casual use of NO_2^+:NO^+ fragmentation ratio as a sole organonitrate proxy. COA was not resolved separately, presumably due to high spectral similarity with BBOA and low mass contribution to SMEAR II aerosol, and is most likely included to the BBOA component. Other, minor aerosol groups that were not included in the model feature characteristics potentially indicative of amine-dominated aerosols, coal combustion aerosol with alkali metals (Rubidium, Cesium), as well as hints of cycloalkanes and organosulphates. We presume many of these observations may arise from by surface ionisation processes, and as such they may not be currently quantifiable in mass. Their corroboration, quantification, and connection to emission sources or thermal ionisation effects require for further study.

We suggest inorganics should be routinely included in factorisation of AMS data due to the high demand of such data in aerosol models. We wish specifically to point out that adding the inorganic information is easy and only requires applying the same tried-and-tested data processing and uses the same error model as for organics. While inclusion of inorganics does diminish the relative weight organics carry in the analysis, and thus may hinder extraction organic factors comprising very low fraction (<5 %) of total mass (Ulbrich et al., 2009), we argue that the added information value of inorganic speciation makes up for this. Compared to organics only analyses, inclusion of inorganic data increases direct usability of AMS data for physicochemical aerosol models. We also demonstrate factorisation-based speciation provide a more realistic and robust, less assumption-dependant and calibration-sensitive, speciation than simplistic ion balance schemes.

The classification methods presented here for evaluating factor analysis output can also have direct use in applications that produce large quantities of discrete aerosol spectral data, such as deriving factorisation error estimates via bootstrapping analysis (Osborne et al., 2014; Brown et al., 2015). With further development, we find it likely a two-step analysis (exploratory factorisation + classification → r-CMB) would be a viable option for increasingly unsupervised and less analyst biased AMS data analysis.

We would also encourage further development of combined statistical methods for improved mass spectral feature extraction and parametrisations for mass spectra, as they will enable future machine learning applications for data analysis. Drawing from the comprehensive information available in current size-resolved aerosol mass spectrometric data, it seems likely that advanced machine learning methods (such as data reduction combined with predictive neural networking (e.g. Burns and Whitesides, 1993; Gasteiger and Zupan, 1993) will likely provide new, improved ways to model aerosol physicochemical properties like hygroscopicity, volatility and optics in the near future.

Data availability

Data used in this study is available from the contact author. r-CMB component profiles will be made available in the AMS spectral database (http://cires1.colorado.edu/jimenez-group/AMSsd/) upon publication.

Author contribution

Conceptualisation: MÄ, ME formulated the study. *Investigation and data curation*: MÄ, HJ, ME collected and curated the experimental data. PA*, ES*, JL*, HL* TP* provided technical support for the experiments. *Formal analysis, methodology, visualisation*: MÄ, supported by KD, designed and performed the statistical analysis and data visualisations. *Validation:* MÄ and KD reviewed the data quality and reproducibility. *Software, methodology*: FC designed and supported the SoFi analysis software. DS*, LW* provided support for Squirrel software. *Writing:* MÄ wrote the original draft, which was reviewed, commented and edited by all the authors. *Funding acquisition, resources, project administration, supervision*: ME, MK, AP, TP and DW* supported and supervised the research.

Contributor roles in *italic* correspond to the taxonomy of CASRAI's CRediT definitions (https://casrai.org/credit/)

* see Acknowledgements

Competing interests

The authors declare they have no conflict of interest.

Acknowledgements

We wish to thank the technical staff at INAR and SMEAR II (Pasi Aalto, Erkki Siivola, Heikki Laakso, Toivo Pohja, Veijo Hiltunen and Janne Levula) for valuable support during the years 2008-2010, in acquiring the data sets analysed here. We thank Prof. Douglas Worsnop for pioneering work in starting the AMS studies at University of Helsinki, and the valuable insightful discussions on AMS data analysis and interpretation. We also gratefully acknowledge the friendly support staff at Aerodyne Research (especially Donna Sueper and Leah Williams) for their help on data analytical questions.

The research was supported by the following programs: the European Commission FP6 projects EUCAARI (036833-2), FP7 ACTRIS (262254), the Horizon 2020 project ACTRIS-2 (654109), ERC Grant COALA (638703), the Finnish COE project CRAICC (272041) and the Academy of Finland COE in Atmospheric Science (2008–2019).

References

Aiken, A. C., DeCarlo, P. F., and Jimenez, J. L.: Elemental analysis of organic species with electron ionization high-resolution mass spectrometry, Anal Chem, 79, 8350-8358, 10.1021/ac071150w, 2007.

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, M. R., Prevot, A. S., 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, Environ Sci Technol, 42, 4478-4485, 2008.

Alfarra, M. R., Coe, H., Allan, J. D., Bower, K. N., Boudries, H., Canagaratna, M. R., Jimenez, J. L., Jayne, J. T., Garforth, A. A., Li, S. M., and Worsnop, D. R.: Characterization of urban and rural organic particulate in the lower Fraser valley using two aerodyne aerosol mass spectrometers, Atmospheric Environment, 38, 5745-5758, 10.1016/j.atmosenv.2004.01.054, 2004.

Alfarra, M. R., Prevot, A. S., Szidat, S., Sandradewi, J., Weimer, S., Lanz, V. A., Schreiber, D., Mohr, M., and Baltensperger, U.: Identification of the mass spectral signature of organic aerosols from wood burning emissions, Environmental science & technology, 41, 5770-5777, 2007.

Allan, J. D., Alfarra, M. R., Bower, K. N., Williams, P. I., Gallagher, M. W., Jimenez, J. L., McDonald, A. G., Nemitz, E., Canagaratna, M. R., and Jayne, J. T.: Quantitative sampling using an Aerodyne aerosol mass spectrometer 2. Measurements of fine particulate chemical composition in two UK cities, Journal of Geophysical Research: Atmospheres, 108, 2003a.

Allan, J. D., Jimenez, J. L., Williams, P. I., Alfarra, M. R., Bower, K. N., Jayne, J. T., Coe, H., and Worsnop, D. R.: Quantitative sampling using an Aerodyne aerosol mass spectrometer 1. Techniques of data interpretation and error analysis, Journal of Geophysical Research: Atmospheres (1984–2012), 108, 2003b.

Allan, J. D., Delia, A. E., Coe, H., Bower, K. N., Alfarra, M. R., Jimenez, J. L., Middlebrook, A. M., Drewnick, F., Onasch, T. B., and Canagaratna, M. R.: A generalised method for the extraction of chemically resolved mass spectra from Aerodyne aerosol mass spectrometer data, Journal of Aerosol Science, 35, 909-922, 2004.

Allan, J. D., Alfarra, M. R., Bower, K. N., Coe, H., Jayne, J. T., Worsnop, D. R., Aalto, P. P., Kulmala, M., Hyotylainen, T., Cavalli, F., and Laaksonen, A.: Size and composition measurements of background aerosol and new particle growth in a Finnish forest during QUEST 2 using an Aerodyne Aerosol Mass Spectrometer, Atmospheric Chemistry and Physics, 6, 315-327, 2006.

Arthur, D., and Vassilvitskii, S.: k-means++: The advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007, 1027-1035,

Bahreini, R., Keywood, M. D., Ng, N. L., Varutbangkul, V., Gao, S., Flagan, R. C., Seinfeld, J. H., Worsnop, D. R., and Jimenez, J. L.: Measurements of secondary organic aerosol from oxidation of cycloalkenes, terpenes, and m-xylene using an Aerodyne aerosol mass spectrometer, Environmental Science & Technology, 39, 5674-5688, 10.1021/es048061a, 2005.

Ball, G. H., and Hall, D. J.: ISODATA, a novel method of data analysis and pattern classification, DTIC Document, 1965.

Berkemeier, T., Ammann, M., Mentel, T. F., Pöschl, U., and Shiraiwa, M.: Organic nitrate contribution to new particle formation and growth in secondary organic aerosols from α-pinene ozonolysis, Environmental science & technology, 50, 6334-6342, 2016.

Bilde, M., Barsanti, K., Booth, M., Cappa, C. D., Donahue, N. M., Emanuelsson, E. U., McFiggans, G., Krieger, U. K., Marcolli, C., and Topping, D.: Saturation vapor pressures and transition enthalpies of low-volatility organic molecules of atmospheric relevance: from dicarboxylic acids to complex mixtures, Chemical reviews, 115, 4115-4156, 2015.

Brown, S. G., Eberly, S., Paatero, P., and Norris, G. A.: Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results, Science of the Total Environment, 518, 626-635, 2015.

Burns, J. A., and Whitesides, G. M.: Feed-forward neural networks in chemistry: mathematical systems for classification and pattern recognition, Chemical Reviews, 93, 2583-2601, 1993.

Canagaratna, M. R., Jayne, J. T., Ghertner, D. A., Herndon, S., Shi, Q., Jimenez, J. L., Silva, P. J., Williams, P., Lanni, T., Drewnick, F., Demerjian, K. L., Kolb, C. E., and Worsnop, D. R.: Chase studies of particulate emissions from in-use New York City vehicles, Aerosol Science and Technology, 38, 555-573, 10.1080/02786820490465504, 2004.

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 Spectrom Rev, 26, 185-222, 10.1002/mas.20115, 2007.

Canonaco, F., Crippa, M., Slowik, J. G., Baltensperger, U., and Prevot, 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., Slowik, J. G., Baltensperger, U., and Prevot, A. S. H.: Seasonal differences in oxygenated organic aerosol composition: implications for emissions sources and factor analysis, Atmospheric Chemistry and Physics, 15, 6993-7002, 10.5194/acp-15-6993-2015, 2015.

Chirico, R., DeCarlo, P., Heringa, M., Tritscher, T., Richter, R., Prévôt, A., Dommen, J., Weingartner, E., Wehrle, G., and Gysel, M.: Impact of aftertreatment devices on primary emissions and secondary organic aerosol formation potential from in-use diesel vehicles: results from smog chamber experiments, Atmospheric Chemistry & Physics Discussions, 10, 2010.

Corrigan, A., Russell, L., Takahama, S., Äijälä, M., Ehn, M., Junninen, H., Rinne, J., Petäjä, T., Kulmala, M., and Vogel, A.: Biogenic and biomass burning organic aerosol in a boreal forest at Hyytiälä, Finland, during HUMPPA-COPEC 2010, Atmospheric Chemistry and Physics, 13, 12233-12256, 2013.

