Methods to extract molecular and bulk chemical information from series of complex mass spectra with limited mass resolution

Harald Stark,^{1,2,3*} Reddy L. N. Yatavelli,^{1,3,§} Samantha L. Thompson,^{1,3} Joel R. Kimmel,^{2,4} Michael J. Cubison,⁴ Puneet S. Chhabra,^{2,†} Manjula R. Canagaratna,² John T. Jayne,² Douglas R. Worsnop,^{2,5} Jose L. Jimenez^{1,3*}

[1] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, 80309 USA

[2] Aerodyne Research Inc., Billerica, MA, 01821 USA

[3] Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO, 80309 USA

[4] TOFWERK AG., 3600 Thun, Switzerland

[5] Department of Physics, University of Helsinki, Helsinki 00014, Finland

§Now at California Air Resources Board, El Monte, CA 91731 USA

†Now at PerkinElmer Inc., Hopkinton, MA 01748 USA

* To whom correspondence should be addressed: harald.stark@aerodyne.com, jose.jimenez@colorado.edu

Graphical Abstract

Keywords

Chemical ionization; high-resolution time-of-flight mass spectrometry; atmospheric measurements; analysis of mass spectra;

Abstract

The resolution of mass spectrometers is often insufficient to conclusively identify all peaks that may be present in recorded spectra. Here, we present new methods to extract consistent molecular and bulk level chemical information by constrained fitting of series of complex organic mass spectra with multiple overlapping peaks. Possible individual peaks in a group of overlapping peaks are identified by both defining a chemical space and by free peak fitting. If simply all possible formulas from the chemical space would be used to fit each peak, the result would not be well constrained. The free peak fitting algorithm provides information about likely peak locations. A new algorithm then reconciles the results of both methods and produces a final peak list for use in subsequent fitting, while using all available experimental constraints. Comparison to ultra-high resolution data suggests that the real peak density is substantially higher than can be resolved with the instrument resolution. Bulk chemical properties such as carbon number (nC) and carbon oxidation state (OS_c) can be calculated from the fit results. For mixtures of compounds dominated by C, H, O and N, bulk properties can be reliably extracted, even though some formula assignments may remain uncertain. This ability to retrieve correct bulk parameters even if not all assigned formulas are correct originates from the relationship between mass defects of individual peaks and the chemical parameters under our CHON composition assumptions. Retrieving consistent bulk parameters across series of many mass spectra is essential for extracting time trends, e.g. for field measurements taking place over several weeks. We illustrate the fitting method using a sample data set from a chemical ionization mass spectrometer with a resolution of approximately 4000 (M/dM), operated using acetate reagent ions. Spectral simulation experiments validate the analysis method by showing good agreement of intensities for many specific ions, as well as for bulk chemical parameters. An alternative method to directly extract bulk chemical information from the raw spectra without the need of any peak assignment or peak fitting is also introduced, which shows good agreement with the peak fitting results. The latter method can be applied very rapidly without the need for complex analysis procedures, e.g. as a quick online diagnostic during data acquisition.

For submission to International Journal of Mass Spectrometry
(<http://www.sciencedirect.com/science/journal/13873806>)

53 1. Introduction

54 Mass spectrometry is a commonly used tool in many research fields. When analyzing complex samples
55 without chromatographic separation, such as organic samples from atmospheric aerosols, mass spectra
56 can contain molecular ions from hundreds to thousands of different organic compounds [1–3]. This is
57 especially true for a soft-ionization technique such as chemical ionization [4,5]. If ion peaks are not fully
58 resolved from each other, as is inevitably the case at higher mass-to-charge ratios (m/z) [6,7], a spectral
59 fitting method is needed. Such a method will help to extract chemical information from the recorded
60 spectra for either molecular or “bulk” properties averaged across all formulas. The term “molecular” is
61 used to represent all possible isomers of a specific sum formula, since mass spectrometry can only
62 provide information about individual exact m/z values, but not about functional groups or molecular
63 structure. Bulk properties, such as carbon oxidation state (OS_c) and number of carbon atoms (nC), are
64 useful indicators of atmospheric processing (i.e. oxidation and/or chemical fragmentation) of organic
65 compounds [6,8]. In addition to retrieving this information from individual spectra, it is often as
66 important to get consistent data from multiple mass spectra, as is the case, for example, for field
67 campaigns, where regularly tens of thousands of mass spectra are acquired over weeks and time series
68 of both molecular and bulk parameters are needed.

69 An example for the large amount of possible chemical formulas present at higher m/z is shown in figure
70 1, which will be discussed in more detail later. The left column shows individual peaks from possible
71 chemical formulas containing C,H,O and N for ions at three different m/z ranges. The number of possible
72 ions increases rapidly with m/z and all possible ions are not resolvable at $m/z > 200$ (bottom row).
73 Assumptions or additional information about the composition of the sample (e.g. elements present) are
74 often useful to limit the number of possible formulas that could be assigned to the ions from the
75 sample. Kind and Fiehn [9] developed a set of rules to determine the most likely formula at a given m/z .
76 However, even such constraints could still lead to ambiguous peak assignment possibilities, for example
77 when overlapping ion signals at even higher m/z are encountered.

78 Over the course of more than 30 years, the analysis of mass spectra has evolved from simple ion peak
79 position algorithms to complex spectral fitting routines. The importance of multiple peaks at identical
80 unit masses, resulting in complex peak shapes was addressed as long ago as 1981 by van Katwijk, using
81 mixtures of organic compounds from petroleum fractions [10], where multi-peak signals were due to
82 isotopes and resolved by successively removing isotopes from ions at lower m/z . Feng et al. [11] used a
83 Gaussian peak shape to determine the center of a peak in a mass spectrum recorded by a quadrupole
84 mass spectrometer. A mass accuracy of 12 ppm was achieved. The situation of overlapping peaks was
85 discussed as a problem for this type of peak fitting, but multi-peak fitting was not used to resolve
86 multiple peaks. The asymmetry of peaks due to tuning issues of the mass spectrometer was addressed
87 by Strittmatter et al.[12], where a multi-variate regression method and a double-Gaussian peak shape
88 were used. The accuracy of 8 ppm from the peak fitting was a significant improvement over the 29 ppm
89 of previously used centroid methods. The concept of baselines as an important component of mass

spectra was discussed by Horn et al [13], which determined baselines from histograms of the ion signals. House et al. [14] applied multi-peak fitting to overlapping features from a Matrix Assisted Laser Desorption Ionization Time-of-Flight (MALDI-TOF) mass spectrometer. That study showed the importance of using an instrument-specific mass resolution function to distinguish single wider peaks from multiple narrower peaks.

A complete procedure to fit multi-peak signals arising from the Aerodyne time-of-flight aerosol mass spectrometer (AMS) was described by DeCarlo et al [15]. The algorithms developed in this study originate from that work. DeCarlo et al [15] also used previously determined peak resolutions and shapes to constrain the multi-peak fitting method. A more recent study by Graus et al [16] described how to identify multi-peak signals from proton-transfer reaction mass spectrometers (PTRMS) based on peak width considerations. They introduced an algorithm for assigning chemical formulas to individual high-resolution peaks based on a list of candidate formulas, constrained by bond valence rules. Subsequent studies [5,17,18] expanded on Graus et al. [15] and developed more refined methods and advanced algorithms for the PTRMS.

Here, a new analysis approach is described which can be used to assign the most likely ion formulas when the resolution of the mass spectrometer is not sufficient to unambiguously assign specific compound formulas. To our knowledge, there are no prior reports of techniques with similar goals in the literature. In this work, the mass spectrometer system was operated with a resolution of approximately 4000, which allows for some degree of separation of the many organic ion fragments but often not enough for unambiguous identification. The molecular and bulk level chemical information obtained with this analysis is demonstrated and evaluated with both field and simulated high-resolution spectra. A new method in which the raw mass spectrum is used to directly determine bulk chemical properties is also introduced. This method calculates bulk properties merely from signal intensity and mass defect without any peak assignment or fitting.

