Supplement of Atmos. Chem. Phys., 16, 12715–12731, 2016
http://www.atmos-chem-phys.net/16/12715/2016/
doi:10.5194/acp-16-12715-2016-supplement

Atmospheric Chemistry and Physics Open Access EGU

Supplement of

Source characterization of highly oxidized multifunctional compounds in a boreal forest environment using positive matrix factorization

Chao Yan et al.

Correspondence to: Chao Yan (chao.yan@helsinki.fi) and Wei Nie (niewei@nju.edu.cn)

1. Estimation of uncertainty from laboratory experiments

The CI-APi-TOF can detect charged ions and clusters in a wide mass range up to ~3000 Th, over which the transmission efficiency (the fraction between the number of ions that reach the detector and the number of ions generated in the chemical ionization source) changes significantly. Its potential influence on signal counting statistics needs examination. Thus, a set of laboratory experiments were conducted to find out a proper equation to describe the detection uncertainty.

The schematic of the experiment set-up was as shown in Fig. S1. A temperature controlled permeation source was connected to the CI-inlet. Nitrogen gas (N_2) was used as both the carrier gas and the dilution air. The optimized flow rate for N_2 through the permeation source was found to be 100 milliliter per minute (mlpm), which ensured that there were enough permeated chemicals being carried out, and in the meanwhile that the temperature inside the permeation source was well controlled. The outflow of the permeation source was further diluted by N_2 flow (~ 10 liter per minute, lpm) before entering the chemical ionization inlet (CI-inlet) as the sample flow. The flow rate of the sample flow was adjusted by varying the difference between the total flow and the sheath flow of the CI-inlet, which were set as 30 lpm and 20 lpm, respectively. All these flow rates were kept unchanged throughout the experiment.

Two different chemicals ($CF_3(CF_2)_2COOH$ and $CF_3(CF_2)_7COOH$) were used in the permeation source, for which the temperature range were 20~60 °C and 30~85 °C, respectively. With the same chemical, we repeated the experiment twice by using different instrument tunings, which had optimized transmission in low-mass range (<200 Th) and high-mass range (>800 Th), respectively. Fig. S2 shows an example of signal variation of major peaks in different temperature stages.

Based on Eq. 6, the σ_{noise} needs to be fixed first. Different from AMS measurement, CI-APi-TOF was running without a routine blank measurement. As an alternative, we used the "blank masses" (800 – 1000 Th), where few peaks are located, to fit the σ_{noise}. Fig. S3 shows σ_{noise} for each mass-to-charge in the blank mass range, calculated from datasets of low-mass setting, high-mass setting, and the setting used in the ambient measurement. The σ_{noise} for each method are in good agreement, indicating that σ_{noise} is mostly independent of instrument tuning. Note that some discrete outliers were observed in high-mass tuning because in this tuning the instrument was sensitive enough to detect some large ions or clusters. We apply a constant value of 0.035 for σ_{noise}, taken as the median of standard deviations calculated for each unit mass in the blank mass range, although a weak decreasing trend was observed with increasing mass-to-charge.

The *a* value in Eq. 7 is derived by fitting the analytical uncertainty to the signal strength. Fig. S4 depicts the analytical uncertainty and signal intensity using all datasets, including different permeation sources and different instrument tunings. In general, the fitting of uncertainties in all experiments follows the same trend, implying an independence of both chemical species and instrument conditions, over a large range of signal intensities between 0.1~10000 cps

(count per second). Since the strongest signal in the ambient measurement is about 20 cps, we fitted the uncertainty only with peaks below this value including isotopic peaks. As shown in Fig. S5, for 300-second integrated data, the best fitted value for $a/\sqrt{t_s}$ is 0.074±0.005 (corresponding to the upper and lower bounds of 95% confidence), and corresponding a value is 1.28±0.1.

2. **Estimation of uncertainty from ambient measurement data**

To assess if the uncertainty derived from the laboratory experiment agrees with what we observe in actual measurement of ambient air, we devised a simple technique to estimate the instrumental noise based on the ambient air data independently. We tested the technique on the same data that was input for PMF, containing 9084 measured time steps and 450 variables from 201 – 650 Th.

The basis of the method is approximating instrument noise as the difference of measured signal (in unit cps) relative to the signal's moving median over 25 minutes, i.e. 5 data points (Fig S6). Assuming changes in the chemical composition generally happen over a longer timescale than the timescale of our measurement (5 minutes), we can consider the deviation from the moving median to result mostly from the uncertainty of the measurement rather than actual chemical changes in the aerosol. To minimize the effect of real atmospheric variations of short duration on the median filter, the median filter for each ion was calculated after omitting the highest 10% of the measurement points.

In order to calculate a single uncertainty estimate using the above method, utilizing the entire data set of 9084 data points for each of the 450 ions, we further divided the signals from each ion into ten bins, or "deciles", according to their signal intensity (cps). Decile 1 corresponds to the lowest (0-10 %), and decile 10 to the highest (90-100 %), observations. An example of the signal bin regions for mass 339 Th is shown in Fig. S6. For each of the 450 ions, we are now left with 10 bins, each containing 9084/10≈908 data points. For each data point we then calculated the deviation from the moving median, yielding 908 observations of measurement uncertainty in each bin.

