Source Finder Professional (SoFi Pro) allowing for time-dependent factor profiles and uncertainty assessment: application to one year of organic aerosol data

F. Canonaco[1,2], A. Tobler[2], G. Chen[2], Y. Sosedova[1], J. G. Slowik[2], C. Bozzetti[1,2], K. R. Daellenbach[3], I. ElHaddad[2], M. Crippa[4], R.-J. Huang[5], M. Furger[2], U. Baltensperger[2], A. S. H. Prévôt[2]

[1]Datalystica Ltd., Park innovAARE, CH-5234 Villigen, Switzerland
[2]Paul Scherrer Institute, Laboratory of Atmospheric Chemistry, CH-5232 Villigen PSI, Switzerland
[3]Institute for Atmospheric and Earth System Research, Helsinki, Finland
[4]European Commission, Joint Research Centre (JRC), Via Fermi, 2749, 21027 Ispra, Italy
[5]State Key Laboratory of Loess and Quaternary Geology, Center for Excellence in Quaternary Science and Global Change, and Key Laboratory of Aerosol Chemistry and Physics, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an 710061, China

A. Result of first approach

The settings of the rejected initial approach are described in Table S1. Performance of the PMF solutions was monitored using two quantities: Q/Q_{exp} as mathematical metric describing the goodness of fit, and the amount of $n_{\text{non-modelled}}$ revealing how many PMF runs do not fulfill the user-defined criteria. Note that the rolling window in the initial approach was not 14 days, as for the final approach in the manuscript, but rather 7 days. Hence, the comparison between the two strategies is only qualitative.

Table S 1. Overview of the rolling mechanism and the repeats of the PMF analysis for the rejected first approach.

Rolling mechanism
➢ a 7-day data window is defined
➢ the window is shifted day-wise over the measured data
PMF analysis
➢ for each window a four- (HOA, COA, BBOA, one OOA) and a five- (HOA, COA, BBOA, two OOAs) factor PMF run is performed, where HOA, COA and BBOA are constrained within the a-value approach
➢ x_{PMF} as free parameter is tested. Every repeat varied the unconstrained information randomly (seed), the a-values for the constrained factor profiles randomly and independently from a = 0 to a = 1 in a resolution of Δa = 0.1 (a-value exploration) and resampling the PMF input using the bootstrap method (resampling method)

Figure S1a) reports the Q/Q_{exp} and S1b) the amount of $n_{\text{non-modelled}}$ as a function of x_{PMF} for the rejected first approach. For a better comparison, the values from the second approach described in the main article are also reported on the right side. Between 50 and 400 repeats for the window-runs show no systematic decrease in Q/Q_{exp} as well as for the number of $n_{\text{non-modelled}}$. This might be due to an insufficient x_{PMF} to properly account for the a-value space (in total 1331 entries with a-values from 0 to 1 with a resolution of 0.1). As a consequence, the Q/Q_{exp} values scatter and the number of $n_{\text{non-modelled}}$ is rather high, with e.g., 50% for 400 repeats for window-run. For ≥700 PMF runs, the amount of $n_{\text{non-modelled}}$ as well as the factor-specific errors (consult Section 3.5 for the definition of the factor-specific error within this study) decrease. For 1500 repeats per window, i.e., a total of more than 1.5 million PMF runs taking 2-3 months of computational time, the number of $n_{\text{non-modelled}}$ is 10 times higher compared to the values from the finally accepted strategy (second approach). To fully account for the a-value space using the resampling and rolling strategy, one would require 1331 (amount of cells in the a-value space) x 100 (typically recommended for bootstrap runs) x 351 (days) x 2 (account for 4 and 5 factor solutions) = 93'436'200 PMF runs. Therefore, even for the highest repeat run conducted within the first approach, i.e., 1500 repeats, only a small fraction of these PMF runs were investigated. Increasing the amount of PMF runs is currently limited by the computational time of regular PCs, given that it took ~2.5 months to get the

results for the 1500 repeats per window-run. This makes this first approach rather inefficient for modern PCs and was therefore replaced by the second approach that led to the solution reported in the main article.