Crippa, M., DeCarlo, P. F., Slowik, J. G., Mohr, C., Heringa, M. F., Chirico, R., Poulain, L., Freutel, F., Sciare, J., Cozic, J., Di Marco, C. F., Elsasser, M., Nicolas, J. B., Marchand, N., Abidi, E., Wiedensohler, A., Drewnick, F., Schneider, J., Borrmann, S., Nemitz, E., Zimmermann, R., Jaffrezo, J. L., Prevot, A. S. H., and Baltensperger, U.: Wintertime aerosol chemical composition and source apportionment of the organic fraction in the metropolitan area of Paris, Atmospheric Chemistry and Physics, 13, 961-981, 10.5194/acp-13-961-2013, 2013.

Crippa, M., Canonaco, F., Lanz, V. A., Aijala, M., Allan, J. D., Carbone, S., Capes, G., Ceburnis, D., Dall'Osto, M., Day, D. A., 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., Mohr, C., Nemitz, E., O'Dowd, C., Ovadnevaite, J., Pandis, S. N., Petaja, T., Poulain, L., Saarikoski, S., Sellegri, K., Swietlicki, E., Tiitta, P., Worsnop, D. R., Baltensperger, U., and Prevot, 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, 10.5194/acp-14-6159-2014, 2014.

Cubison, M. J., Ortega, A. M., Hayes, P. L., Farmer, D. K., Day, D., Lechner, M. J., Brune, W. H., Apel, E., Diskin, G. S., Fisher, J. A., Fuelberg, H. E., Hecobian, A., Knapp, D. J., Mikoviny, T., Riemer, D., Sachse, G. W., Sessions, W., Weber, R. J., Weinheimer, A. J., 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, 10.5194/acp-11-12049-2011, 2011.

Daellenbach, K. R., Stefenelli, G., Bozzetti, C., Vlachou, A., Fermo, P., Gonzalez, R., Piazzalunga, A., Colombi, C., Canonaco, F., and Hueglin, C.: 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.

Drewnick, F., Hings, S. S., DeCarlo, P., Jayne, J. T., Gonin, M., Fuhrer, K., Weimer, S., Jimenez, J. L., Demerjian, K. L., Borrmann, S., and Worsnop, D. R.: A new time-of-flight aerosol mass spectrometer (TOF-AMS) - Instrument description and first field deployment, Aerosol Science and Technology, 39, 637-658, 10.1080/02786820500182040, 2005.

Drewnick, F., Diesch, J.-M., Faber, P., and Borrmann, S.: Aerosol mass spectrometry: particle–vaporizer interactions and their consequences for the measurements, Atmospheric Measurement Techniques, 8, 3811-3830, 2015.

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, Atmos. Chem. Phys., 11, 1155-1165, 10.5194/acp-11-1155-2011, 2011.

Elsasser, M., Crippa, M., Orasche, J., DeCarlo, P. F., Oster, M., Pitz, M., Cyrys, J., Gustafson, T. L., Pettersson, J. B. C., Schnelle-Kreis, J., Prevot, A. S. H., and Zimmermann, R.: Organic molecular markers and signature from wood combustion particles in winter ambient aerosols: aerosol mass spectrometer (AMS) and high time-resolved GC-MS measurements in Augsburg, Germany, Atmospheric Chemistry and Physics, 12, 6113-6128, 10.5194/acp-12-6113-2012, 2012.

Farmer, D. K., Matsunaga, A., Docherty, K. S., Surratt, J. D., Seinfeld, J. H., Ziemann, P. J., and Jimenez, J. L.: Response of an aerosol mass spectrometer to organonitrates and organosulfates and implications for atmospheric chemistry, Proc Natl Acad Sci U S A, 107, 6670-6675, 10.1073/pnas.0912340107, 2010.

Fortier, J., and Solomon, H.: Clustering procedures, Multivariate analysis, 1966.

Fountoukis, C., and Nenes, A.: ISORROPIA II: a computationally efficient thermodynamic equilibrium model for K+–Ca 2+–Mg 2+–NH 4+–Na+–SO 4 2−–NO 3−–Cl−–H 2 O aerosols, Atmospheric Chemistry and Physics, 7, 4639-4659, 2007.

Freutel, F., Drewnick, F., Schneider, J., Klimach, T., and Borrmann, S.: Quantitative single-particle analysis with the Aerodyne aerosol mass spectrometer: development of a new classification algorithm and its application to field data, Atmospheric Measurement Techniques, 6, 3131-3145, 10.5194/amt-6-3131-2013, 2013.

Friedlander, S. K.: Chemical element balances and identification of air pollution sources, Environmental Science & Technology, 7, 235-240, 1973.

Gasteiger, J., and Zupan, J.: Neural networks in chemistry, Angewandte Chemie International Edition in English, 32, 503-527, 1993.

Gordon, G. E.: Receptor models, Environmental Science & Technology, 22, 1132-1142, 1988.

Gysel, M., Crosier, J., Topping, D., Whitehead, J., Bower, K., Cubison, M., Williams, P., Flynn, M., McFiggans, G., and Coe, H.: Closure study between chemical composition and hygroscopic growth of aerosol particles during TORCH2, Atmospheric Chemistry and Physics, 7, 6131-6144, 2007.

Hao, L., Kortelainen, A., Romakkaniemi, S., Portin, H., Jaatinen, A., Leskinen, A., Komppula, M., Miettinen, P., Sueper, D., and Pajunoja, A.: Atmospheric submicron aerosol composition and particulate organic nitrate formation in a boreal forestland–urban mixed region, Atmospheric Chemistry and Physics, 14, 13483-13495, 2014.

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

Haynes, W. M.: CRC handbook of chemistry and physics, CRC press, 2014.

Hersey, S. P., Craven, J. S., Schilling, K. A., Metcalf, A. R., Sorooshian, A., Chan, M. N., Flagan, R. C., and Seinfeld, J. H.: The Pasadena Aerosol Characterization Observatory (PACO): chemical and physical analysis of the Western Los Angeles basin aerosol, Atmospheric Chemistry and Physics, 11, 7417-7443, 10.5194/acp-11-7417-2011, 2011.

Hoffmann, D., Tilgner, A., Iinuma, Y., and Herrmann, H.: Atmospheric stability of levoglucosan: a detailed laboratory and modeling study, Environmental science & technology, 44, 694-699, 2009.

Hong, J., Äijälä, M., Häme, S. A., Hao, L., Duplissy, J., Heikkinen, L. M., Nie, W., Mikkilä, J., Kulmala, M., and Prisle, N. L.: Estimates of the organic aerosol volatility in a boreal forest using two independent methods, Atmospheric Chemistry and Physics, 17, 4387-4399, 2017.

Hopke, P. K.: Receptor modeling for air quality management, Elsevier, 1991.

Hopke, P. K.: Review of receptor modeling methods for source apportionment, Journal of the Air & Waste Management Association, 66, 237-259, 2016.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., and Nishioka, T.: MassBank: a public repository for sharing mass spectral data for life sciences, J Mass Spectrom, 45, 703-714, 10.1002/jms.1777, 2010.

Irei, S., Shimono, A., Hikida, T., Kuramoto, K., Suzuki, Y., and Takami, A.: Qualitative evaluation of m/z 85, 87, and 133 signals in organic aerosol mass spectra of fly ash produced by coal combustion, Aerosol and Air Quality Research, 14, 406-412, 2014.

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

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S., 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.

Kiendler-Scharr, A., Mensah, A. A., Friese, E., Topping, D., Nemitz, E., Prevot, A., Äijälä, M., Allan, J., Canonaco, F., and Canagaratna, M.: Ubiquity of organic nitrates from nighttime chemistry in the European submicron aerosol, Geophysical Research Letters, 43, 7735-7744, 2016.

Kim, S., Koo, I., Wei, X., and Zhang, X.: A method of finding optimal weight factors for compound identification in gas chromatography-mass spectrometry, Bioinformatics, 28, 1158-1163, 10.1093/bioinformatics/bts083, 2012.

Kowalski, B. R.: Measurement analysis by pattern recognition, Analytical Chemistry, 47, 1152A-1162a, 1975.

Kroll, J. H., and Seinfeld, J. H.: Chemistry of secondary organic aerosol: Formation and evolution of low-volatility organics in the atmosphere, Atmospheric Environment, 42, 3593-3624, 2008.

Kulmala, M., Rannik, Ü., Pirjola, L., Dal Maso, M., Karimäki, J., Asmi, A., Jäppinen, A., Karhu, V., Korhonen, H., and Malvikko, S.-P.: Characterization of atmospheric trace gas and aerosol concentrations at forest sites in southern and northern Finland using back trajectories, Boreal environment research, 5, 315-336, 2000.

Kulmala, M., Asmi, A., Lappalainen, H., Carslaw, K., Pöschl, U., Baltensperger, U., Hov, Ø., Brenquier, J.-L., Pandis, S. N., and Facchini, M.: Introduction: European Integrated Project on Aerosol Cloud Climate and Air Quality interactions (EUCAARI)–integrating aerosol research from nano to global scales, Atmospheric Chemistry and Physics, 9, 2825-2841, 2009.