We quantified the uncertainty related to each of the 450x10 bins by assuming the deviations are normally distributed, and fitted a Gaussian distribution to the data points in each bin (see example in Fig. S7). We took the standard deviation σ of each fit as a single parameter measure of the uncertainty of the corresponding bin. Next, we parameterized the noise dependence as a function of the signal for each ion, by plotting σ vs the average signal of each bin (see Fig S8 for examples), and constructing a (weighted non-linear) least squares fit to the data according to

$$f(a,e) = c\sqrt{s} + e \quad \text{(Eq. S1)}$$

Parameter e can be understood as the electronic noise of the instrument, and parameter c is similar to the $\frac{a}{\sqrt{t_s}}$ parameter in equation (Eq. 6), which defines the square root dependence constant. As the fitted data use the same integration time (300 seconds), we fit c and e as constant parameters.

After obtaining 450 values for e and c, we excluded values from a few fits with clearly non-physical outcome (such as negative c or e, or clear outliers outside of two standard deviations from the mean). For the remaining values, we

took the mean (weighted by the inverses of their uncertainties) over all the ions as the final value, and the standard deviation as its respective uncertainty. The final expression for the uncertainty estimation was

$$f(a,e) = (0.0628 \pm 0.0168)\sqrt{s} + (0.0206 \pm 0.0119) \quad \text{(Eq. S2)}$$

3. Data censoring

In the data matrix, there were many points (i.e. mass j at time i) below the detection limit, defined as $3\sigma_{noise}$.The data quality for these points were questionable and we therefore "censored" these data by replacing the X_{ij} and S_{ij} with σ_{noise} and $6\sigma_{noise}$, respectively. A similar approach was suggested by Polissar et al (1998), and has been adopted in many studies since then. In this work, the signal-to-noise ratio (SNR) for censored data were 0.17 ("bad signal"), most of which will be further downweighted by a factor of 10, and will thus have little influence on the model results.

This practice is still under debate, and it is criticized by the developer of the PMF model P. Paatero (*cite the comment*) for creating ghost factors and causing bias in the results. To check if the results in this work were affected by censoring data, we also ran PMF with uncensored data and errors. We then use the uncentered correlation between solutions (here we use the factor spectra) from censored and uncensored data to evaluate their similarity, and the coefficients are given in Table S1. The solutions appear to be almost identical, meaning that censoring data did not bias our results, nor did it provide any benefit. Thus, we would also not recommend censoring data in future studies using similar data sets.

93 Table S1. Uncentered correlation coefficients between PMF solutions using censored data and uncensored data.

Number of factors	factor type	UC coefficient
2	daytime	0.9997
	nighttime	0.9998
3	daytime	0.9998
	nighttime	0.9998
	201 Th	0.9991
4	daytime type-1	0.9980
	daytime type-2	0.9834
	nighttime	0.9998
	201 Th	0.9989
5	daytime type-1	0.9994
	daytime type-2	0.9981
	daytime type-3	0.9981
	nighttime	0.9997
	201 Th	0.9978
6	daytime type-1	0.9995
	daytime type-2	0.9982
	daytime type-3	0.9991
	nighttime type 1	0.9998
	nighttime type-2	0.9997
	transport	0.9988