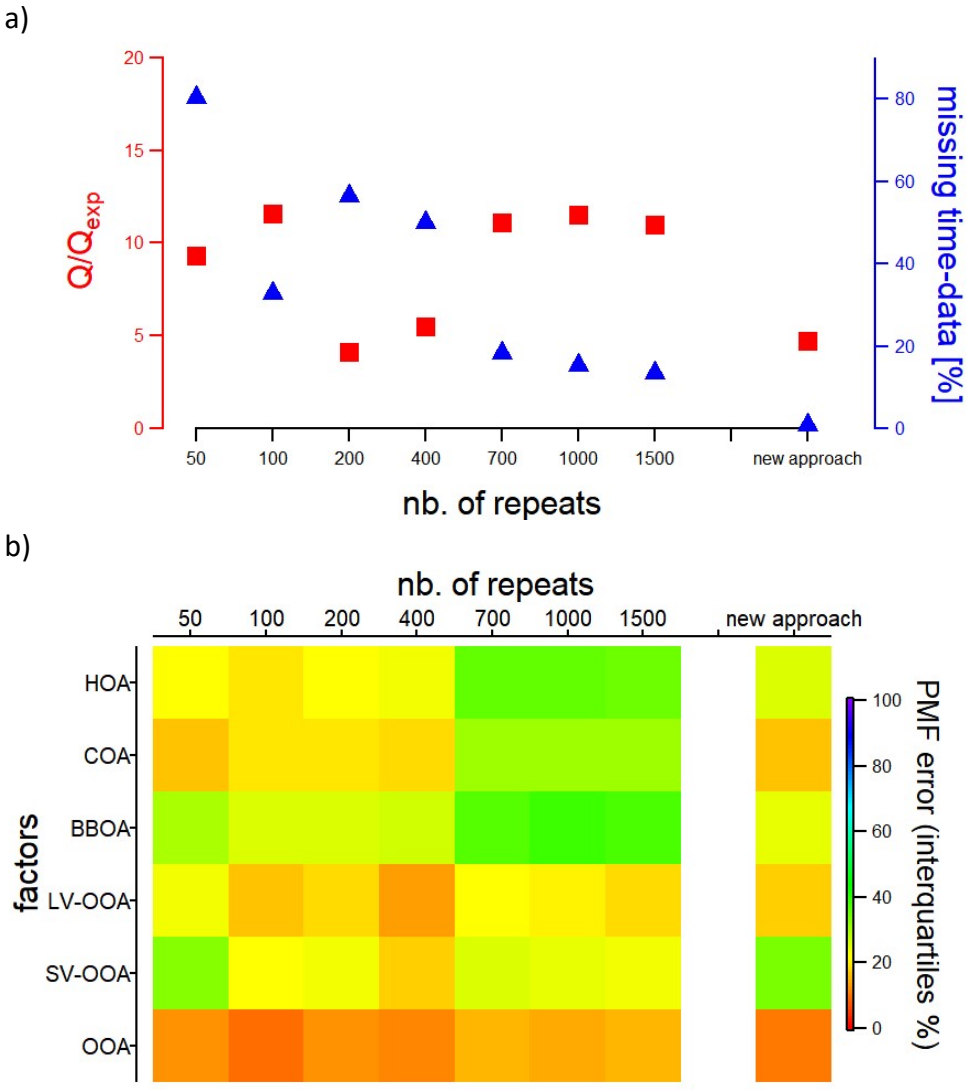


Figure S1. a) The mathematical metric Q/Q_{exp} and the amount of $n_{non\text{-}modelled}$ and b) the PMF error for all factors as a function of the x_{PMF} for the window-runs.

B. Factor time series

The weekly cycles of HOA, NO_x and eBC_{tr} tend to be higher during the weekdays and lower during the weekend. COA peaks on Saturdays, possibly driven by barbequing during the warm season and in winter due to increased outdoor meal consumptions, as all restaurants, pubs and coffees are open and most of the people in Zurich and surroundings have the day off. Other than OOA, the factors BBOA, LV-OOA and SV-OOA show good agreement with their tracers, eBC_{wb}, SO_4^{2-} and NO_3^-, respectively.