Kulmala, M., Asmi, A., Lappalainen, H. K., Baltensperger, U., Brenguier, J. L., Facchini, M. C., Hansson, H. C., Hov, O., O'Dowd, C. D., Poschl, U., Wiedensohler, A., Boers, R., Boucher, O., de Leeuw, G., van der Gon, H. A. C. D., Feichter, J., Krejci, R., Laj, P., Lihavainen, H., Lohmann, U., McFiggans, G., Mentel, T., Pilinis, C., Riipinen, I., Schulz, M., Stohl, A., Swietlicki, E., Vignati, E., Alves, C., Amann, M., Ammann, M., Arabas, S., Artaxo, P., Baars, H., Beddows, D. C. S., Bergstrom, R., Beukes, J. P., Bilde, M., Burkhart, J. F., Canonaco, F., Clegg, S. L., Coe, H., Crumeyrolle, S., D'Anna, B., Decesari, S., Gilardoni, S., Fischer, M., Fjaeraa, A. M., Fountoukis, C., George, C., Gomes, L., Halloran, P., Hamburger, T., Harrison, R. M., Herrmann, H., Hoffmann, T., Hoose, C., Hu, M., Hyvarinen, A., Horrak, U., Iinuma, Y., Iversen, T., Josipovic, M., Kanakidou, M., Kiendler-Scharr, A., Kirkevag, A., Kiss, G., Klimont, Z., Kolmonen, P., Komppula, M., Kristjansson, J. E., Laakso, L., Laaksonen, A., Labonnote, L., Lanz, V. A., Lehtinen, K. E. J., Rizzo, L. V., Makkonen, R., Manninen, H. E., McMeeking, G., Merikanto, J., Minikin, A., Mirme, S., Morgan, W. T., Nemitz, E., O'Donnell, D., Panwar, T. S., Pawlowska, H., Petzold, A., Pienaar, J. J., Pio, C., Plass-Duelmer, C., Prevot, A. S. H., Pryor, S., Reddington, C. L., Roberts, G., Rosenfeld, D., Schwarz, J., Seland, O., Sellegri, K., Shen, X. J., Shiraiwa, M., Siebert, H., Sierau, B., Simpson, D., Sun, J. Y., Topping, D., Tunved, P., Vaattovaara, P., Vakkari, V., Veefkind, J. P., Visschedijk, A., Vuollekoski, H., Vuolo, R., Wehner, B., Wildt, J., Woodward, S., Worsnop, D. R., van Zadelhoff, G. J., Zardini, A. A., Zhang, K., van Zyl, P. G., Kerminen, V. M., Carslaw, K. S., and Pandis, S. N.: General overview: European Integrated project on Aerosol Cloud Climate and Air Quality interactions (EUCAARI) - integrating aerosol research from nano to global scales, Atmospheric Chemistry and Physics, 11, 13061-13143, 10.5194/acp-11-13061-2011, 2011.

Lanz, V. A., Alfarra, M. R., Baltensperger, U., Buchmann, B., Hueglin, C., and Prevot, 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, 2007a.

Lanz, V. A., Alfarra, M. R., Baltensperger, U., Buchmann, B., Hueglin, C., Szidat, S., Wehrli, M. N., Wacker, L., Weimer, S., and Caseiro, A.: Source attribution of submicron organic aerosols during wintertime inversions by advanced factor analysis of aerosol mass spectra, Environmental science & technology, 42, 214-220, 2007b.

Lee, B. H., Mohr, C., Lopez-Hilfiker, F. D., Lutz, A., Hallquist, M., Lee, L., Romer, P., Cohen, R. C., Iyer, S., and Kurtén, T.: Highly functionalized organic nitrates in the southeast United States: Contribution to secondary organic aerosol and reactive nitrogen budgets, Proceedings of the National Academy of Sciences, 113, 1516-1521, 2016.

Li, J., Pósfai, M., Hobbs, P. V., and Buseck, P. R. J. J. o. G. R. A.: Individual aerosol particles from biomass burning in southern Africa: 2, Compositions and aging of inorganic particles, 108, 2003.

Liang, Z., Chen, L., Alam, M. S., Rezaei, S. Z., Stark, C., Xu, H., and Harrison, R. M.: Comprehensive chemical characterization of lubricating oils used in modern vehicular engines utilizing GC× GC-TOFMS, Fuel, 220, 792-799, 2018.

Lindon, J. C., Tranter, G. E., and Koppenaal, D.: Encyclopedia of spectroscopy and spectrometry, Academic Press, 2016.

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.

MacQueen, J.: Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967, 281-297,

Mclafferty, F. W.: Mass Spectrometric Analysis - Molecular Rearrangements, Analytical Chemistry, 31, 82-87, DOI 10.1021/ac60145a015, 1959.

McLafferty, F. W., and Turecek, F.: Interpretation of mass spectra, University science books, 1993.

Mcquitty, L. L.: Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data, Educational and Psychological Measurement, 26, 825-&, Doi 10.1177/001316446602600402, 1966.

Miller, M., Friedlander, S., and Hidy, G.: A chemical element balance for the Pasadena aerosol, in: Aerosols and atmospheric chemistry, Elsevier, 301-312, 1972.

Mohr, C., Huffman, A., Cubison, M. J., Aiken, A. C., Docherty, K. S., Kimmel, J. R., Ulbrich, I. M., Hannigan, M., and Jimenez, J. L.: Characterization of primary organic aerosol emissions from meat cooking, trash burning, and motor vehicles with high-resolution aerosol mass spectrometry and comparison with ambient and chamber observations, Environ Sci Technol, 43, 2443-2449, 2009.

Mohr, C., DeCarlo, P. F., Heringa, M. F., Chirico, R., Slowik, J. G., Richter, R., Reche, C., Alastuey, A., Querol, X., Seco, R., Penuelas, J., Jimenez, J. L., Crippa, M., Zimmermann, R., Baltensperger, U., and Prevot, A. S. H.: Identification and quantification of organic aerosol from cooking and other sources in Barcelona using aerosol mass spectrometer data, Atmospheric Chemistry and Physics, 12, 1649-1665, 10.5194/acp-12-1649-2012, 2012.

Murphy, S. M., Sorooshian, A., Kroll, J. H., Ng, N. L., Chhabra, P., Tong, C., Surratt, J. D., Knipping, E., Flagan, R. C., and Seinfeld, J. H.: Secondary aerosol formation from atmospheric reactions of aliphatic amines, Atmospheric Chemistry and Physics, 7, 2313-2337, 2007.

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., Prevot, 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., Zhang, Q., Ulbrich, I. M., and Worsnop, D. R.: Real-time methods for estimating organic component mass concentrations from aerosol mass spectrometer data, Environ Sci Technol, 45, 910-916, 10.1021/es102951k, 2011.

Niemi, J. V., Saarikoski, S., Aurela, M., Tervahattu, H., Hillamo, R., Westphal, D. L., Aarnio, P., Koskentalo, T., Makkonen, U., Vehkamaki, H., and Kulmala, M.: Long-range transport episodes of fine particles in southern Finland during 1999-2007, Atmospheric Environment, 43, 1255-1264, 10.1016/j.atmosenv.2008.11.022, 2009.

Orlando, J. J., and Tyndall, G. S.: Laboratory studies of organic peroxy radical chemistry: an overview with emphasis on recent issues of atmospheric significance, Chemical Society Reviews, 41, 6294-6317, 2012.

Osborne, J. W., Costello, A. B., and Kellow, J. T.: Best practices in exploratory factor analysis, CreateSpace Independent Publishing Platform Charleston, SC, 2014.

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, 1994.

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

Paatero, P.: The multilinear engine - A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model, Journal of Computational and Graphical Statistics, 8, 854-888, Doi 10.2307/1390831, 1999.

Paatero, P.: User's guide for positive matrix factorization programs PMF2 and PMF3, Helsinki: University of Helsinki, 2000.

Patokoski, J., Ruuskanen, T. M., Kajos, M. K., Taipale, R., Rantala, P., Aalto, J., Ryyppo, 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, 10.5194/acp-15-13413-2015, 2015.

Peräkylä, O., Vogt, M., Tikkanen, O.-P., Laurila, T., Kajos, M. K., Rantala, P. A., Patokoski, J., Aalto, J., Yli-Juuti, T., and Ehn, M.: Monoterpenes' oxidation capacity and rate over a boreal forest, Boreal Environment Research, 2014.

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., and Vlachou, A.: Inorganic salt interference on CO2+ in aerodyne AMS and ACSM organic aerosol composition studies, Environmental science & technology, 50, 10494-10503, 2016.

Rasulev, U. K., and Zandberg, E. Y.: Surface ionization of organic compounds and its applications, Progress in Surface Science, 28, 181-412, 1988.

Rebotier, T. P., and Prather, K. A.: Aerosol time-of-flight mass spectrometry data analysis: a benchmark of clustering algorithms, Anal Chim Acta, 585, 38-54, 10.1016/j.aca.2006.12.009, 2007.

Reilly, P., and Wood, R. H.: Prediction of the properties of mixed electrolytes from measurements on common ion mixtures, The Journal of Physical Chemistry, 73, 4292-4297, 1969.

Robinson, N., Hamilton, J., Allan, J., Langford, B., Oram, D., Chen, Q., Docherty, K., Farmer, D., Jimenez, J., and Ward, M.: Evidence for a significant proportion of Secondary Organic Aerosol from isoprene above a maritime tropical forest, Atmospheric Chemistry and Physics, 11, 1039-1050, 2011.

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

Schneider, J., Weimer, S., Drewnick, F., Borrmann, S., Helas, G., Gwaze, P., Schmid, O., Andreae, M. O., and Kirchner, U.: Mass spectrometric analysis and aerodynamic properties of various types of combustion-related aerosol particles, International Journal of Mass Spectrometry, 258, 37-49, 10.1016/j.ijms.2006.07.008, 2006.

Shafizadeh, F.: The chemistry of pyrolysis and combustion, The chemistry of solid wood, 489-529, 1984.

Sofowote, U. M., Su, Y., Dabek-Zlotorzynska, E., Rastogi, A. K., Brook, J., and Hopke, P. K.: Constraining the factor analytical solutions obtained from multiple-year receptor modeling of ambient PM2. 5 data from five speciation sites in Ontario, Canada, Atmospheric Environment, 108, 151-157, 2015.

Sogacheva, L., Dal Maso, M., Kerminen, V.-M., and Kulmala, M.: Probability of nucleation events and aerosol particle concentration in different air mass types arriving at Hyytiälä, southern Finland, based on back trajectories analysis, Boreal environment research, 10, 479-491, 2005.

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

Steinhaus, H.: Sur la division des corp materiels en parties, Bull. Acad. Polon. Sci, 1, 801-804, 1956.

Sun, Y., Zhang, Q., Schwab, J., Yang, T., Ng, N., and Demerjian, K.: Factor analysis of combined organic and inorganic aerosol mass spectra from high resolution aerosol mass spectrometer measurements, Atmospheric Chemistry and Physics, 12, 8537-8551, 2012.

Swietlicki, E., Hansson, H.-C., Hämeri, K., Svenningsson, B., Massling, A., McFiggans, G., McMurry, P., Petäjä, T., Tunved, P., and Gysel, M.: Hygroscopic properties of submicrometer atmospheric aerosol particles measured with H-TDMA instruments in various environments—a review, Tellus B: Chemical and Physical Meteorology, 60, 432-469, 2008.