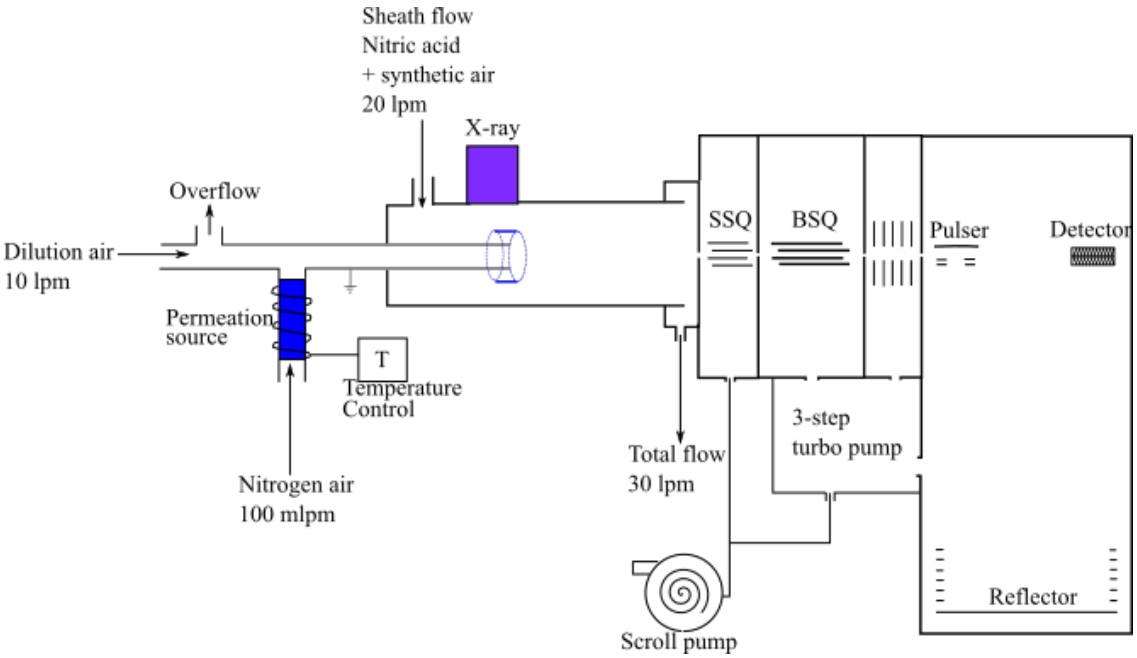


Fig. S1. The schematic of the laboratory experiment assembly. All the flows were constant throughout the experiments, while different chemicals, temperatures and instrument tunings were tested.

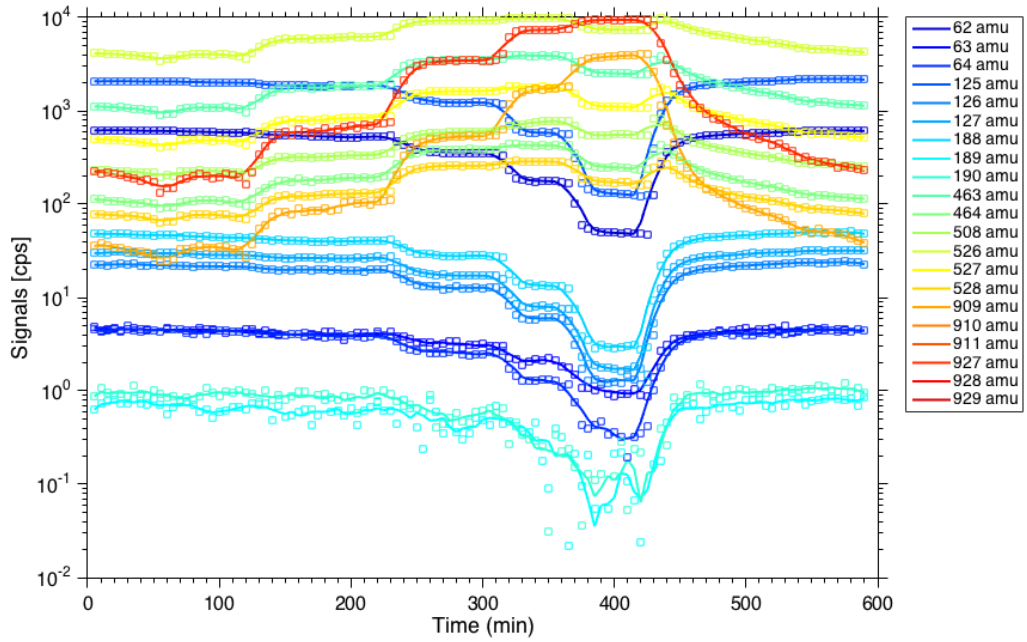


Fig. S2. An example of signal variations at different temperatures in the experiment using $CF_3(CF_2)_7COOH$ and high-mass tuning. The temperatures increased stepwise (i.e. 30, 40, 50, 60, 70, and 85 °C), and the signals showed stepwise change simultaneously. For further error fitting (Fig. S4 and Fig. S5), only steady-state data were used.

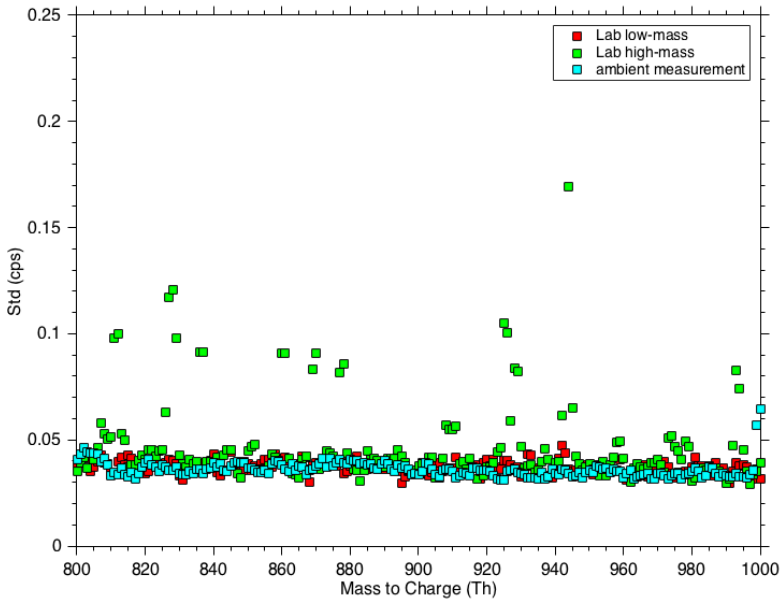


Fig. S3. Background estimation for data from low-mass tuning (red), high-mass tuning (green), and tuning for ambient measurement. 800~1000 amu was selected as the 'blank mass' though some peaks can be observed in high mass tuning.