2. Methods

2.1. Experimental methods

All experiments used a high-resolution time-of-flight chemical ionization mass spectrometer, coupled to a micro-orifice volatility impactor inlet (MOVI-HRToF-CIMS, Aerodyne Research, Billerica, MA) [19,20]. The experimental data analyzed here were acquired during the BEACHON-ROMBAS field study in 2011 to provide organic composition information for gas and particle phase constituents. The measurement location was a pine forested area in the Colorado Rocky Mountains [21]. Details of the experimental setup are described elsewhere [22]. Briefly, ambient air was sampled using a custom inlet and aerosol particles were collected by impaction on a post while chemically analyzing the gas phase (henceforth referred to as “gas-phase cycle”). Then the particle composition was measured as the post was slowly heated leading to thermal desorption of the collected particles during the “particle cycle”. Acetate (CH_3COO^-) was used as the reagent ion, which was produced by flowing acetic anhydride through a ^{210}Po ionizer [23] to selectively detect and quantify organic acids. As tuned, the mass spectrometer had a

maximum resolution of 4000 and operated under “high declustering” conditions to avoid measuring ambient compounds clustered with reagent ions [24].

The methods developed here were applied to 15-second average spectra (from 1-second raw data) generated from the BEACHON-RoMBAS dataset. (20-30 August 2011). A “final peak list” was generated using highly averaged spectra (typically, 24-hour averages, more than 80,000 mass spectra) to improve signal-to-noise (S/N) ratio and because some ions may not be present during all times of the day. It contained all possible peaks encountered from both gas- and particle-phase cycles. This list was eventually used with a strongly-constrained fit algorithm to determine signal intensities for all measured raw spectra. The different components of the analysis steps are described below in detail. An overview of the data analysis flow is shown in figure 2.

2.2. Structure of the multi-peak fitting methods

All data analysis presented here was performed within the data analysis software *Igor Pro* (versions 6.2-6.3, Wavemetrics, Lake Oswego, OR, USA). The customized analysis routines are distributed by Tofwerk (Thun, Switzerland) and Aerodyne Research Inc. (Billerica, MA, USA) under the name “Tofware” (<https://sites.google.com/site/citofms/analysis-software>). The routines are designed for compatibility with Tofwerk TOF data files (HDF 5.0).

The method developed here is based on constrained fitting of all ion signals in a mass spectrum. The signal as a function of m/z of a group of multiple overlapping ion peaks (“multi-peak”) in a mass spectrum can be described mathematically as a linear combination of individual peaks of the same shape but different locations and amplitudes:

$$I(m/z) = \sum_{i=1}^n A_i \cdot S_i(m/z) \quad \text{Equation 1}$$

With: n : number of individual ions in multi-peak

$I(m/z)$: Signal at individual m/z

A_i : amplitude of peak i

$S_i(m/z)$: peak shape function, centered at peak i , normalized to unity peak height

While this equation describes how the multi-peak is composed of the individual peaks, the same equation is used as a fit function to find individual peak positions in a multi-peak. It should be noted that the actual fits were always performed in time-of-flight (TOF) space and the results converted to m/z space. The reason for this is that the peak width and shape are properties of the measurement distribution in TOF space and translating them to m/z would be very difficult owing to the non-linear conversion between TOF and m/z . Examples of measured multi-peaks and fits to them following the above equation are shown in Figure 1. If one tried to determine the number of peaks, peak locations, amplitudes, widths, shapes, and the m/z scale from each fit, the fits would be very under-determined and the results highly uncertain. Constraining ahead of time all parameters of the fit that can be determined a priori reduces the uncertainty of the fit results and allows extraction of the most

information from mass spectra. This includes determining the m/z calibration, peak width and shape, and the number and location of peaks that are part of a multi-peak group.

2.3. Mass-to-charge ratio (m/z) calibration

A high accuracy m/z calibration is needed to have confidence in peak assignments, especially when multiple overlapping peaks are present. A post-measurement m/z calibration procedure was applied to all spectra, using signals from ion peaks deemed to be isolated. For the dataset used in this study, seven isolated ion peaks were identified as suitable for m/z calibration (listed in table 1) based on sufficient intensity across all measured spectra and lack of substantial interfering peaks. The assigned chemical formulas for the two compounds at high m/z (>300) were determined by their isotopic ratios (mainly ^{13}C : ^{12}C and ^{37}Cl : ^{35}Cl) together with their approximate mass defects. 15-sec data were averaged to 1-minute to increase signal-to-noise for all m/z calibrations. Low signal-to-noise can degrade the quality of the m/z calibration and the peak fits [25]. The instrument m/z calibration typically varies over much slower time scales. The assigned time of flight for each m/z -calibration ion was determined by fitting a Gaussian function to the measured data points. Then, a power law equation was fitted to all m/z -TOF data pairs:

$$\text{TOF} = p_1 (m/z)^{p_2} + p_3 \quad \text{Equation 2}$$

The exponent p_2 may deviate slightly from the theoretical value of 0.5 due to non-ideal behavior in the extraction process [15]. Table 1 also shows accuracies of the calibration m/z values with regard to the m/z calibration function from a 5-day period on 25-30 August 2011. Accuracies are reported as parts-per-million of the m/z value (ppm), calculated as:

$$\text{Accuracy [ppm]} = (m/z_{\text{comp}} - m/z_{\text{cal}}) / (m/z_{\text{cal}}) \times 10^6 \quad \text{Equation 3}$$

With: m/z_{comp} : peak position

m/z_{cal} : calibration value (from equation 2)

The average accuracy achieved for all seven m/z -calibration peaks is 3.8 ppm, determined from a weighted average of all m/z calibration accuracies and their standard deviations in table 1. The ions used in the calibration will naturally have the lowest errors, thus the accuracy for ions not used in the m/z calibration must also be evaluated. Importantly, several overlapping peaks are recorded at most integer m/z , and multi-peak fitting has to be conducted to determine their individual signal intensities. This procedure may introduce additional uncertainty in the determined peak positions beyond the m/z accuracy of isolated peaks. In order to estimate the accuracy for typical measured ions, multi-peak fits of peaks with known individual peak positions were performed while allowing the m/z calibration to shift as an additional fit parameter. The m/z shift value that resulted in the lowest squares of residuals was considered the best estimate of the m/z scale for that multi-peak group. Repeating this process for a long time series of acquired spectra allows quantification of the average bias and m/z accuracy from the average and standard deviation, respectively. Table 2 shows the average biases and m/z accuracies for 4 ions, determined from a time series of several hours of individual spectra. The table lists values for both

isolated peaks and for those in multiple-peak groups. The average accuracies estimated for the isolated peaks are similar to those for the m/z calibration peaks, but an order-of-magnitude larger for the multiple peak situation, reflecting the additional uncertainties due to the proximity of other peaks [25]. This is the critical reason why we perform the m/z calibration only on individual peaks and then use it as the m/z base for all peak fits, so that uncertainties arising from multi-peak group fitting does not deteriorate the m/z calibration.

2.4. Instrument functions

In order to better constrain the multi-peak fitting, it is advantageous to first determine three additional instrument functions: the peak width (resolution) as a function of m/z , the isolated peak shape, and the ion signal baseline. These parameters allow the precise definition of an isolated peak at each m/z . This is a required input of our constrained multi-peak fitting routine and results in a much better estimate of the ion peak position and intensity than if one attempted to determine all parameters from each multi-peak fit. Instrument functions, calculated every two hours over a two day period, were found to be stable. Therefore, daily averaged instrument functions were used for the entire campaign.