Takami, A., Miyoshi, T., Shimono, A., Kaneyasu, N., Kato, S., Kajii, Y., and Hatakeyama, S.: Transport of anthropogenic aerosols from Asia and subsequent chemical transformation, Journal of Geophysical Research: Atmospheres, 112, 2007.

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.

Williams, J., Crowley, J., Fischer, H., Harder, H., Martinez, M., Petaja, T., Rinne, J., Back, J., Boy, M., Dal Maso, 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., 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., de Arellano, J. V. G., 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, 11, 10599-10618, 10.5194/acp-11-10599-2011, 2011.

Vogel, A. L., Äijälä, M., Corrigan, A. L., Junninen, H., Ehn, M., Petäjä, T., Worsnop, D. R., Kulmala, M., Russell, L. M., and Williams, J.: In situ submicron organic aerosol characterization at a boreal forest research station during HUMPPA-COPEC 2010 using soft and hard ionization mass spectrometry, Atmospheric Chemistry and Physics, 13, 10933-10950, 2013.

Worton, D. R., Isaacman, G., Gentner, D. R., Dallmann, T. R., Chan, A. W., Ruehl, C., Kirchstetter, T. W., Wilson, K. R., Harley, R. A., and Goldstein, A. H.: Lubricating oil dominates primary organic aerosol emissions from motor vehicles, Environmental science & technology, 48, 3698-3706, 2014.

Yatavelli, R. L., Mohr, C., Stark, H., Day, D. A., Thompson, S. L., Lopez-Hilfiker, F. D., Campuzano-Jost, P., Palm, B. B., Vogel, A. L., and Hoffmann, T.: Estimating the contribution of organic acids to northern hemispheric continental organic aerosol, Geophysical Research Letters, 42, 6084-6090, 2015.

Ye, Q., Robinson, E. S., Ding, X., Ye, P., Sullivan, R. C., and Donahue, N. M.: Mixing of secondary organic aerosols versus relative humidity, Proceedings of the National Academy of Sciences, 113, 12649-12654, 2016.

Zhang, Q., Alfarra, M. R., Worsnop, D. R., Allan, J. D., Coe, H., Canagaratna, M. R., and Jimenez, J. L.: Deconvolution and quantification of hydrocarbon-like and oxygenated organic aerosols based on aerosol mass spectrometry, Environ Sci Technol, 39, 4938-4952, 2005.

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, Anal Bioanal Chem, 401, 3045-3067, 10.1007/s00216-011-5355-y, 2011.

Zieger, P., Aalto, P., Aaltonen, V., Äijälä, M., Backman, J., Hong, J., Komppula, M., Krejci, R., Laborde, M., and Lampilahti, J.: Low hygroscopic scattering enhancement of boreal aerosol and the implications for a columnar optical closure study, Atmospheric Chemistry and Physics, 15, 7247-7267, 2015.

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.

S.1 Chemical Mass Balance model

Chemical mass balance (CMB) models are typically used as receptor models for cases where source profiles are known, and only the mass loading information needs resolving (Friedlander, 1973; Gordon, 1988;Hopke, 1991a; Miller et al., 1972). In receptor modelling based on mass conservation, the observed loading x of variable v at sampling time t can be modelled as a sum of contributions from all sources $s = 1 \ldots p$ (e.g. Hopke, 1991b):

$$x_{v,t} = \sum_{s=1}^{p} f_{v,s} m_{s,t} + \varepsilon_{v,t}. \qquad \text{(S.1)}$$

The contribution from a single source ($f_{v,s} m_{s,t}$) is the total mass, $m_{s,t}$, emitted from that source multiplied by the fraction of variable v in the emission profile of said source, $f_{v,s}$. The residual $\varepsilon_{v,t}$ encompasses the difference between observation and the model.

S.2 Mass scaling optimisation

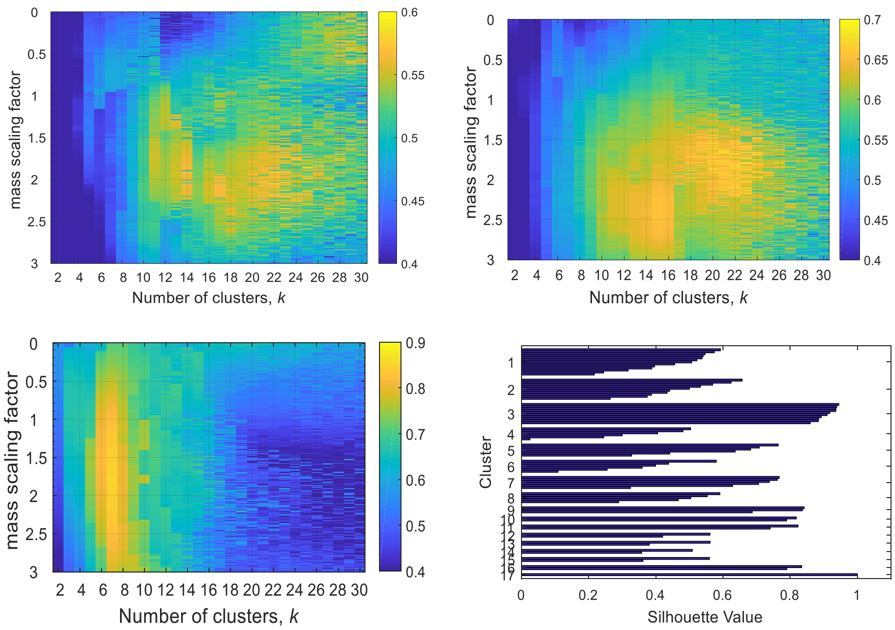


Figure S.1. Clustering silhouette scans for phases I (upper left), II (upper right) and III (lower left) panels. x-axis comprises cluster number (k) and y-axis the mass scaling factor s_m (Equation 4). Brighter colour (higher silhouette) indicates better solution quality. Note the different colour scales in panels. Stable silhouette maxima (i.e. not sensitive to minor, 0.01 changes in s_m) solutions were preferred. The solutions we selected were for P-I: k=17, s_m=2.12, for P-II: k=15, s_m=2.41 and for P-III: k=7, s_m=1.81). In lower right panel: object specific silhouettes (s_i) by cluster for the selected P-I solution.

Beside the selected solutions described in Sections 3.1.1 to 3.1.3, there are other solutions of mathematically (near) equal quality available, resulting in slightly different classifications. When deciding on clustering solutions, we also explored the highest, alternative local maxima. While a comprehensive description of the clustering solution space is out of the scope of what we can present here, we note the main differences between the solutions relate to outlier groups and how they can be combined or separated from the main classes. In phase I, the other viable solution would have been the maxima at k = 14 (s_m = 1.88 gives approximately the same classification as the selected solution at k = 17 (s_m = 2.12). We selected the latter solution since it yields a higher silhouette a) overall (0.5645 vs 0.5628) and b) for the AN cluster specifically (0.48 vs 0.35).

For both P-I and P-II, there exist a number of unstable (in relation to minor s_m change), high k solutions. However, upon inspection, their high silhouettes (compared to lower k solutions) seem to result from division of outlier clusters (n = 2…3, with low within-cluster silhouettes), into singleton (n = 1) clusters. We find that including solutions with such further divisions does not incorporate new information, but rather smears the picture when aiming to interpret the outlier clusters (Section 3.4).

S.3 Ion pairing scheme

In the pairing schemes of Gysel et al. (2007) and Hong et al. (2017), nitrate, sulphate or ammonium are not allowed to exist separately from inorganic salts (e.g. organonitrate, organic sulphate or organic acid salt with ammonium). Since there is evidence of especially organonitrates representing a considerable fraction of total particulate nitrate (e.g. Kiendler-Scharr et al., 2016), we modify the Hong et al. scheme to allow organic nitrate and “excess” ammonium (i.e. NH_4 existing with other species, such as organics). Organic sulphate was not included since it is not possible to differentiate between the two possible forms of “excess sulphur” (from the perspective of inorganic salt balance): sulphuric acid and organosulphate. To maintain compatibility with previous schemes, a sulphuric acid class was included. Any potential organosulphates would also fall into the sulphuric acid class.

Our ion balance scheme, modified from Hong et al., (2017), is divided into three cases:

Case 1: Aerosol is acidic, not enough [NH_4^+] to convert all sulphate to NH_4HSO_4. Sulphate exists as sulphuric acid and ammonium bisulphate. All nitrate is organic, as sulphate formation is assumed preferential. All [NH_4^+] goes to ammonium bisulphate.

$$\frac{[NH_4^+]}{[SO_4^{2-}]} < 1$$

sulphuric acid: $$[\mathrm{H_2SO_4}] = [SO_4^{2-}] - [NH_4^+] \qquad \text{(S.2)}$$

ammonium bisulphate: $$[\mathrm{NH_4}HSO_4] = [NH_4^+]$$

ammonium sulphate: $[(\mathrm{NH_2})\mathrm{SO_4}] = 0$

ammonium nitrate: $[\mathrm{NH_4NO_3}] = 0$

organonitrate: $[org\ \mathrm{NO_3}] = [NO_3^-]$

excess ammonium: $[\text{excess NH}_4] = 0$

Case 2: Aerosol is partly neutralised, sulphate exists as ammonium bisulphate and ammonium sulphate. Any nitrate is organic. All of [NH_4^+] associates with sulphate.

$$1 \le \frac{[NH_4^+]}{[SO_4^{2-}]} < 2$$

sulphuric acid: $[\mathrm{H_2SO_4}] = 0$ (S.3)

ammonium bisulphate: $[\mathrm{NH_4}H\mathrm{SO_4}] = \left(2 - \frac{[NH_4^+]}{[SO_4^{2-}]}\right) * [SO_4^{2-}]$

ammonium sulphate: $[(\mathrm{NH_4})\mathrm{SO_4}] = (\frac{[NH_4^+]}{[SO_4^{2-}]} - 1) * [SO_4^{2-}]$

ammonium nitrate: $[\mathrm{NH_4NO_3}] = 0$

organonitrate: $[org\ \mathrm{NO_3}] = [NO_3^-]$

excess ammonium: $[\text{excess NH}_4] = 0$

Case 3: Aerosol is fully neutralised, sulphate exists as ammonium sulphate. Leftover [NH_4^+] from sulphate neutralisation goes with [NO_3^-]. In case all of [NO_3^-] can be neutralised by [NH_4^+], leftover [NH_4^+] is labelled "excess NH_4^+", and assumed to bind with other species (e.g. organic acids).

$$2 < \frac{[NH_4^+]}{[SO_4^{2-}]}$$

sulphuric acid: $[\mathrm{H_2SO_4}] = 0$ (S.4)

ammonium bisulphate: $[\mathrm{NH_4}H\mathrm{SO_4}] = 0$

ammonium sulphate: $[(\mathrm{NH_2})\mathrm{SO_4}] = [SO_4^{2-}]$

ammonium nitrate: $[\mathrm{NH_4NO_3}] = \max([NO_3^-], [NH_4^+] - 2 * [SO_4^{2-}])$

organonitrate: $[org\ \mathrm{NO_3}] = [NO_3^-] - [\mathrm{NH_4NO_3}]$

excess ammonium: $[\text{excess NH}_4] = \min(0, [NH_4^+] - 2 * [SO_4^{2-}] - [NO_3^-])$

We echo Hong et al. (2017) in emphasising that this approximation is only a first estimate of the dynamics of inorganic salt formation; the assumption of perfect internal mixing is likely unrealistic for atmospheric aerosol particles, due to the nature of atmospheric processing often affecting the particle outer surface, resulting in layered structures (e.g. Buajarern et al., 2007). The model also does not adequately describe the thermodynamic equilibriums of salt formation or model the competition between [SO_4^{2-}] and [NO_3^-] for [NH_4^+] realistically.