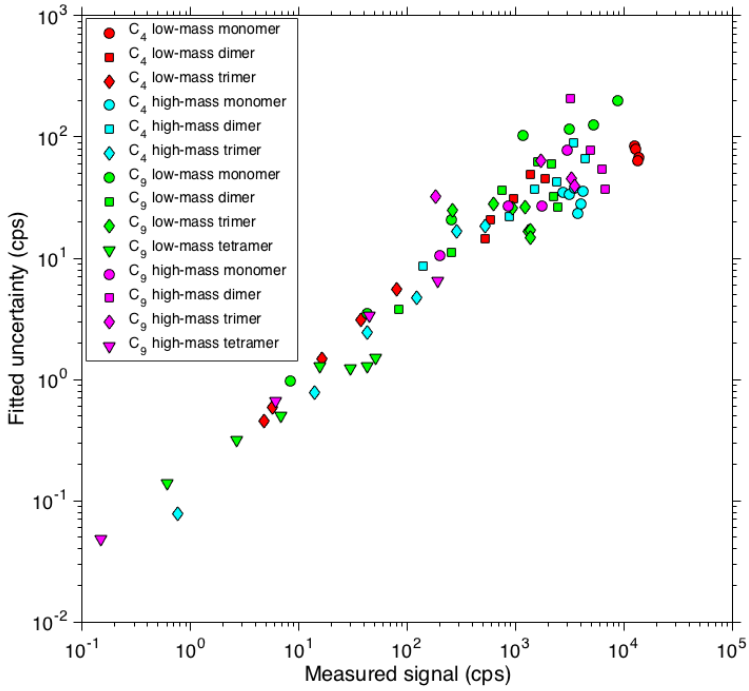
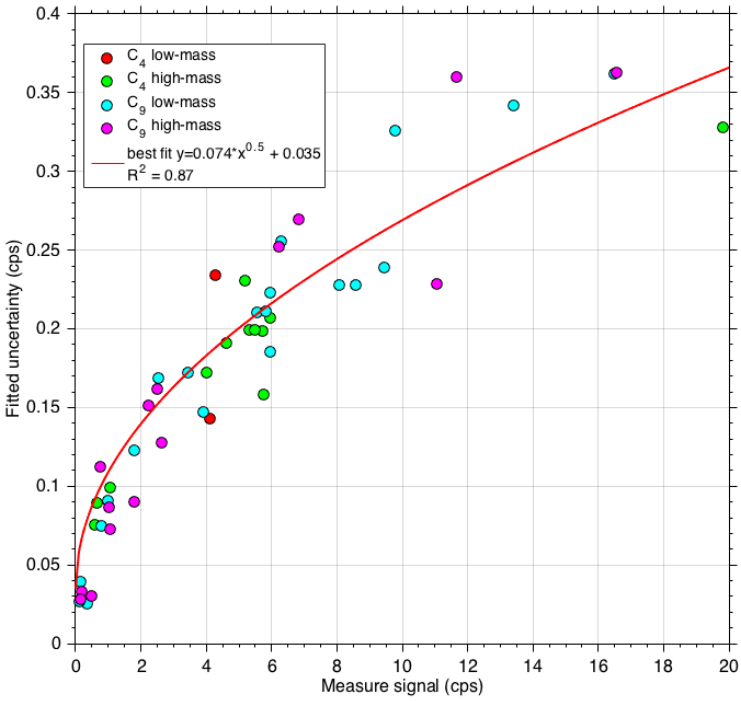


Fig. S4. The analytical uncertainty versus signal strength for different chemicals and instrument tunings. Different combinations of a certain chemical and a certain tuning are marked with different color. Within each combination, different shapes are used to mark different chemical oligomers or the reagent ions.



a t

Fig. S5. Fitting of uncertainty versus signal based on Eq. 6. Only signals smaller than 20 cps (in the typical atmospheric level) were used. The same color code was used as in Figure S4.

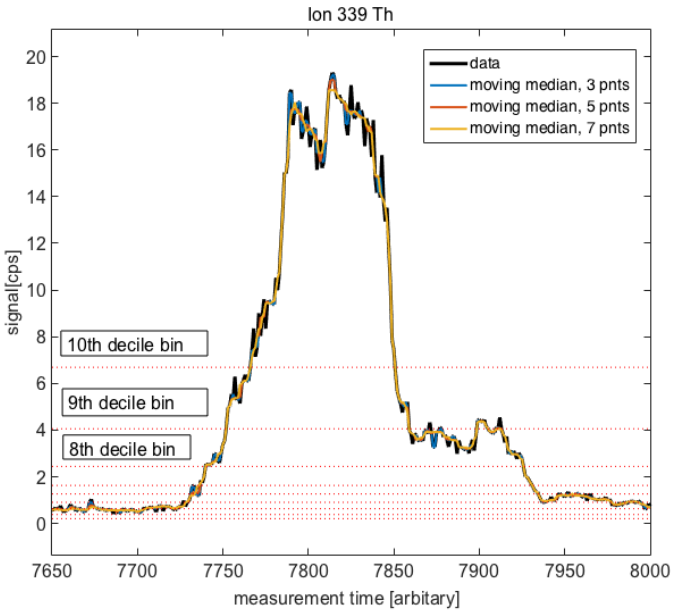


Fig. S6. Example of estimating noise in ambient data, shown for a single UMR signal (339 Th) over approximately one diurnal cycle. Measured data is shown as a black solid line while the colored solid lines correspond to using running median filter with 3 (blue), 5 (red) or 7 (yellow) points. For the final analysis, a window of 5 points was selected. Point assignation to decile bins according to their signal is additionally illustrated by the red dotted lines.