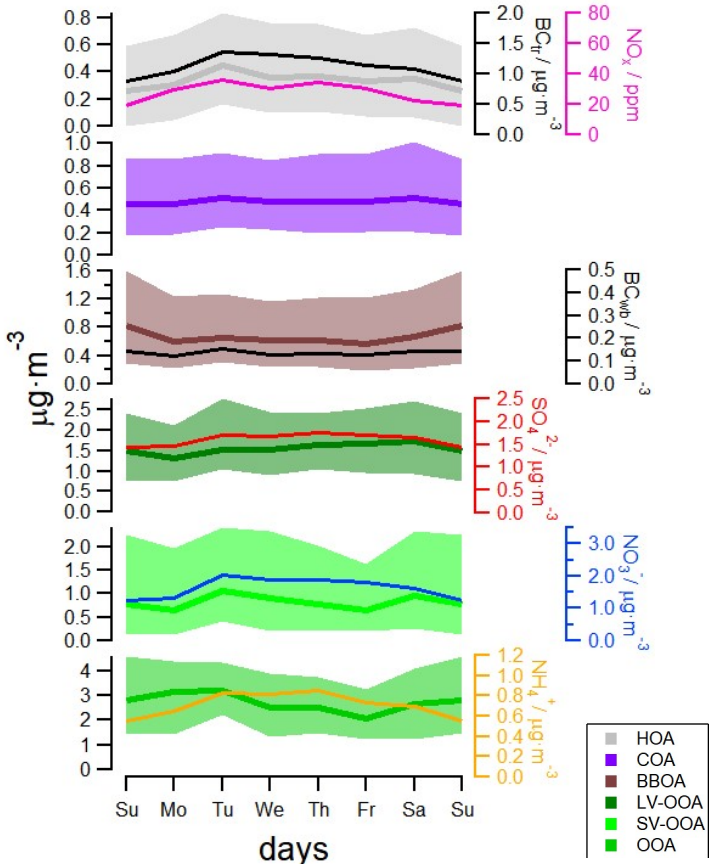
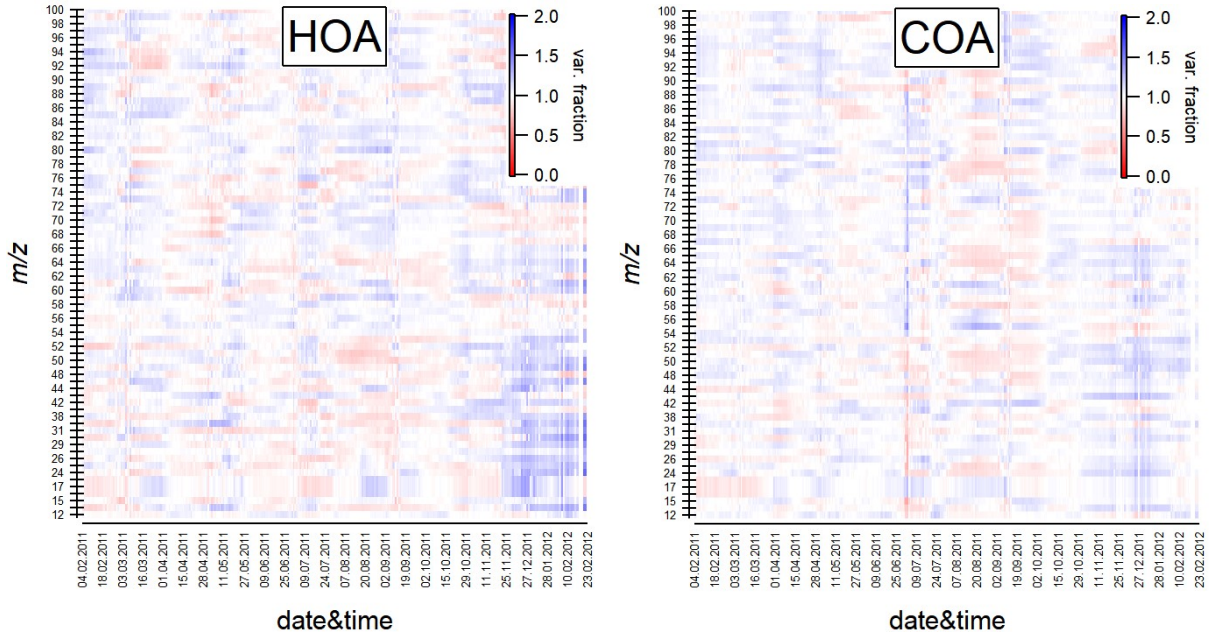


Figure S2. Weekly cycles of the factors for the entire period (February 2011 – February 2012). The thick lines represent the median and the shaded areas span the interquartile range. Typical external tracers are also shown for comparison, i.e., eBC_{tr} and NO_x for HOA, eBC_{wb} for BBOA, SO_4^{2-} for LV-OOA, NO_3^- for SV-OOA and NH_4^+ for OOA.