S.4 Relative ionisation efficiencies (RIEs) and ion pairing scheme sensitivity to RIE_{NH4} changes

To convert from nitrate equivalent mass units (used in factorisation steps until P-III), the factorisation output data needs to be corrected for species-specific Relative Ionisation Efficiency (RIE; Allan et al., 2003; Jimenez et al., 2003). We thus summed the signals from individual ion species, and corrected them for RIE. Due to high uncertainties (standard deviation 51% of mean RIE) and for some data sets unavailability of relative ionisation efficiency calibrations for NH_4, we generally did not apply RIE correction for data sets individually, but instead used the default value of RIE_{NH4} of 4.0 for AMS instruments. Ammonium sulphate calibrations were not routinely performed, so RIE_{SO4} data was not available – we thus used the default RIE_{SO4} of 1.2. For September 2008 (data set II) we had IE and RIE_{NH4} data available, but it similarly showed a very high variability between subsequent estimates for RIE. For the 12 brute-force-single-particle (BFSP; Drewnick et al., 2015) type of calibration results available, the average ± standard deviation of RIE observed was 3.39 ± 1.20 (± 35 %; min 1.03, max 5.83). Especially for BFSP-type NH_4NO_3 calibration, uncertainty and/or bias for RIE_{NH4} is high, due to low ions-per-particle (IPP) rate for NH_4 fragment ions, and sensitivity to threshold set for the detector (AMS best practices, Jayne, 2018).

To evaluate results for our inorganic r-CMB components, we calculated ion balance ratios for AN and AS, and compared them with theoretical values derived from compound molecular weights. The theoretical mass ratio of NH_4 to NO_3 for NH_4NO_3 is 0.29. With default RIE_{NH4} of 4.0 the r-CMB AN factor yields a ratio of 0.46, and with RIE_{NH4} from calibrations (3.39 ± 1.20), an average ratio of 0.54 (min 0.40, max 0.84). Results by Sun et al. (2012) produce a value of 0.36, so overall there remains discrepancy between the theoretical and observed ratios. The reason for it remains unclear, although one should note that the uncertainties involved in such estimates are generally very large if RIE uncertainty is high. The respective theoretical mass ratio for ammonium sulphate, NH_4 : SO_4, is between 0.18 (NH_4HSO_4) and 0.36 ($(NH_4)_2SO_4$). Our AS r-CMB component features a ratio of 0.20 for RIE_{NH4} = 4.0, and 0.24 for RIE_{NH4} = 3.39. The min/max estimates due to uncertainty of RIE_{NH4} (± 1.20) are (0.17…0.39). The ratio reported by Sun et al (2012) is 0.34.

As RIE is not incorporated in the factorisation phase, uncertainty in RIE does not play a role in PMF/ME-2 outcome. We did briefly test the sensitivity of the inorganic apportionment scheme to changes in RIE of NH_4 ion species, by running the model with RIE_{NH4} ± 33% for the September 2008 data set (data set III; also presented in Section 3.2.2). The resulting model output changes are shown below, in Figure S.2 and Table S.1.

Table S.1. RIE$_{NH4}$ sensitivity test results for the inorganic apportionment scheme. ± estimate is given as ((max-min) / 2) / default result.

RIE$_{NH4}$	AS (μg/m^3)	AN (μg/m^3)	orgNO$_3$ (μg/m^3)	excess_NH$_4$ (μg/m^3)
4.00 (default)	1.562	0.139	0.105	0.032
2.68 (- 33 %)	1.473	0.031	0.202	0.003
5.32 (+ 33 %)	1.620	0.186	0.063	0.108
+/-	4.7 %	55.8 %	66.2 %	164.1 %

It thus seems the inorganic model AS component is rather insensitive to 33 % change in RIE$_{NH4}$. However, the nitrate components are much more sensitive (AN ± 56 % and orgNO$_3$ ± 66 %), and the sensitivity of "excess_NH$_4$" class to a 33 % change in RIE$_{NH4}$ is extremely high (± 164 % in mass concentrations). The changes, however, seem connected mostly to differences in scaling rather than different temporal variability, so the RIE$_{NH4}$ sensitivity does not seem to explain the difference between the inorganic ion balance scheme and the r-CMB model result. While the relative uncertainty becomes large for AN, orgNO$_3$ and excess_NH$_4$, we note this may be partly due to the generally low nitrate concentrations at SMEAR II, and the overall large uncertainty (arising from other error sources, e.g. fragmentation table calculations) in measuring low concentrations of NH$_4$ typical of the site.

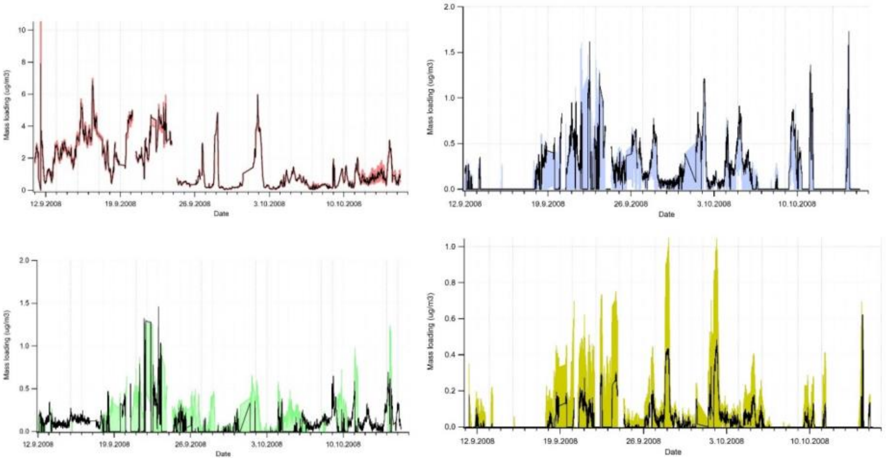


Figure S.2. RIE$_{NH4}$ sensitivity test results for the inorganic apportionment model (Sect 2.4.1), tested for data set III. Black line indicates the result with default RIE$_{NH4}$ of 4.0. Coloured areas corresponds to the min/max results for when RIE$_{NH4}$ was changed by ± 33 %. Components are colour-coded similar to main text Figure 10. (red: AS, blue: AN, green: orgNO$_3$, yellow: excess NH$_4$).

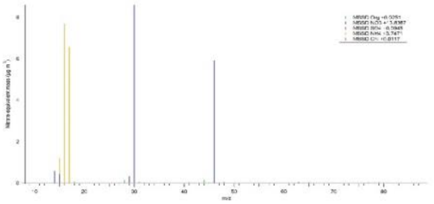
S.5 Approximation for error propagation for signal ratios

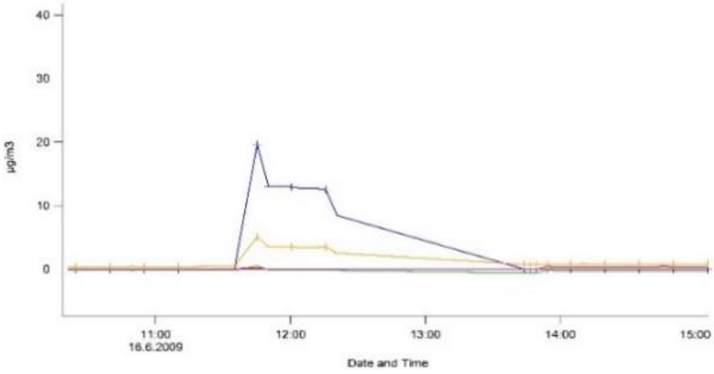
Propagation of uncertainty can be approximated by the variance formula (Ku, 1966), which yields for a ratio of two signals (s_1 and s_2) and their uncertainties (σ_1 and σ_2) an uncertainty

$$\sigma(s_1 : s_2) = \sqrt{\frac{1}{s_2^2} * \sigma_1^2 + \frac{s_1^2}{s_2^4} * \sigma_2^2}\,. \qquad \text{(S.5)}$$