2.4.1. Resolution

The resolution of a peak is a unitless value and is defined as the ratio of its m/z to its full width at half maximum (FWHM):

$$R = (m/z) / (m/z)_{FWHM} \quad \text{Equation 4}$$

Since many peaks in a mass spectrum do not arise from individual ion signals, especially for the 'CHON' rich compounds measured in this field study, an algorithm has been developed to identify peaks that appear to be isolated and determine the peak width as a function of m/z . The algorithm uses the observed linear relationship for this instrument between peak width and m/z with R as the slope. It fits a line of all measured widths versus m/z and iteratively removes the widest peak (relative to the linear fit) from the spectrum and recalculates the linear fit until the number of peaks below the line (narrow peaks) are below a preselected value. This results in a selection of individual peaks across a wide m/z range. We note that this method results in a less stringent criterion for isolated peaks than needed for the m/z calibration, which is acceptable because R is derived from the linear fit and outliers are mostly ignored. Due to the automated peak selection process, different isolated ions were used for different days. Typically, 16 peaks were used and average resolution was 4000 above m/z 200 (see inset in panel A of figure S1).

2.4.2. Peak shape

Ideally, a Gaussian peak shape would be observed in a mass spectrometer. However, many nonlinear effects, such as imperfections in the initial distribution of ion positions and energies, a finite time needed to raise the pulsed extraction voltages, effects of temperature on instrument geometry and electronics, changes in tuning due to user input and formation of insulating layers (organic films) on instrument electrodes can lead to asymmetric peak shapes [15]. Therefore, the actual peak shape of the instrument for a given dataset needed to be determined to better constrain multi-peak fitting. An

average peak shape for the entire m/z range was determined for each 24-hour period by manually selecting individual peaks, normalized to their maximum height and FWHM. Figure S1B shows an example peak shape. Note that the non-Gaussian left-hand tail of the peak shape is needed to account properly for ions at the 3% level on that side of large peaks.

2.4.3. Baseline

Baselines under peaks, e.g. from electronic noise, have to be subtracted before any peak fitting can be conducted. Here, baseline regions were defined to be centered at the half point between integer m/z , with decreasing widths at increasing m/z , to account for an increasing ion variety resulting in wider mass defect spreads. These baseline regions were averaged and then interpolated across the peak integration regions to calculate a baseline for the entire m/z range (fig S1C).

2.5. Formula generator

A custom algorithm was used to generate a list of all possible chemical formulas, tailored specifically for the acetate reagent ion chemistry. The formula list was then used to assign formulas to exact m/z peak positions from measured spectra. The algorithm starts with a “seed” formula and uses three progression and three termination rules. The progression rules were: (a) addition of an O atom; (b) addition of a methylene group (CH_2); and (c) subtraction of 2 hydrogens (H_2). Progression rule (c) can be seen as generating an unsaturation ($\text{C}=\text{C}$ or $\text{C}=\text{O}$ double bond) or ring in the compound, while the other rules don’t change the number of rings plus double bonds (RPDB). Therefore, the algorithm starts with the most saturated compounds and generates formulas with the same or increasing RPDB. A formula generation branch was terminated if one of the terminating rules was reached or exceeded: (I) H:C below threshold; (II) O:C above threshold; and (III) maximum m/z or maximum number of C atoms. Although the algorithm generates chemically possible bound organic formulas, it is not possible to determine if the compounds corresponding to these formulas are chemically stable or not.

2.6. Determining number, position and height of peaks

The final critical component that allows one to later perform accurate constrained peak fitting is the precise determination of the number, positions and ion peak heights in the multi-peak spectra. Two peak fitting algorithms were used, one that allows a free fit to determine number and position of peaks and one algorithm that determines only the peak heights while number and position of peaks are constrained. An iterative algorithm is used to assign peaks determined by the first algorithm as input to the second algorithm.

2.6.1. Free peak list algorithm

The “free peak list” algorithm determines both amplitudes and positions of a variable number of peaks. It uses an iterative Levenberg-Marquardt routine built into Igor that minimizes the squared residuals of the fit, constrained by the peak shapes and widths determined from the instrument functions. The resulting list of peak positions (m/z ’s) from this algorithm is named “free peak list” or “free list”. Note that some peak positions determined by this algorithm may not correspond to possible chemical formulas within the accuracy of the mass spectrometer m/z calibration. Also, due to the complex iterative fitting, singularities can be encountered frequently and computation times can be long, about 200 times longer than with the fixed peak algorithm. This would mean that fitting of the entire dataset,

which took about 12 hours with the fixed peak algorithm, would take about 3 months using the free peak list algorithm. Thus the free peak algorithm was only used for helping constrain the peak positions, and not on the final fits to determine chemical information from the spectra, which are based on the fixed peak list algorithm described next.

2.6.2. Fixed peak list algorithm

This peak fitting algorithm is used to determine only peak heights by minimizing the residuals in equation 1, after the number of peaks and their positions are fixed, e.g. with the methods in the previous section. A method for the matrix solution of an over-determined system of linear equations is used. “Over-determined” in this case means that there are more individual data points than peak heights to be fit. Equation 1 can be expressed in matrix form as:

$$I = A S \quad \text{Equation 5}$$

As with linear regression, minimization of least-square differences leads to the optimal solution of the problem. Assuming fixed values for $(m/z)_0$ (constrained peak positions), the linear least-squares solution to this equation is:

$$A = (S^T S)^{-1} S^T I \quad \text{Equation 6}$$

This solution yields unique values for A as long as the number of time-of-flight data points is greater than or equal to the number of ion peaks. As in the fitting method described in section 2.2, the results were obtained in TOF space and converted to m/z . The solution was constrained to only contain positive peak heights by iteratively removing ion peaks with negative peak heights.

2.6.3. Iterative peak assignment algorithm

As discussed below, the density of peaks produced by the formula generator at higher m/z greatly exceeds the number of peaks that can be reliably retrieved from the CIMS spectra (e.g. the number of peaks from the free peak list algorithm). Thus, a method is needed to produce a “final peak list” that contains a smaller number of peaks and that can be used for reliable data analysis. A peak assignment method is proposed here that combines the free peak positions determined by the free peak algorithm and the full formula list from the formula generator to produce the final list of peak positions with their respective sum formulas.

This method employs an iterative algorithm to automatically assign formulas. It is assumed that the number of peaks that can be fit at a given m/z (i.e. the number of degrees of freedom in the spectra) is similar to the number of peaks retrieved by the free peak algorithm. The algorithm is carried out at each integer m/z (multi-peak group) and can be described as a 9-step process:

1. Start at the lowest integer m/z , and add all peaks arising from isotopes from formulas at lower m/z . We note that the intensity of these ions will be predetermined as a fraction of the signal at the lower m/z per their isotopic fractions, and thus this step does not add additional degrees of freedom to the eventual multi-peak fit. A manual check is required to ensure that signal from isotopes at lower m/z does not exceed the total signal at the current m/z .

2. Use the free peak list algorithm (with constrained isotopes from lower m/z) to determine free peak positions. The fit is considered successful when two conditions are met: (a) peaks are spaced by at least a minimum distance (typically around 0.5 peak half-widths) and (b) the remaining residual is of the magnitude expected due to random errors.
3. Assign position of most intense peak in the free peak list to the chemical formula at the closest m/z from the full formula list, if there is at least one m/z in the full formula list within high accuracy tolerance (e.g. 10 ppm). If the most intense peak can't be assigned within accuracy tolerance, assign second most intense peak, and so on.
4. Regenerate free peak list at current m/z , using the already assigned peak positions from step 3 as constraints, continuing to constrain isotopes from lower m/z .
5. Repeat steps 3 and 4 until no more free peaks above minimum signal level (determined from expected noise level) can be found or assigned within current accuracy tolerance.
6. Increase accuracy tolerance (e.g. 10 $\rightarrow$ 20 ppm), if data is not described quantitatively (root mean square residual above expected noise level) by currently assigned chemical formulas.
7. Repeat from step 3 until the accuracy tolerance limit (e.g. 70 ppm) is reached or all free peaks are assigned.
8. Add isotope peaks from newly assigned formulas to all higher m/z .
9. Increment m/z by 1, repeat from step 2 until maximum m/z is reached.