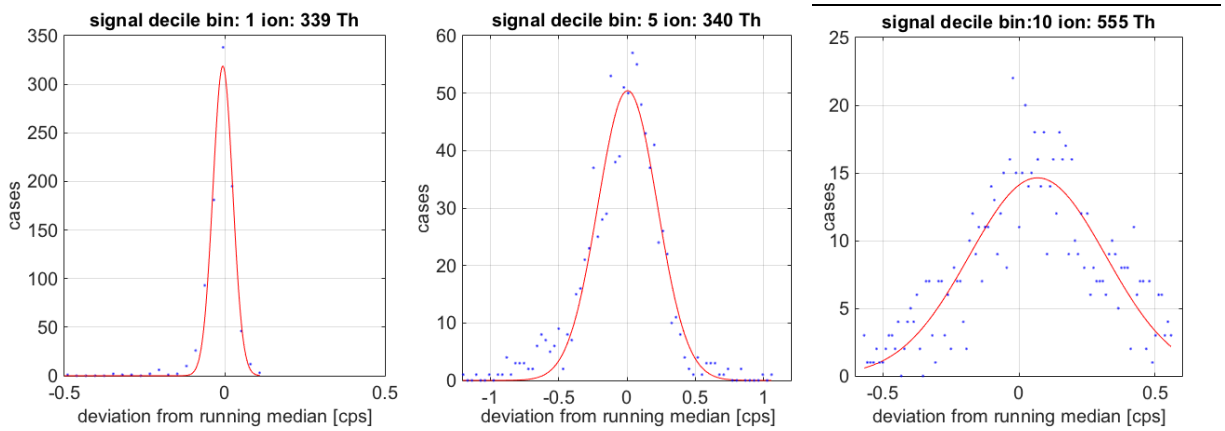


Fig. S7. Examples of a histogram of the deviations between ion signal and the five point moving median for ions at 339 (1st signal decile), 340 (5th decile) and 555 (10th decile) Th. The median points (difference = 0) are excluded. Also shown are the least squares Gaussian fits, from which the standard deviation σ (along with its uncertainty) is extracted.

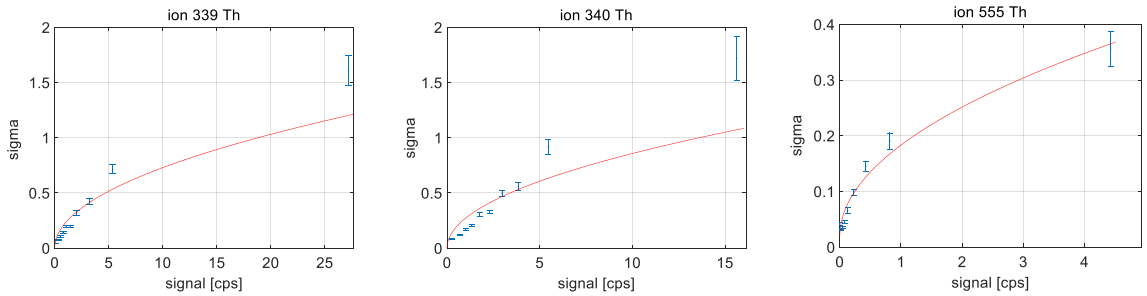


Fig. S8. The normal distribution (see Figure S7) standard deviations and their 95% confidence limits associated with the signal decile bins of ions at 339, 340 and 555 Th. The best (weighted non-linear least squares) fit for $c\sqrt{s} + e$ is shown in red, depicting our model for the signal dependence of the error.