5 In this work the above is applied in estimating the uncertainties of various signal ratios, such as σ (NO_2^+:NO^+).

S.6 Ammonium nitrate and ammonium sulphate calibrations and the CO_2^+ artefact





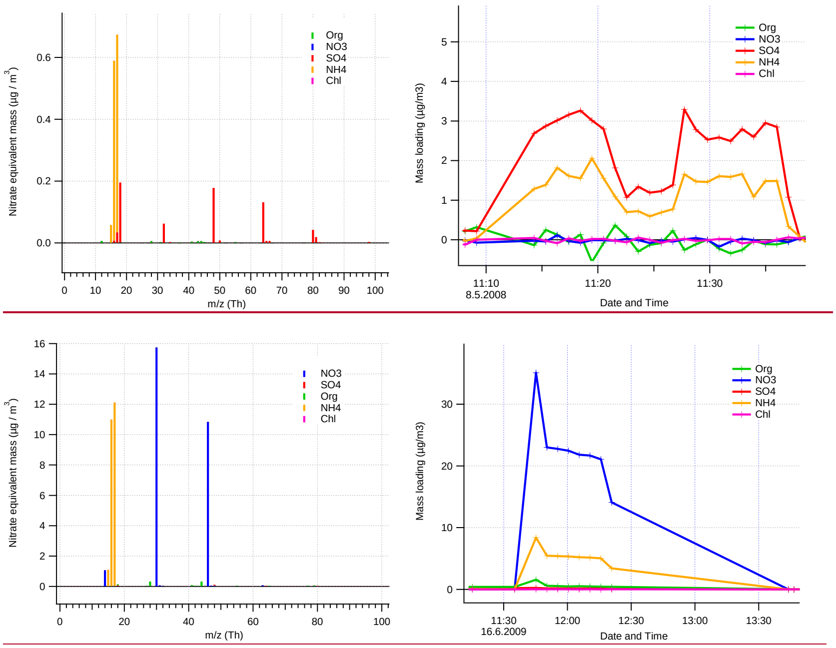
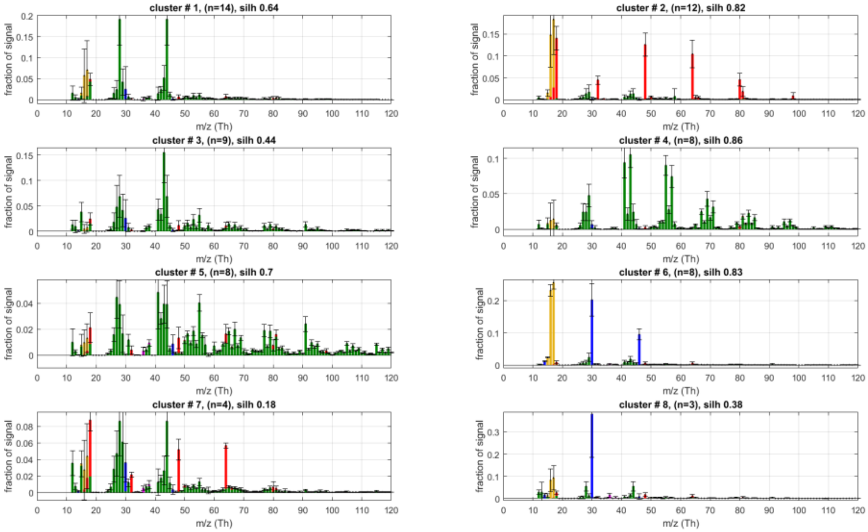


Figure S.3. AS and AN calibrations in 'MS' mode~~ (unprocessed data).~~. Left: mass spectra, right: time series of a calibration. The data is not quality assured and was not used for (R)IE analysis.

Pieber et al (2016) discuss the CO_2^+ artefact induced by thermal decomposition of inorganic salts. They find a *m/z* 44 Th artefact with a magnitude of 3.4 % of total nitrate signal. For the phase III solution we obtain a corresponding value (for signal ratio of *m/z* 44 Th to NO_3) of 2.9 %, suggesting most of the *m/z* 44 Th signal may indeed be explained by this artefact. When accounting for fragmentation-table-calculated *m/z* 44 Th derivatives, the same applies most of total organic signal for AN. For AS (P-III), we observe a similar ratio (*m/z* 44 Thto SO_4) of 2.8 %, when Pieber (2016) estimate a contribution between 0.1 to 0.3 % for AS. Similar *m/z* 44 Th artefact is also seen for the AN calibration data (Figure S.3) above. For ON, the observed signal ratio of *m/z* 44 Th to NO_3 (P-III) is 22 %. Notably, the ON solution for P-I differs from the P-III solution here, with a ratio of 8.8 %. For KNO_3 Pieber et al (2016) report a ratio of 4.5 %. This also gives rise to the speculation discussed in Section 3.2.2, that the ON factor might arise from thermal decomposition of KNO_3. Furthermore, Pieber et al (2016) importantly note that thermal decomposition of KNO_3 is temperature dependent, so vaporiser temperatures above 600 degrees Celsius would increase the observed ratio. While the standard operating temperature of the vaporiser (as per instrument readout) was used by default, oven temperature calibrations were not performed. Thus we do not have vaporiser power data available for the various campaigns to further examine the possibility of a bias in oven temperature readout, and if such a bias could connect to our observations of ON.

S.7 Phase II clustering solution

Phase II solution was obtained for $k = 15$, $s_m = 2.41$ (silh 0.6557), and is presented below.



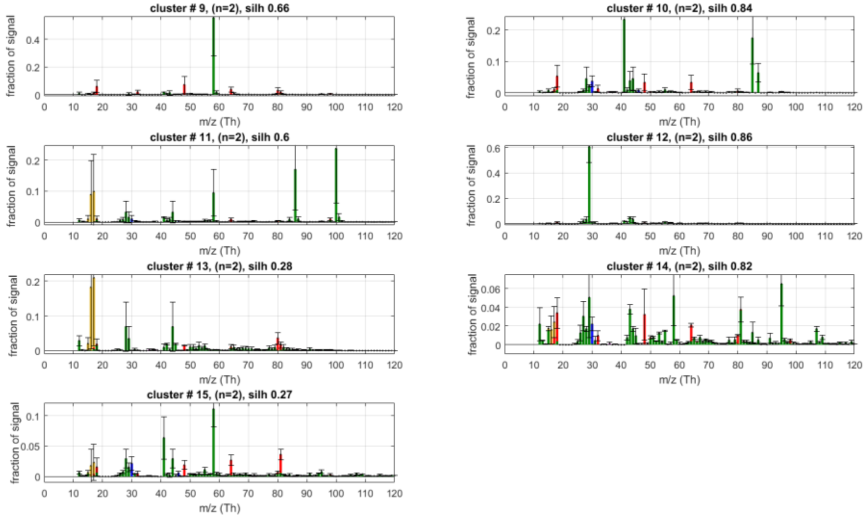
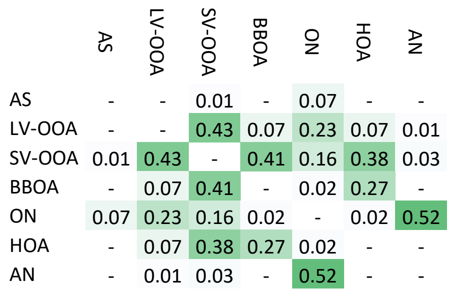


Figure S.4. P-II cluster centroids and variabilities are silhouette-weighted averages and standard deviations for the cluster members. The main SOA types were identified as cluster #1 ('low-volatile oxidised organic aerosol', LV-OOA) clusters #3 ('semi-volatile oxidised organic aerosol', SV-OOA) and cluster #7 ('organic nitrogen', ON).

S.8 CMB component inter-correlations

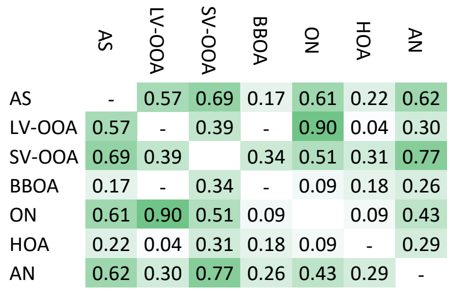
We calculated the mass-scaled r-CMB component profile inter-correlations to evaluate r-CMB component similarities. The result is shown in Tables S.2 (full spectrum) and S.3 (organics only). Since the scaled correlation similarity (r_s^2) also corresponds to the quantity that the clustering algorithm aims to minimise, high values for r_s are not to be expected of a robust (hard) classification solution. From the resulting similarities, it seems the distinction between AN and ON seems to be the hardest for the algorithm (with the selected metric). Also, SV-OOA shares many similar features with the LV-OOA, BBOA and HOA.

Table S.2. Squared correlation coefficients (r_s^2; $s_m = 1.81$) for CMB component spectra inter-correlations, with both organics and inorganics included. Only correlations with p-values < 0.05 are shown.



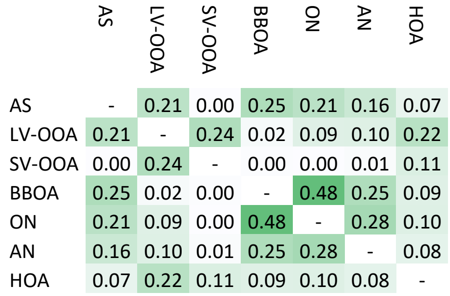
	AS	LV-OOA	SV-OOA	BBOA	ON	HOA	AN
AS	-	-	0.01	-	0.07	-	-
LV-OOA	-	-	0.43	0.07	0.23	0.07	0.01
SV-OOA	0.01	0.43	-	0.41	0.16	0.38	0.03
BBOA	-	0.07	0.41	-	0.02	0.27	-
ON	0.07	0.23	0.16	0.02	-	0.02	0.52
HOA	-	0.07	0.38	0.27	0.02	-	-
AN	-	0.01	0.03	-	0.52	-	-

Table S.3. Squared correlation coefficients (r_s^2; $s_m = 1.81$) for CMB component spectra inter-correlations, for organics only. Only correlations with p-values < 0.05 are shown.