3. Results and discussion

3.1. Determination of the Peak List for Speciated Composition

3.1.1. Generation of full formula list

We used the algorithm described in section 2.5 to generate the full formula list. For organic acids, four different seed formulas were used, formic acid (HCOOH) and formate (HCOO) for organic acids and NCO and HNCO for nitrogen-containing organic acid formula lists. Preliminary analysis of the spectra at even m/z had shown that ions without deprotonation were also present at low ion signal levels, likely formed via e^- transfer or other mechanisms such as fragmentation. Therefore, complete acid formulas were also used included in the peak list. Algorithm termination rules were: (I) $H:C < 0.2$; (II) $O:C > 2$; and (III) maximum m/z larger than 500 or maximum number of C atoms above 30. A relatively high value of $O:C > 2$ was chosen for termination rule (II) to allow the possibility of highly oxidized compounds from terpene oxidation [26]. The lists based on the four seed formulas were combined into one full formula list and contained many thousands of possible formulas (see table 3), increasing rapidly with maximum m/z considered. To illustrate this large number of possible ions compared to the instrument resolution, figure 3 shows peak densities (number of peaks per half-widths at half maximum (HWHM)), determined by the instrument functions for the free, full, and final lists. It can be seen that the peak density of both the free and the final peak lists are on the order of 1-2 per half-width, while the full list reaches values well above 10 per half-width. A similar picture arises when plotting peak distances of adjacent peaks in the form of histograms, also shown in figure 3. While the full list shows very small distances of under 0.2 for most peaks, the free and final lists have values around 0.5-1, consistent with the fact that 1-2 peaks per half-width can be fitted reliably [25], which demonstrates the need for an assignment method.

3.1.2. Application of peak assignment algorithm

Four highly averaged spectra were used for peak assignment, 2 gas and 2 particle cycles from two 24-hr periods on 26 and 29 August, 2011. The spectra were empirically divided into 3 regions where different algorithm variants were applied: (a) low m/z : 40-99, (b) mid m/z : 100-199, (c) high m/z : 200-500.

The free peak list for the low m/z region contains roughly 130 peaks, mostly well separated. Therefore, we were able to determine exact m/z positions from the free peak method and manually assign chemical formulas. This allowed us to also consider inorganic or halogen-containing formulas, which were needed for some peaks in this region (e.g. NO_3^- , Br^- , SO_4^-). We used two composition finding tools: (1) a custom tool included in Tofware; and (2) an online tool ("ChemCalc" [27]). Relative isotope abundance was used to constrain ion peak formulas while rejecting formulas that would require a higher isotope signal than was observed in the spectrum. The assignment was done in an iterative fashion by generating a new free peak list for each integer m/z after assigning one peak location and its formula (which was then constrained as a fixed peak position in the next free fit iteration). This resulted in a final list of 147 chemical formulas for the low m/z region. Figure 1 shows examples of the different peak assignment methods used in the 3 different m/z regions. It can be seen that the peak number stays constant in the low m/z 40-99 range, consistent with the manual assignment based on the free peak list. Based on the formulas, an upper limit of about 11% of the total peak area in this range may be due to non-acid peaks, either due to alternative ionization mechanisms (e.g. Cl_2^-) or fragmentation of fragile acid ions (e.g. $\text{C}_2\text{H}_3\text{O}^-$).

In the mid (100-199) and high (200-500) m/z regions, 16000 possible peaks were generated in the full formulas list, an average of 40 per integer m/z (see middle panel in figure 3) which precludes reliable quantitative fits. This compares to only 1200 peaks which were found using the free fit method. The iterative algorithm described in section 2.6.3 was applied in these regions. The minimum peak distance was set to be 0.5 peak half-widths (calculated from the instrument resolution). The accuracy tolerance range used was 10-70 ppm. Also, a few formulas containing inorganic sulfur and sulfur- and halogen-containing organic compounds were first manually entered to the initial lists. The halogen-containing compounds do not seem time-correlated with atmospheric processes and their formulas contain at least 2 oxygen atoms. They can therefore be considered to be mostly background compounds from the mass spectrometer that were also used for m/z calibration. Further, the complete series of alkanolic acid anions ($\text{CH}_3-(\text{CH}_2)_x-\text{COO}^-$) were added, because all members of the series were identified at lower m/z [22] and we expect those compounds to be present in the atmosphere [28,29]. Presence of heteroatoms other than N cannot be ruled out. Nevertheless, based on known organic aerosol elemental composition [30][3,31], an assumption that the measured signal is mainly from C, H, O, and N is reasonable. This assumption is also consistent with the lack of substantial signals at larger mass defects in this m/z range, except for the sulfur- and halogen-containing compounds that were identified manually. The middle row of figure 1 (mid m/z) shows a slight decrease in peak numbers in the final list, which is most likely due to the fact that the numerical free peak list algorithm found a number of small peaks at mass defect values outside the range of the chemical space constructed by the formula generator. These peaks were considered unimportant for the further analysis of the organic compounds. As shown in the bottom row of figure 1, the high m/z region shows many fewer peaks in the final peak list than in the full formula list

as a result of the iterative assignment method. However, the final peak list contains more peaks than the free peak list due to the fact that it is produced by regenerating the free peak list multiple times during the iterative assignment process.

The progress and success of the peak-assignment method can be evaluated from statistical properties. Figure 4A shows two statistical properties, the fraction of fit area compared to total area and the average root mean square (RMS) residual. The RMS residual was calculated by:

$$RMS = \sqrt{\frac{1}{n} \sum_{i=1}^n (S_i - F_i)^2} \quad \text{Equation 7}$$

With: n : number of data points considered

S_i : Measured signal at data point n

F_i : Fitted signal at data point n

Figure 4B shows the number of assigned peaks from all three lists and from an exploratory set of algorithm runs of the assignment algorithm in which the tolerance was kept constant during steps 3-5. Both the fraction of the fit area and the RMS residual reach plateaus around 50-70 ppm and 750 assigned peaks. This means that an increase in the tolerance limit would only lead to an increase in the number of peaks in the final list but not in a more quantitative fit of the spectra. Instead, the fits would start to be under-constrained, leading to non-representative peak heights for some assigned formulas. The upper limit of this plateau was chosen for the tolerance limit used in the assignment process described above. This value of 70 ppm is also equivalent to the 2- σ value of the fit uncertainties of multi-peak fits shown in table 2 (rows 1 and 4), which suggests that this value is a reasonable choice for determining the final peak list. Also shown on figure 4 is the number of compounds in the full formula list. The number of assigned peaks is only about 5% of the total number of possible formulas over the whole m/z range. This has two possible meanings: (a) the number of distinct atmospheric compounds that have appreciable signal is substantially smaller than the full formula list; (b) the resolution of the mass spectrometer only allows separate identification and quantification of fewer compounds than might be present in the atmosphere at appreciable concentrations.

A mass spectrometer with a higher resolution would allow better separation and identification of the compounds. However, a very high resolution of up to 250,000 (more than 60 times higher than the current instrument resolution of about 4,000) would be required to separate all peaks within 1 half peak width from the full formula list at $m/z = 400$, and thus to unequivocally determine their presence or absence with the algorithm used here. Some ultra-high resolution instruments are capable of reaching such resolution values [3]. They can only be operated, however, in a laboratory environment. Such instruments could be used in future studies to analyze filter samples collected in parallel with an HRToF-CIMS to shed further light on which of the possibilities, (a) or (b) above, may be playing a larger role. In addition, we note that a single chemical formula may correspond to multiple structural isomers that cannot be separated with the technique used here. An alternative way to get a more accurate compound identification would be to combine the mass spectrometer with an orthogonal measurement technique such as gas chromatography [32] or ion mobility spectrometry [33] to allow further

separation. However, separation techniques can introduce changes in the chemical species present in the sample (e.g. due to heating in a GC) or not transmit all compounds present [34].