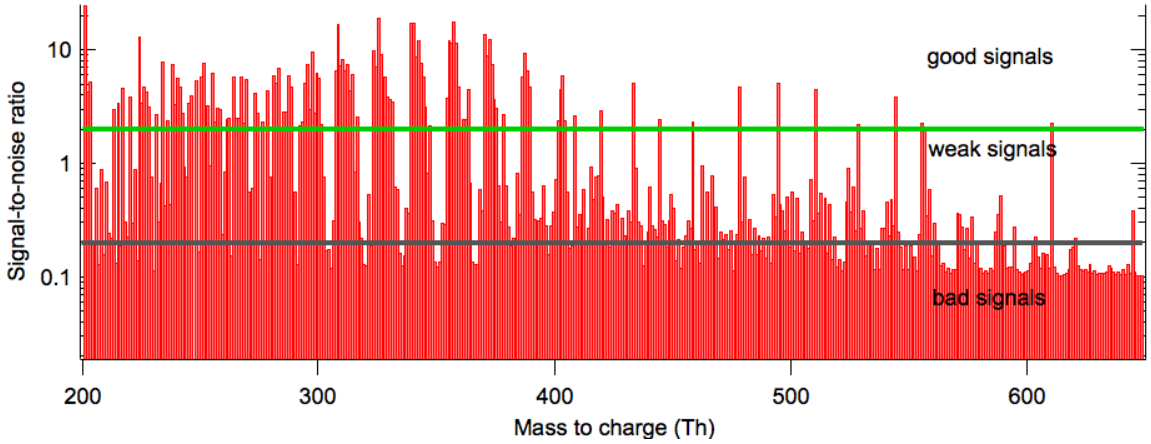


Fig. S9. Signal-to-noise ratio (SNR) of all variables (masses). SNR ≥ 2 is considered as "good signal", 2 > SNR ≥ 0.2 is considered "weak signal", and SNR < 0.2 is considered as "bad signal". In total, 173 signals are "weak", 152 signals are "bad", and the rest are good signals.

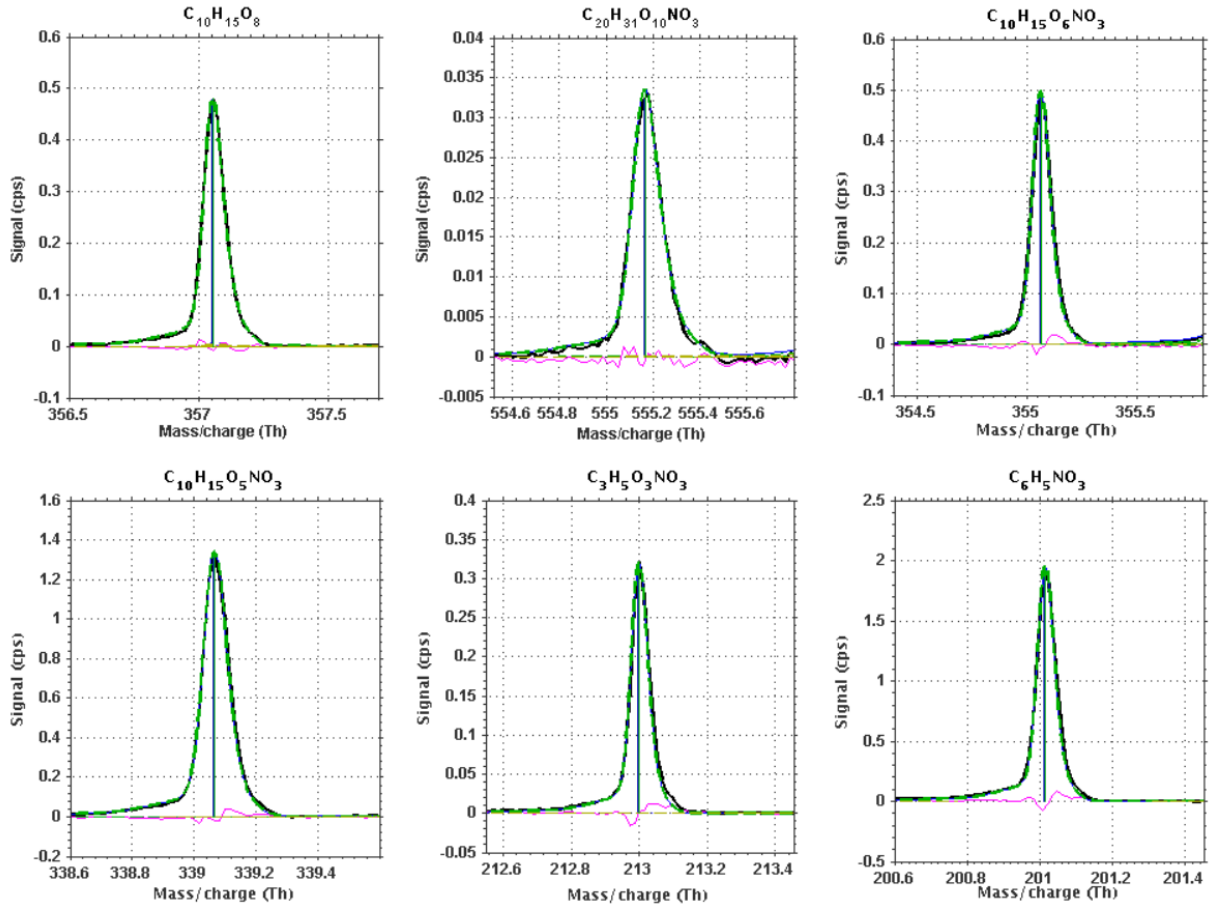


Fig. S10. Examples of peak fitting. The black solid line is the measured signal, the green dashed line denotes the fitted peak, and the purple one is the residual. The six examples correspond to the fingerprint molecules chosen from the 6 factors (marked with * in Table 1).