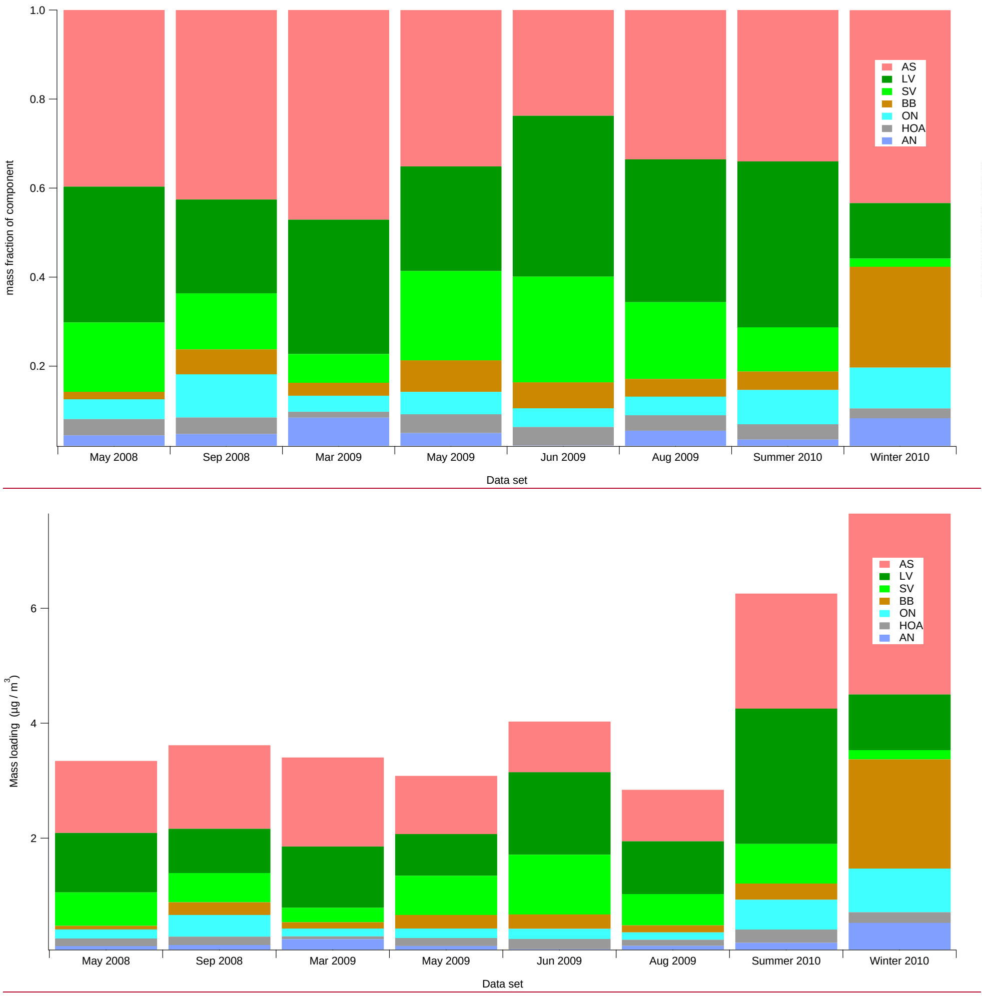
	AS	LV-OOA	SV-OOA	BBOA	ON	HOA	AN
AS	-	0.57	0.69	0.17	0.61	0.22	0.62
LV-OOA	0.57	-	0.39	-	0.90	0.04	0.30
SV-OOA	0.69	0.39		0.34	0.51	0.31	0.77
BBOA	0.17	-	0.34	-	0.09	0.18	0.26
ON	0.61	0.90	0.51	0.09		0.09	0.43
HOA	0.22	0.04	0.31	0.18	0.09	-	0.29
AN	0.62	0.30	0.77	0.26	0.43	0.29	-

Table S.3. Pearson correlation coefficients for CMB component time series inter-correlations. Only correlations with p-values < 0.05 are shown.



	AS	LV-OOA	SV-OOA	BBOA	ON	AN	HOA
AS	-	0.21	0.00	0.25	0.21	0.16	0.07
LV-OOA	0.21	-	0.24	0.02	0.09	0.10	0.22
SV-OOA	0.00	0.24	-	0.00	0.00	0.01	0.11
BBOA	0.25	0.02	0.00	-	0.48	0.25	0.09
ON	0.21	0.09	0.00	0.48	-	0.28	0.10
AN	0.16	0.10	0.01	0.25	0.28	-	0.08
HOA	0.07	0.22	0.11	0.09	0.10	0.08	-

S.9 Mass attribution by campaign



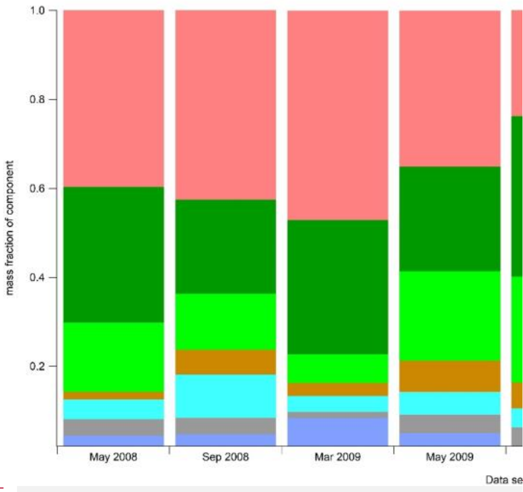


Figure S.5. ~~M~~Mass fractions (upper panel) and absolute mass concentrations (lower panel) ~~-~~of r-CMB component by data set. Data set VIII (Winter 2010) results are less reliable due to pronounced surface ionisation effects.

As discussed in 3.2.1, we consider results for data set VIII less reliable due to effects very likely stemming from pronounced surface ionisation processes. The high uncertainty is also reflected in phase I (PMF) high object function values ($Q/Q_{expected}$; see Paatero et al., 2014). The (P-I) Q/Q_{exp} diagnostic value, optimally approaching unity, was 15.29 for data set VIII, while for the other 7 data sets the mean ± standard deviation of Q/Q_{exp} was 1.32 ± 0.53, indicating the PMF model performance was good for data sets I to VII, but poor for set VIII. While a high object function value alone does not disqualify a result, it does indicate the uncertainties in data were likely underestimated (e.g. Norris et al., 2008), and the model does not capture the uncertainty-weighted data variation properly, even with 10 unconstrained factors.

S.10 Cluster projections for Phase I (unconstrained PMF) result

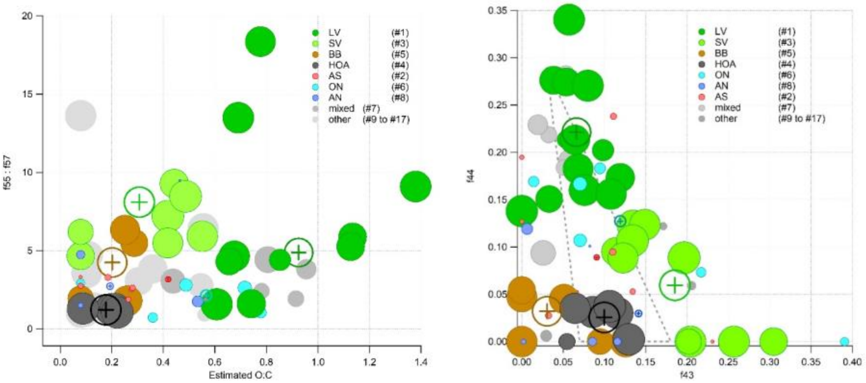


Figure S.6. (left panel) P-I solution - cluster projections onto a *f55/f57* (Mohr et al., 2012), *O:C* (estimated, Aiken et al., 2008) plane. Circles correspond to the members of the cluster and the cross markers to cluster centroids. Marker size indicates organic mass fraction in spectra. Axes are truncated. (right panel) P-I solution, projected onto *f44, f43* plane (i.e. the 'Sally's triangle' plot; Ng et al, 2010). Circles correspond to objects in cluster and the cross markers to cluster centroids. Marker size indicates organic mass fraction in spectra. Dotted line marks the area where most laboratory data for organic aerosol falls (Ng et al., 2010).

S.11 Potassium signal and surface ionisation effects

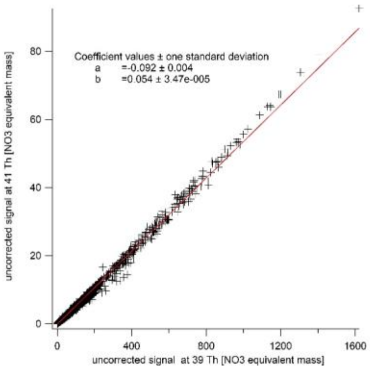


Figure S.7. X-axis: measured signal at m/z 39 Th (mainly K-39). Y-axis: measured signal at m/z 41 Th. Data is for Winter 2010 campaign (data set VIII), and contains the highest K-39 signals observed. Although corrected for in fragmentation table, the K-41 isotope correction does not fully negate the effect of potassium signal on m/z 41, which may partly explain the high m/z 41 Th signal observed affecting e.g. the BBOA component in data set VIII.

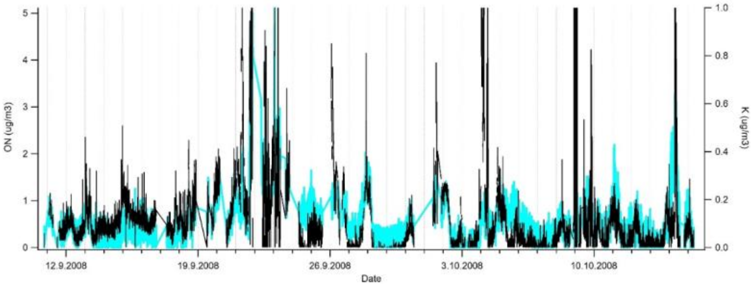


Figure S.8. On left axis: orgNO3 mass estimate from the Kiendler-Scharr parameterisation (Eq. 7) in turquoise. On the right axis: total Potassium signal in black. Although the similar temporal behaviour is at least to some extent explained by total aerosol loading, it does raise the question if surface ionisation effects could contribute to the ON component, e.g. in form of KNO_3 signal.

S.12 WD and mass spectral analysis for Rb-containing outlier factor

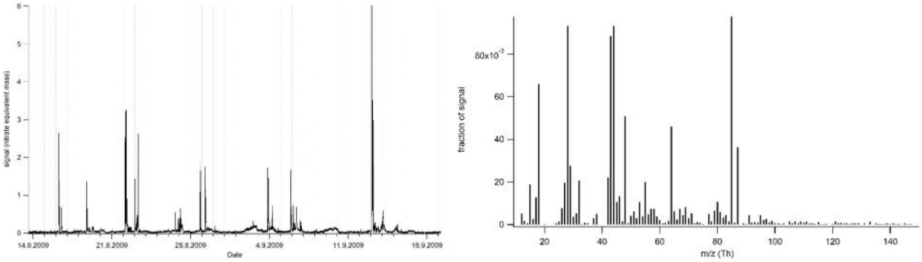


Figure S.9. Time series (left; P-I) and mass spectrum (right, P-I)) of a Rb-containing factor from data set VI.