Lists in the high m/z region generated separately for the gas-phase cycles contained fewer formulas (around 400) than for particle cycles (around 800). This is likely due to the lower vapor pressure of the higher molecular-weight (higher m/z) compounds, which tend to preferentially partition to the particle phase and may not be detectable in the gas phase [22]. Over 75% of the formulas extracted from the gas-phase spectra were also found in the particle spectra. This indicates that similar compounds were measured in both gas and particle phase and adds confidence to the identity of the assigned peaks. The compounds that were in the gas phase but not in the particle list were lower in carbon number and slightly higher in OS_c , consistent with a higher vapor pressure.

3.1.3. Evaluation of the results of the peak assignment algorithm

After generating peak lists for all 4 average spectra, the lists were combined into one final peak list of 1374 chemical formulas. Figure 3, panel B shows a plot of mass defect versus m/z for peak densities per half-width of the instrument resolution. Final peak densities are around 1 peak per half-width, a peak density that allows quantitative fitting. Table 3 lists statistics and properties of the different lists for all three m/z regions. Due to the presence of more inorganic compounds in the low m/z region, the full formula list has fewer formulas than the final list. The opposite is true for the high and mid m/z regions. Here, the full formula list contains many more formulas than were assigned in the final list. This is a result of the iterative assignment algorithm and reflects the limitation that only peaks further apart than approximately one peak half-width can be fitted quantitatively. We tested the completeness of the overall fit using the final peak list by comparing the areas of spectra retrieved using the final peak list to the total peak area in the raw spectra. The area fitted with the final peak list always accounted for more than 95% of the raw peak area, indicating that no major features were missed due to this procedure. Another view on the completeness of the fits is a cumulative signal plot, shown in figure S5. It shows that more than 100 ion signals are needed to account for more than 90% of the total ion signal for the particles. Conversely, only about 10-20 peaks are needed in the gas phase spectra to account for 90% of the signal. Particularly, one ion ($HCOO^-$, formate), contributes to about 80% of the total gas phase signal.

The peak assignment algorithm can be further evaluated by comparing the peak lists to data from an ultra-high resolution (UHR) electrospray ionization (ESI) mass spectrometer, acquired from filter samples in a forested region in the Rocky Mountains near the study area [3]. This instrument can be expected to resolve all individual ions present in the mass spectrum with no or little overlap. However, not all molecules present in the sample can be ionized by ESI, so the UHR-MS peak list must be considered a lower limit on the number of ions present. The UHR-MS peak list contained around 2700 molecules, a value in between the number of peaks in the full and the final lists. This indicates that the full list may contain more ions than present in the atmosphere at concentrations detectable by ESI. This is also consistent with the assumption that the CIMS high-resolution mass spectra from this study very likely contain more ions than can be separately quantified. As shown in figure 3, panel A, the calculated peak density from the UHR peak list, converted to the lower resolution of the CIMS, is most similar to the full peak list and cannot be resolved at the CIMS resolution level. Figure S2 shows a graphical representation of the peak assignment. A mass defect plot for each of the three lists (free peaks, full

formulas, final list) is shown, where all peaks are plotted as a function of mass defect and m/z in a 2-dimensional plot. The number and location of the peaks in the free peak list (left panel) is mostly reproduced in the final peak list (right panel), confirming the aim of the algorithm to fully represent the raw data. Also, approximate carbon oxidation state (OS_C) is equally captured in both the full formulas and the final list, as shown by the similar color-coding of the data. This is the case even though the full list (middle panel) shows a much higher point density than the final list.

3.2.Determination of Bulk Chemical Properties

3.2.1. Bulk Composition Determined from Formulas

Bulk chemical properties, such as average nC or OS_C , can be calculated for the results of fits to the spectra using the final peak list. Calculation of these average values requires knowledge of the relative sensitivities of all peaks used to calculate the average, which are not available for the large majority of the peaks in the mass spectra of this study. Instead, we assumed equal sensitivity for all peaks. Further studies quantifying more sensitivities for a wide range of compounds and possibly allowing the calculation of sensitivities for all peaks in the mass spectrum will allow a more accurate calculation of these average parameters. Most properties (nC, O:C and H:C ratios) can be directly calculated from the assigned formulas and their fitted peak intensities. The approximate carbon oxidation state (OS_C) can then be calculated using the following equation [6]:

$$OS_C = 2 \times O:C - H:C \quad \text{Equation 8}$$

Formulas containing nitrogen were included in the calculation of approximate OS_C . The influence of N on OS_C varies depending on functional groups, which cannot be determined from formulas alone. Therefore, no modification to the above equation was made for such formulas. The overall contribution of N-containing formulas in this study was very small (<0.1% of total signal), which did not lead to any additional uncertainty in OS_C . Overall, we expect that complex spectra from other instruments and applications can also be analyzed successfully with this method to provide similar information, as long as the contribution from heteroatoms, including N, is small.

Table 4 shows bulk parameters for gas-phase and particle spectra from the campaign average spectra. Average nC is higher in the particle phase, consistent with expectations. Gas-phase OS_C is higher than particle phase because of the fact that long-chained molecules have a lower vapor pressure and higher H:C and partition into the particle phase before getting oxidized further, while short-chain molecules with higher O:C are more common in the gas phase. In this study, gas-phase spectra were dominated by formulas containing one or two carbon atoms such as formic (CH_2O_2) and oxalic acid ($C_2H_2O_4$).

3.2.2. Evaluation of the Retrieved Bulk Composition using Synthetic Data

To evaluate the accuracy of the peak assignment method for determining bulk chemical properties, the final peak list was fitted to a set of simulated spectra, generated from the full formula list. Random peak heights were assigned to all peaks in the full formula list. The overall peak intensity distribution was scaled to be similar to that observed for atmospheric samples in this study, which was approximately exponentially decreasing with increasing m/z . To approximate the observed diversity of peak intensities, an averaged atmospheric mass spectrum was used and a stepped probability distribution function was

developed to simulate the statistical distribution of the observed ion peaks. The probability distribution function was initialized at each m/z and resulted in 70% of the peaks in the full formula list having a zero peak height, 20% having 20% of the expected signal intensity (from the atmospheric spectrum) at the respective m/z , and 10% of the peaks having the full expected signal intensity. This exercise allowed comparison of the “fitted” properties, such as fit quality and elemental parameters, to the “true” values from the simulated spectrum, from here on referred to as “input” values or spectra. A set of 100 input spectra was used to avoid statistical errors and to estimate the uncertainty of the derived properties.

Spectra and fits at individual m/z are shown in figure S3. Excellent fits are achieved for the low m/z region, where the input and fitted number of peaks is identical, as expected. In the mid m/z range, the fit quality is still very good. As the m/z increases the fits diverge from the simulated data, particularly above m/z 300. A poor fit does not indicate a problem of the peak assignment method but merely shows that some of the formulas used in the simulation did not occur when generating the final peak list from the atmospheric spectra. This results in a more stringent test of the accuracy of the bulk parameter determination, as it simulates the fact that a different formula was assigned than was in the simulated input spectrum.

Input and fitted spectra are compared in figure 5 as Kroll diagrams (approximate OS_c vs. nC) [6]. While individual data points differ between input and fitted spectra, overall trends such as OS_c as a function of nC and total ion signal versus OS_c agree quite well. More detailed plots of trends in elemental parameters are shown in figure 6 for OS_c vs. nC , ion signal vs. nC , and ion signal vs. OS_c . These graphs include results from all 100 simulations. Error bars in the plots are 90% confidence limits (2σ) from the simulations. Overall uncertainties of various elemental parameters are calculated by statistical analysis of the differences of these parameters between simulated and fitted spectra (see table 5). In general, differences between simulated data and fits are small, which confirms that OS_c can be reproduced reliably by fitting complex spectra with the simplified final list. Figure 6 also includes a comparison table for overall averages for elemental parameters, which agree quite well between input and fit. The average bias between simulation and fit is shown in figure S4 (red traces) as histograms from all simulations for both OS_c and nC . It can be seen that, while OS_c and nC determined by the fit are close to the simulation input values, they are slightly over- and underestimated, respectively.