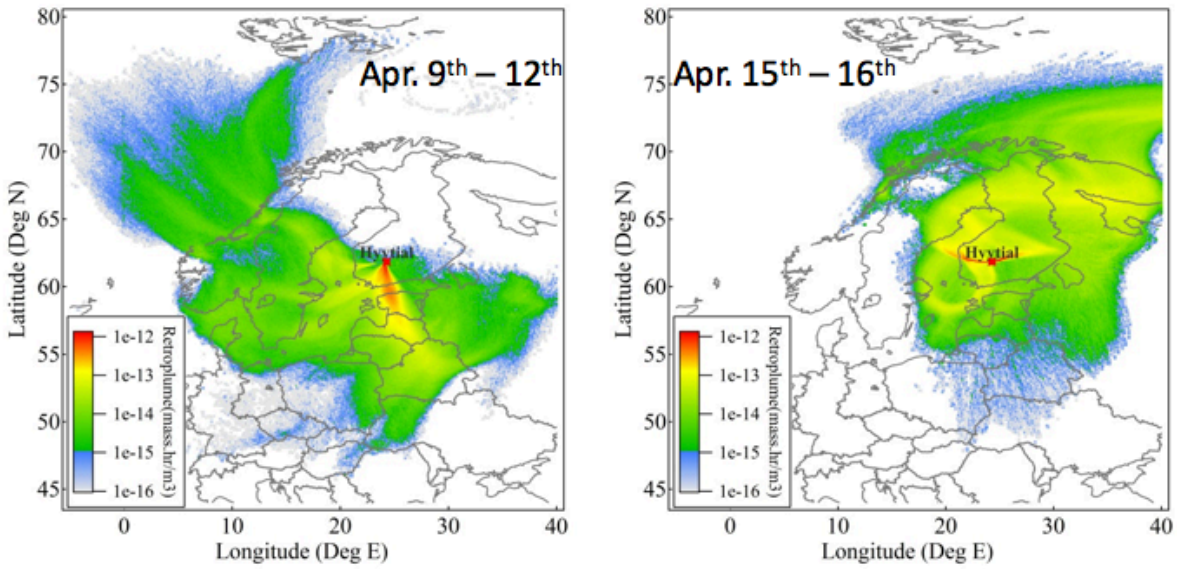


Fig. S11. Air mass analysis using backward Lagrangian particle dispersion model (LPDM). The shown results are based on 500m altitude calculation. The plot on the left shows that air masses were mainly from Eastern Europe on Apr. 9th – Apr.12th, while the plot on the right shows that air masses were from Northern Europe on most other days, for example Apr. 15th – Apr. 16th.