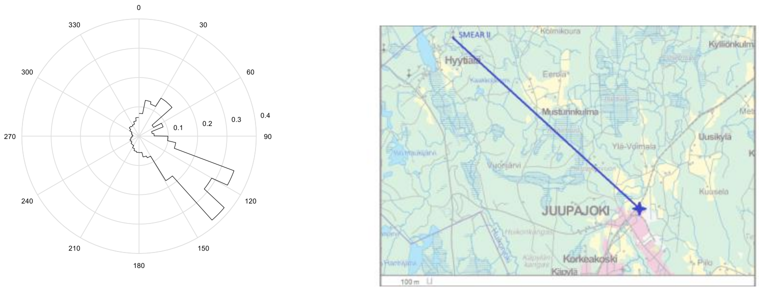


Figure S.10. Left: wind rose for the Rb containing factor from data set VI (P-I) average signal for the Rb-containg factor for wind sectors. Right: map showing the locations for SMEAR II and the Juupajoki district heating plant, co-located with a sawmill.

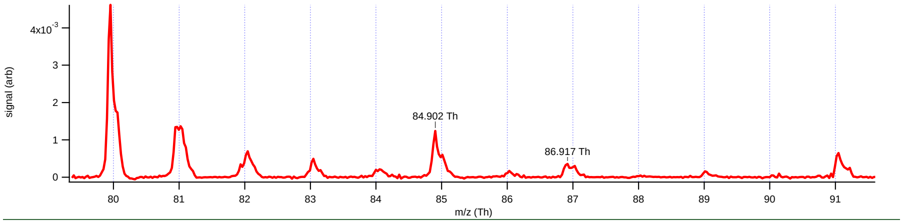


Figure S.11. Mass calibrated raw mass spectral signal for a plume event in Aug 14th 2010 (data set VII). While the mass resolution of the instrument (~500 *m/dm*) is not enough for proper high-resolution analysis, we can confirm existence of ions with clearly negative mass defect at m/z 85 and 87 Th. The highest m/z bins (tagge~~d~~s in figure) agree with exact masse~~s~~es for ^{85}Rb^{+} (84.912 a.m.u.) and ~~^{85}~~^{7}Rb^{+} (86.909 a.m.u.) ions~~.~~.

S.13 Diurnal cycles of r-CMB components

Daily cycles of r-CMB components are presented in Figure S.12. Due to aerosol generally being transported over long distances to the site, clear source-related peaks of e.g. rush hours for HOA are not resolved. Besides temporal behaviour of source emissions, the observed daily cycles can be connected to aerosol temperature response or boundary layer dynamics.

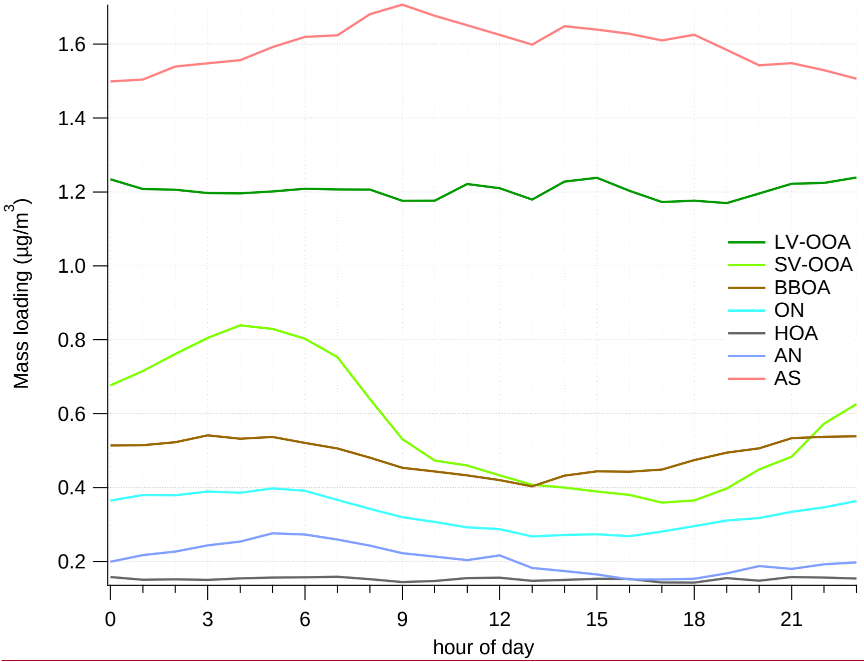


Figure S.12. Average loading of r-CMB model components by hour of day, over all of the data sets. Since the data spans all seasons, and the length of day varies considerably, the variability of average loading is generally high.

S.14. Estimated variability for the final r–CMB components

Formatted: Heading 1

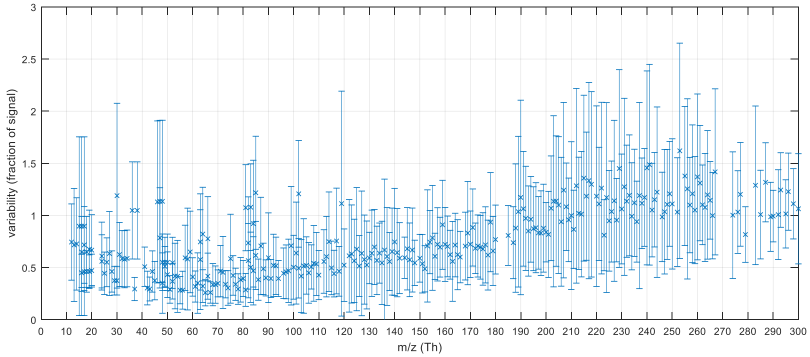


Figure S.13. Variability estimate as a function of m/z ratio for the final r-CMB model components. Mean (x) and standard deviation (error bar) of the variability-to-signal ratio, calculated over all the 7 P-III clusters.

Formatted: Normal

Jayne J.: AMS best practices, AMS user's meeting presentation, 2018. Available online at http://cires1.colorado.edu/jimenez-group/UsrMtgs/UsersMtg19/Jayne_CalibrationOverview.pdf, viewed 19.Sep.2018.

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, M. R., Prevot, A. S., 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, Environ Sci Technol, 42, 4478-4485, 2008.

Allan, J. D., Jimenez, J. L., Williams, P. I., Alfarra, M. R., Bower, K. N., Jayne, J. T., Coe, H., and Worsnop, D. R.: Quantitative sampling using an Aerodyne aerosol mass spectrometer 1. Techniques of data interpretation and error analysis, Journal of Geophysical Research: Atmospheres (1984–2012), 108, 2003.

Buajarern, J., Mitchem, L., and Reid, J. P.: Characterizing the formation of organic layers on the surface of inorganic/aqueous aerosols by Raman spectroscopy, The Journal of Physical Chemistry A, 111, 11852-11859, 2007.

Drewnick, F., Diesch, J.-M., Faber, P., and Borrmann, S.: Aerosol mass spectrometry: particle–vaporizer interactions and their consequences for the measurements, Atmospheric Measurement Techniques, 8, 3811-3830, 2015.

Friedlander, S. K.: Chemical element balances and identification of air pollution sources, Environmental Science & Technology, 7, 235-240, 1973.

Gordon, G. E.: Receptor models, Environmental Science & Technology, 22, 1132-1142, 1988.

Gysel, M., Crosier, J., Topping, D., Whitehead, J., Bower, K., Cubison, M., Williams, P., Flynn, M., McFiggans, G., and Coe, H.: Closure study between chemical composition and hygroscopic growth of aerosol particles during TORCH2, Atmospheric Chemistry and Physics, 7, 6131-6144, 2007.

Hong, J., Äijälä, M., Häme, S. A., Hao, L., Duplissy, J., Heikkinen, L. M., Nie, W., Mikkilä, J., Kulmala, M., and Prisle, N. L.: Estimates of the organic aerosol volatility in a boreal forest using two independent methods, Atmospheric Chemistry and Physics, 17, 4387-4399, 2017.

Hopke, P. K.: Receptor modeling for air quality management, Elsevier, 1991a.

Hopke, P. K.: An introduction to receptor modeling, Chemometrics and Intelligent Laboratory Systems, 10, 21-43, 1991b.

Jimenez, J. L., Jayne, J. T., Shi, Q., Kolb, C. E., Worsnop, D. R., Yourshaw, I., Seinfeld, J. H., Flagan, R. C., Zhang, X. F., Smith, K. A., Morris, J. W., and Davidovits, P.: Ambient aerosol sampling using the Aerodyne Aerosol Mass Spectrometer, Journal of Geophysical Research-Atmospheres, 108, Artn 8425

10.1029/2001jd001213, 2003.

Kiendler-Scharr, A., Mensah, A. A., Friese, E., Topping, D., Nemitz, E., Prevot, A., Äijälä, M., Allan, J., Canonaco, F., and Canagaratna, M.: Ubiquity of organic nitrates from nighttime chemistry in the European submicron aerosol, Geophysical Research Letters, 43, 7735-7744, 2016.

Ku, H. H.: Notes on the use of propagation of error formulas, Journal of Research of the National Bureau of Standards, 70, 1966.

Miller, M., Friedlander, S., and Hidy, G.: A chemical element balance for the Pasadena aerosol, in: Aerosols and atmospheric chemistry, Elsevier, 301-312, 1972.

Mohr, C., DeCarlo, P. F., Heringa, M. F., Chirico, R., Slowik, J. G., Richter, R., Reche, C., Alastuey, A., Querol, X., Seco, R., Penuelas, J., Jimenez, J. L., Crippa, M., Zimmermann, R., Baltensperger, U., and Prevot, A. S. H.: Identification and quantification of organic aerosol from cooking and other sources in Barcelona using aerosol mass spectrometer data, Atmospheric Chemistry and Physics, 12, 1649-1665, 10.5194/acp-12-1649-2012, 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., Prevot, 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.

Norris, G., Vedantham, R., Wade, K., Brown, S., Prouty, J., and Foley, C.: EPA positive matrix factorization (PMF) 3.0 fundamentals & user guide, Prepared for the US Environmental Protection Agency, Washington, DC, by the National Exposure Research Laboratory, Research Triangle Park, 2008.

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.

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., and Vlachou, A.: Inorganic salt interference on CO2+ in aerodyne AMS and ACSM organic aerosol composition studies, Environmental science & technology, 50, 10494-10503, 2016.

Sun, Y., Zhang, Q., Schwab, J., Yang, T., Ng, N., and Demerjian, K.: Factor analysis of combined organic and inorganic aerosol mass spectra from high resolution aerosol mass spectrometer measurements, Atmospheric Chemistry and Physics, 12, 8537-8551, 2012.