The fact that good agreement in bulk parameters is observed despite the fact that not all ions are identified correctly is related to the fact that mass defect and elemental parameters, particularly OS_c , are correlated. The approximate relationship between these two properties across the m/z range is apparent in figure 7 for both the full and the final peak lists, also confirmed by r^2 values above 0.6. In the next section we explore an alternative method where we exploit this relationship to estimate OS_c directly from the raw mass spectra.

3.2.3. Determination of bulk chemical properties from raw mass spectra

It is likely that the agreement in bulk compositional parameters between simulated and fitted spectra is helped by the fact that many elemental parameters are related to the mass defect at each m/z . In other words, an ion located in a given relative position in a multi-peak group will be representative of the average OS_c (and other elemental parameters such as O:C and H:C ratios) of the actual compound(s)

contributing to the measured signal at its approximate mass defect value, even if the assignment is not correct or if there are smaller ion signals nearby that our procedure (given the limited resolution of the spectra) cannot quantify. Under the assumptions that chemical formulas only contain C, H, N and O, some bulk compositional parameters can be determined from the mass defect at each m/z in a raw spectrum, without assigning and fitting any peaks. Given the complex and laborious nature of the peak list determination, such a procedure would be useful to provide quick estimates of bulk chemical parameters, e.g. during data acquisition.

To allow such estimation, we generated empirical functions for OS_c and carbon number (nC) as a function of m/z and mass defect. These functions were derived from multi-variate fits of 2-D plots of the respective parameter. The full formula list was used to generate these 2-D plots, because it contained a larger range of possible formulas, therefore spanning the entire chemical CHON space. In particular, the two following equations were used:

$$OS_c = A * MD + B * m/z + C * [MD / (m/z)] + D \quad \text{Equation 9}$$

$$nC = A' * MD + B' * m/z + D' \quad \text{Equation 10}$$

With:

A, A', B, B', C, D, D': adjustable fit parameters

MD: mass defect

The fit parameters and their errors are listed in table 6. Figure 8A compares the result of the multivariate fit for OS_c to the values for individual peaks. The fact that marker colors blend into the background colors shows a good overall agreement between multivariate fit and data. A similar result is achieved for the carbon number (figure 8B), except that the deviations are larger, reflecting the fact that carbon number is not as tightly constrained as OS_c by mass defect and m/z . Further, only carbon numbers up to about 15 can be reliably extracted from the raw spectra due to the limited m/z range up to 400 analyzed here, indicated by a white space in figure 8B. Lower carbon numbers at a given m/z are observed at lower mass defects, so for $nC > 15$ using the data up to m/z 400 would result in a positive bias on nC . Therefore, demonstration of the method is limited to $nC \leq 15$.

To test this method, equation 9 is used to calculate signal-weighted OS_c for the average raw spectra from all particle and gas-phase cycles. The entire spectral range was used for this calculation, but some integer m/z known to contain large ion signals of inorganic compounds were excluded. The estimated OS_c are 0.38 ± 0.12 and 1.17 ± 0.18 , for particle and gas-phase cycles, respectively, compared with the values calculated from the results of the peak fitting procedure of 0.15 and 1.54, respectively.

Uncertainties were calculated by error propagation of the uncertainties in the individual fit parameters. Additional uncertainty arises from isotopes at adjacent m/z 's that are not considered explicitly by this direct method. However, this effect can be considered minor since the mass defect for the isotope signals is quite similar to that of the parent ion and therefore the contribution to OS_c will be similar as well.

Average nC estimated using equation 10 are 4.3 ± 0.1 and 2.1 ± 0.2 for particle and gas phase spectra, respectively, which compare reasonably well to the final peak list values of 3.9 and 1.5, respectively, especially when considering the worse two-dimensional fit for nC (see above).

To evaluate this method, we compared estimated OS_c and nC to the simulated spectra discussed in section 3.2.2. The results are shown in figures 6 and S4. Also shown are results from the raw spectra method that were generated by removing the $nC \leq 15$ constraint. The agreement between input and directly calculated parameters is quite good, considering that no peak fitting or assignment was used to arrive at this result. The missing nC information is most apparent in the signal vs OS_c trace, where a slightly lower signal at lower OS_c is apparent. This can be explained by the fact that formulas with $nC > 15$ have similar OS_c values.

This method should give valid results for mass spectra from all CIMS instruments. In the case of clustering reagent ions, such as iodide or nitrate, the reagent ion has to be removed from the formulas before determining mass defect and bulk properties. Since the empirical fit parameters were determined for CHON, they are only valid for spectra in which these elements dominate.

4. Conclusions

We present new methods to analyze complex mass spectra when the instrument resolution is not sufficient to identify and quantify all individual peaks present. First, a method for peak identification is developed to provide the best formulas incorporating all known constraints. The method starts with an algorithm to calculate all possible chemical formulas, and separately identifying possible peak positions in the multi-peak systems encountered in a mass spectrum. The chemical space and the peak positions in a mass spectrum are then combined by an iterative assignment process to obtain an optimum list of formulas to be used for quantitative fitting of the mass spectrum. We applied this method to spectra acquired from the atmosphere in a pine forest in Colorado in 2011, but these methods should be applicable to spectra from complex mixtures from a wide range of applications and instrument resolutions, [requiring recalculation of all possible formulas \("full formula list"\) when using different ion chemistries](#). The number of peaks that can be constrained is much lower than the number of chemically possible formulas.

Second, we introduce a method of deriving bulk chemical information from the spectra. This method uses the results of the peak identification and fitting methods. Simulations using synthetic data were conducted to validate this bulk method. While not all peaks from the simulation input are retrieved correctly by the speciated method, we show that bulk chemical information, such as oxidation state, carbon number and functions thereof, can be quantitatively determined using the results of the peak identification and fitting methods. The success of the bulk results despite the limitations of peak identification is related to the fact that oxidation state and carbon number are related to the mass defect at each m/z .

Finally, we make use of the relationship between mass defect and bulk chemical parameters to develop a second method in which bulk composition parameters are determined without any peak assignment

or fitting. Again, comparison to simulated data allows validation of this method. It shows slightly higher uncertainty than the assignment and fitting method, but has the advantage that it can be used for rapid determination of elemental parameters, for example as an online method to get chemical information while the spectra are being acquired. This method should be applicable to any mass spectra dominated by C,H,O and N-containing formulas, regardless of ionization technique. It could also be extended to obtain accurate values for any bulk parameters that correlate directly with mass defect at every m/z .

Acknowledgements

We thank Prof. Lynn Mazzoleni for providing the ultra-high resolution data shown in figure 3.

This work was partially supported by DOE SBIR DE-SC0011105, DOE BER/ASR DE-SC0006035, NSF AGS-1243354, and NOAA NA13OAR4310063. We thank the ToF-CIMS user community for useful discussions.