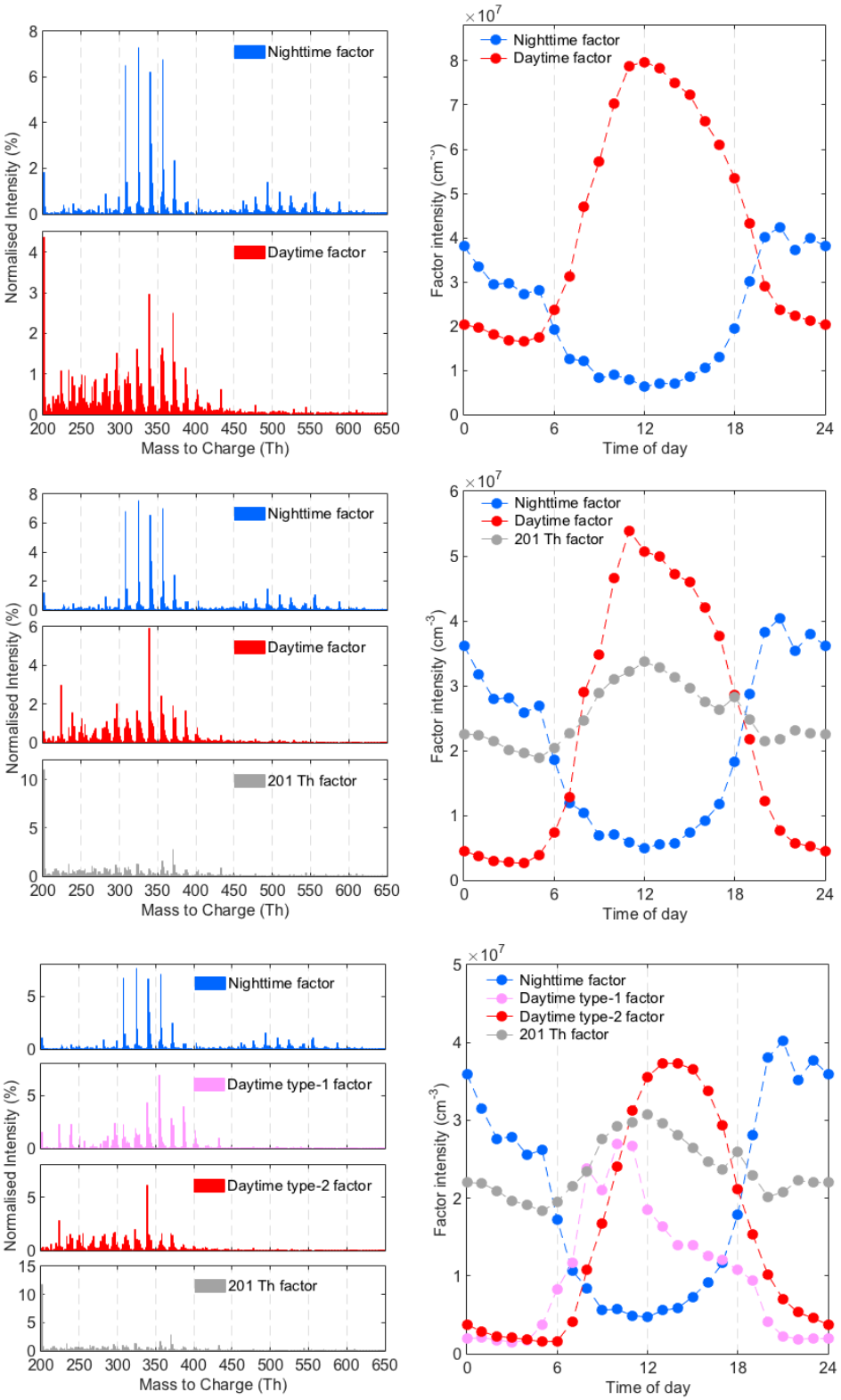
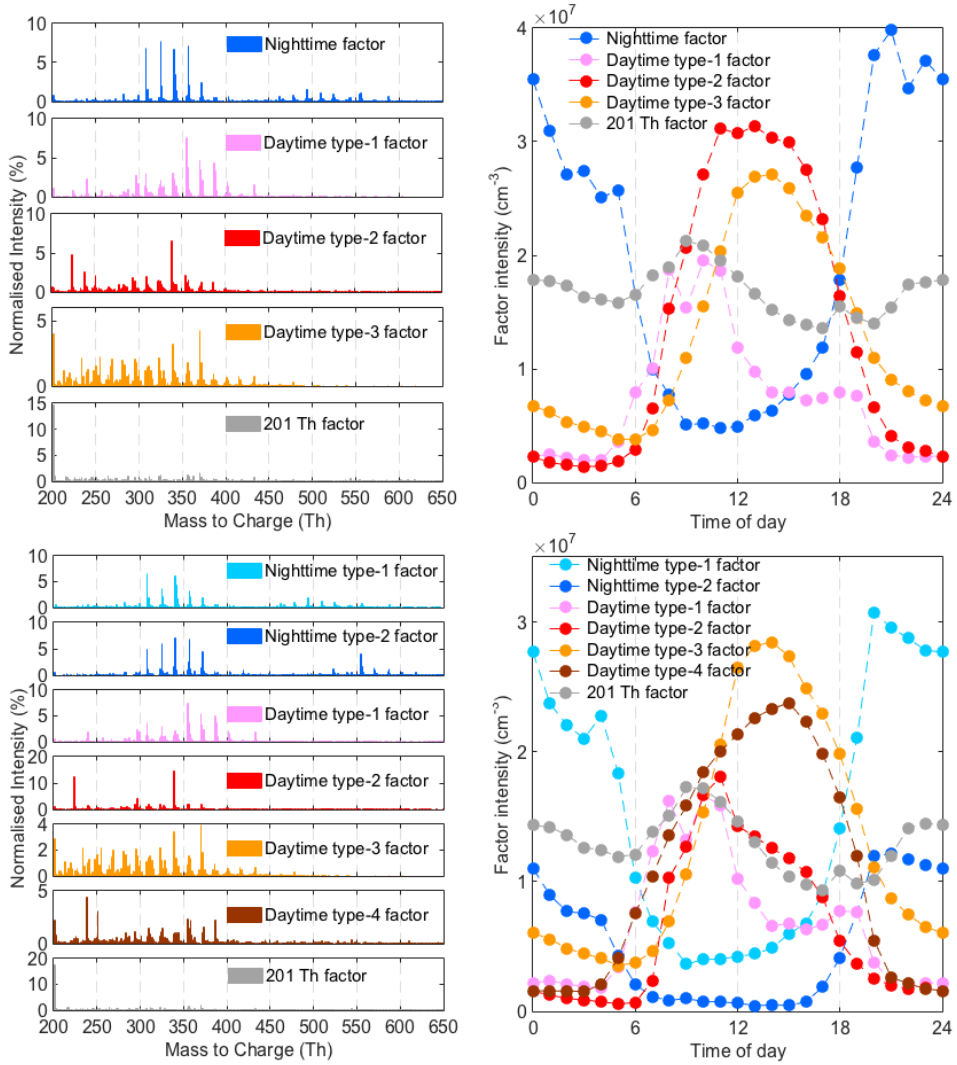


Fig. S12. Profile (left panels) and diurnal variation (right panels) of PMF factors. The top panels show the 2-factor case, the mid panels denote the 3-factor case, and the bottom panels demonstrate the 4-factor case.



148

149 **Fig. S12** (continued). Profile (left panels) and diurnal variation (right panels) of PMF factors. The top panels show
150 the 5-factor case, and the bottom panels demonstrate the 7-factor case. Note that the optimal solution with 6 factors
151 are shown in Fig. 5 and Fig. 6.

152

153