5. References

- [1] J.F. Hamilton, P.J. Webb, A.C. Lewis, J.R. Hopkins, S. Smith, P. Davy, Partially oxidised organic components in urban aerosol using GCXGC-TOF/MS, *Atmos. Chem. Phys.* 4 (2004) 1279–1290. doi:10.5194/acp-4-1279-2004.
- [2] A.H. Goldstein, I.E. Galbally, Known and unexplored organic constituents in the earth's atmosphere, *Environ. Sci. Technol.* 41 (2007) 1514–1521. doi:10.1021/es072476p.
- [3] L.R. Mazzoleni, P. Saranjampour, M.M. Dalbec, V. Samburova, A.G. Hallar, B. Zielinska, et al., Identification of water-soluble organic carbon in non-urban aerosols using ultrahigh-resolution FT-ICR mass spectrometry: organic anions, *Environ. Chem.* 9 (2012) 285–297. doi:10.1071/en11167.
- [4] D. Aljawhary, A.K.Y. Lee, J.P.D. Abbatt, High-resolution chemical ionization mass spectrometry (ToF-CIMS): application to study SOA composition and processing, *Atmos. Meas. Tech.* 6 (2013) 3211–3224. doi:10.5194/amt-6-3211-2013.
- [5] L. Cappellin, F. Biasioli, P.M. Granitto, E. Schuhfried, C. Soukoulis, F. Costa, et al., On data analysis in PTR-TOF-MS: From raw spectra to data mining, *Sensors and Actuators B-Chemical*. 155 (2011) 183–190. doi:10.1016/j.snb.2010.11.044.
- [6] J.H. Kroll, N.M. Donahue, J.L. Jimenez, S.H. Kessler, M.R. Canagaratna, K.R. Wilson, et al., Carbon oxidation state as a metric for describing the chemistry of atmospheric organic aerosol, *Nat. Chem.* 3 (2011) 133–139. doi:10.1038/Nchem.948.
- [7] T. Kind, O. Fiehn, Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm, *BMC Bioinformatics*. 7 (2006). doi:10.1186/1471-2105-7-234.

- 652 [8] N.M. Donahue, J.H. Kroll, S.N. Pandis, A.L. Robinson, A two-dimensional volatility basis set – Part
653 2: Diagnostics of organic-aerosol evolution, *Atmos. Chem. Phys.* 12 (2012) 615–634.
654 doi:10.5194/acp-12-615-2012.
- 655 [9] T. Kind, O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by
656 accurate mass spectrometry, *BMC Bioinformatics*. 8 (2007). doi:10.1186/1471-2105-8-105.
- 657 [10] J. Van Katwijk, Improved evaluation of high-resolution mass spectra through the investigation of
658 peak shapes and isotopic contributions, *Int. J. Mass Spectrom. Ion Phys.* 39 (1981) 287–310.
659 doi:10.1016/0020-7381(81)87004-0.
- 660 [11] R. Feng Konishi, Y. Bell, A.W., High Accuracy Molecular Weight Determination and Variation
661 Characterization of Proteins Up To 80 ku by Ionspray Mass Spectrometry, *J. Am. Soc. Mass*
662 *Spectrom.* 2 (1991) 381–401. doi:10.1016/1044-0305(91)85005-Q.
- 663 [12] E.F. Strittmatter, N. Rodriguez, R.D. Smith, High mass measurement accuracy determination for
664 proteomics using multivariate regression fitting: Application to electrospray ionization time-of-
665 flight mass spectrometry, *Anal. Chem.* 75 (2003) 460–468. doi:10.1021/ac026057g.
- 666 [13] D.M. Horn, R.A. Zubarev, F.W. McLafferty, Automated reduction and interpretation of high
667 resolution electrospray mass spectra of large molecules, *J. Am. Soc. Mass Spectrom.* 11 (2000)
668 320–332. doi:10.1016/s1044-0305(99)00157-9.
- 669 [14] L.L. House, M.A. Clyde, R.L. Wolpert, Bayesian Nonparametric Models for Peak Identification in
670 MALDI-TOF Mass Spectroscopy, *Ann. Appl. Stat.* 5 (2011) 1488–1511. doi:10.1214/10-aos450.
- 671 [15] P.F. DeCarlo, J.R. Kimmel, A. Trimborn, M.J. Northway, J.T. Jayne, A.C. Aiken, et al., Field-
672 deployable, high-resolution, time-of-flight aerosol mass spectrometer, *Anal. Chem.* 78 (2006)
673 8281–8289. doi:10.1021/ac061249n.
- 674 [16] M. Graus, M. Mueller, A. Hansel, High Resolution PTR-TOF: Quantification and Formula
675 Confirmation of VOC in Real Time, *J. Am. Soc. Mass Spectrom.* 21 (2010) 1037–1044.
676 doi:10.1016/j.jasms.2010.02.006.
- 677 [17] M. Mueller, C. George, B. D’Anna, Enhanced spectral analysis of C-TOF Aerosol Mass
678 Spectrometer data: Iterative residual analysis and cumulative peak fitting, *Int. J. Mass Spectrom.*
679 306 (2011) 1–8. doi:10.1016/j.ijms.2011.04.007.
- 680 [18] M. Mueller, T. Mikoviny, W. Jud, B. D’Anna, A. Wisthaler, A new software tool for the analysis of
681 high resolution PTR-TOF mass spectra, *Chemom. Intell. Lab. Syst.* 127 (2013) 158–165.
682 doi:10.1016/j.chemolab.2013.06.011.
- 683 [19] R.L.N. Yatavelli, J.A. Thornton, Particulate Organic Matter Detection Using a Micro-Orifice
684 Volatilization Impactor Coupled to a Chemical Ionization Mass Spectrometer (MOVI-CIMS),
685 *Aerosol Sci. Technol.* 44 (2010) 61–74. doi:10.1080/02786820903380233.

- 686 [20] R.L.N. Yatavelli, F. Lopez-Hilfiker, J.D. Wargo, J.R. Kimmel, M.J. Cubison, T.H. Bertram, et al., A
687 Chemical Ionization High-Resolution Time-of-Flight Mass Spectrometer Coupled to a Micro
688 Orifice Volatilization Impactor (MOVI-HRToF-CIMS) for Analysis of Gas and Particle-Phase Organic
689 Species, *Aerosol Sci. Technol.* 46 (2012) 1313–1327. doi:10.1080/02786826.2012.712236.
- 690 [21] J. Ortega, A. Turnipseed, A.B. Guenther, T.G. Karl, D.A. Day, D. Gochis, et al., Overview of the
691 Manitou Experimental Forest Observatory: site description and selected science results from
692 2008 to 2013, *Atmos. Chem. Phys.* 14 (2014) 6345–6367. doi:10.5194/acp-14-6345-2014.
- 693 [22] R.L.N. Yatavelli, H. Stark, S.L. Thompson, J.R. Kimmel, M.J. Cubison, D.A. Day, et al.,
694 Semicontinuous measurements of gas–particle partitioning of organic acids in a ponderosa pine
695 forest using a MOVI-HRToF-CIMS, *Atmos. Chem. Phys.* 14 (2014) 1527–1546. doi:10.5194/acp-14-
696 1527-2014.
- 697 [23] P. Veres, J.M. Roberts, C. Warneke, D. Welsh-Bon, M. Zahniser, S. Herndon, et al., Development
698 of negative-ion proton-transfer chemical-ionization mass spectrometry (NI-PT-CIMS) for the
699 measurement of gas-phase organic acids in the atmosphere, *Int. J. Mass Spectrom.* 274 (2008)
700 48–55. doi:10.1016/j.ijms.2008.04.032.
- 701 [24] T.H. Bertram, J.R. Kimmel, T.A. Crisp, O.S. Ryder, R.L.N. Yatavelli, J.A. Thornton, et al., A field-
702 deployable, chemical ionization time-of-flight mass spectrometer, *Atmos. Meas. Tech.* 4 (2011)
703 1471–1479. doi:10.5194/amt-4-1471-2011.
- 704 [25] M.J.. D.S. and J.L.J. Cubison, Statistical precision of the intensities retrieved from constrained
705 fitting of overlapping peaks in high-resolution mass spectra., *Atmos. Meas. Tech. Discuss.* 7
706 (2014) 12617–12647. doi:10.5194/amtd-7-12617-2014.
- 707 [26] M. Ehn, J.A. Thornton, E. Kleist, M. Sipila, H. Junninen, I. Pullinen, et al., A large source of low-
708 volatility secondary organic aerosol, *Nature.* 506 (2014) 476–+. doi:10.1038/nature13032.
- 709 [27] L. Patiny, A. Borel, ChemCalc: A Building Block for Tomorrow’s Chemical Infrastructure, *J. Chem.*
710 *Inf. Model.* 53 (2013) 1223–1228. doi:10.1021/ci300563h.
- 711 [28] I.G. Kavouras, N. Mihalopoulos, E.G. Stephanou, Formation of atmospheric particles from organic
712 acids produced by forests, *Nature.* 395 (1998) 683–686. doi:10.1038/27179.
- 713 [29] S. Samy, L.R. Mazzoleni, S. Mishra, B. Zielinska, A.G. Hallar, Water-soluble organic compounds at
714 a mountain-top site in Colorado, USA, *Atmos. Environ.* 44 (2010) 1663–1671.
715 doi:10.1016/j.atmosenv.2010.01.033.
- 716 [30] A.C. Aiken, P.F. Decarlo, J.H. Kroll, D.R. Worsnop, J.A. Huffman, K.S. Docherty, et al., O/C and
717 OM/OC ratios of primary, secondary, and ambient organic aerosols with high-resolution time-of-
718 flight aerosol mass spectrometry, *Environ. Sci. Technol.* 42 (2008) 4478–4485.
719 doi:10.1021/Es703009q.

- 720 [31] K.S. Docherty, A.C. Aiken, J.A. Huffman, I.M. Ulbrich, P.F. DeCarlo, D. Sueper, et al., The 2005
721 Study of Organic Aerosols at Riverside (SOAR-1): instrumental intercomparisons and fine particle
722 composition, *Atmos. Chem. Phys.* 11 (2011) 12387–12420. doi:10.5194/acp-11-12387-2011.
- 723 [32] B.J. Williams, A.H. Goldstein, N.M. Kreisberg, S. V Hering, An in-situ instrument for speciated
724 organic composition of atmospheric aerosols: Thermal Desorption Aerosol GC/MS-FID (TAG),
725 *Aerosol Sci. Technol.* 40 (2006) 627–638. doi:10.1080/02786820600754631.
- 726 [33] A.B. Kanu, P. Dwivedi, M. Tam, L. Matz, H.H. Hill Jr., Ion mobility-mass spectrometry, *J. Mass*
727 *Spectrom.* 43 (2008) 1–22. doi:10.1002/jms.1383.
- 728 [34] B.J. Williams, J.T. Jayne, A.T. Lambe, T. Hohaus, J.R. Kimmel, D. Sueper, et al., The First Combined
729 Thermal Desorption Aerosol Gas Chromatograph—Aerosol Mass Spectrometer (TAG-AMS),
730 *Aerosol Sci. Technol.* 48 (2014) 358–370. doi:10.1080/02786826.2013.875114.

731

731

732 **6. Tables:**

Exact m/z	Assigned formula	Accuracy $\pm 1 \sigma$ (ppm)
36.9659	$^{37}\text{Cl}^-$	3 ± 2
41.9985	NCO^-	5 ± 4
45.9935	NO_2^-	3 ± 1
71.0138	$\text{C}_3\text{H}_3\text{O}_2^-$	4 ± 2
119.0350	$\text{C}_4\text{H}_7\text{O}_4^-$	7 ± 2
304.9100	$\text{C}_{12}\text{H}_5\text{Cl}_4\text{O}^-$	5 ± 4
338.8710	$\text{C}_{12}\text{H}_4\text{Cl}_5\text{O}^-$	4 ± 3

733 Table 1. Isolated peaks used for m/z calibration

Exact m/z	Assigned formula	Average bias $\pm 1\sigma$ Accuracy (ppm)	Comments
63.9619	SO_2^-	-3 ± 34	Multiple peaks
69.9377	Cl_2^-	3.2 ± 3.4	Single peak
88.9875	C_2HO_4^-	0.3 ± 3.0	Single peak
217.0712	$\text{C}_9\text{H}_{13}\text{O}_6^-$	3 ± 34	Multiple peaks

734 Table 2. Mass accuracies of select peaks

	Low m/z (0-99)	Mid m/z (100-199)	High m/z (200-500)	All m/z
Total area fit (%)	104*	101	100	
Number of free peaks	136	361	802	1299
Number of Full formulas	67	819	15208	16094
Number of peaks assigned in the final list	149	306	919	1374
Number of inorganic/hetero ⁺	46	13	11	70

735 Table 3. Results of different fit methods over the three m/z regions. *value above 100% due to
 736 overestimation of acetate reagent ion isotopes due to saturation effects at m/z 59. ⁺atoms other than
 737 C,H,O,N

738

Parameter	Gas Phase	Particle Phase
H:C	1.7	1.3
O:C	1.6	0.73
nC	1.5	3.9
OS _c	1.5	0.14

739 Table 4: Bulk elemental parameters from analysis of atmospheric spectra, campaign averages

Parameter	Precision (1 σ)	Accuracy (fit – simulation)
H:C	0.011	-0.035
O:C	0.008	0.011
nC	0.083	-0.520

OS _c	0.010	0.057
-----------------	-------	-------

Table 5: Uncertainties of elemental parameters from analysis of simulated spectra

Fit parameter	OS _c		Fit parameter	nC	
	Value	Uncertainty		Value	Uncertainty
A	2.796E+00	6.12E-01	A'	2.188E+01	1.16E+00
B	-1.304E-03	3.03E-04	B'	4.677E-02	1.16E-3
C	-4.054E+03	1.63E+02			
D	1.295E+00	7.9E-02	D'	-8.011E-01	2.906E-01

Table 6: Fit parameters from multivariate fit to determine approximate carbon oxidation state (OS_c) and carbon number (nC) as function of mass defect and m/z

7. Figures:

Figure 1: Raw mass spectra at m/z 89, 189 and 255 with multi-peak fits. Rows correspond to the three m/z regions used in this work. Columns correspond to the three peak lists used in this work.

Figure 2: Data analysis flow used in this study

756

757

758

759 Figure 3: Panel A: histogram of distances between adjacent peaks, panel B1-3: peak density (colors,
 760 black = zero density) per instrument peak half-width as function of mass defect and m/z . Values are for
 761 the three peak lists used in this work

762

763

Figure 4. Statistical parameters for the automatic peak assignment method when used for high m/z region (200-400). Bottom axes are accuracy tolerance for peak assignment. Panel A shows merit functions: average root mean square fit residual ("RMS") and total area fit; the approximate free peak m/z accuracy is shown as a vertical shaded area. Panel B: number of assigned peaks from different lists and from the iterative algorithm with fixed tolerance

Figure 5: Kroll diagram (approximate carbon oxidation state vs. carbon number) comparison of input spectra from simulation (A) and fitted to final list (B). Markers colored and sized by log(peak area). Grey background reflects chemical space from formula generator

Figure 6: Comparison of simulation input and fitted bulk chemical parameters from both final peak list and raw spectrum method: oxidation state vs. carbon number (A), signal vs. carbon number (B) and signal vs. approximate oxidation state (C). The top sections of each graph show residuals between simulation and final list fits from one individual (highlighted) and all 100 simulations; bottom sections show simulation input and results from final list and raw spectrum method. Error bars are 2 standard deviations from all simulation residuals. The table gives an overview of average bulk parameters for entire spectra.

790

791

792

793 Figure 7: Relationship between oxidation state and mass defect shown for four m/z values. Lines are
 794 linear least square fits. Pearson r^2 correlation coefficients are listed in legends.

795

Figure 8: Mass defect plot and multi-variate fit of OS_c (panel A) and carbon number (panel B). Markers are the values for individual peaks, colored by OS_c and carbon number, respectively; background images are results of multi-variate fits, also colored by OS_c and carbon number using the same color scales as for the markers.