C. Time-dependent factor profiles

The temporal variation of the key variables for each factor presented in the main text in Section 3.3 is only a subsection of the entire mass spectrum. For completeness, the full mass spectra over time are reported in Figure S4. The image plots reveal that the oxygenated factors (LV-OOA, SV-OOA and OOA) undergo the strongest variation, as shown by the more intense red and blue colors. Among these factors, apart from the key variables already discussed in Section 3.3 many more variables show strong seasonal changes, expressed in complete color inversions (blue to red or *vice versa*), e.g., *m/z* 57, 61 or 73 for SV-OOA and LV-OOA. The image plots in Figure S4 reveal strong temporal changes of the factor profiles and underline the importance of the rolling approach for long-term SA analysis.

a) **b)**



c) **d)**

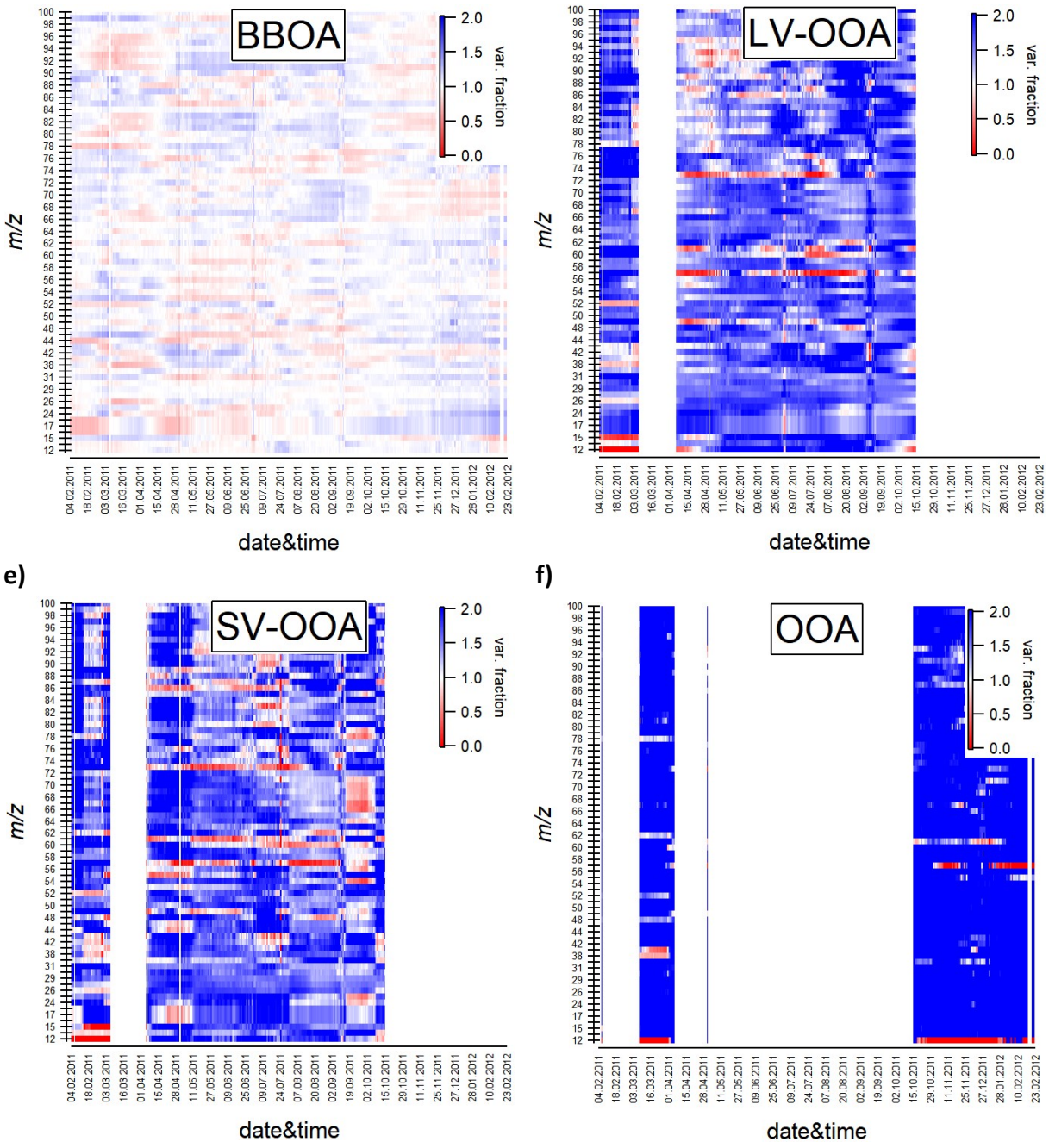


Figure S3. The mass spectra for the six factors as a function of time. Each variable is normalized by its mean to better stress its temporal variation. The color-code is truncated at 2, and for higher values the dark blue color is kept constant giving less weight to few transient events.

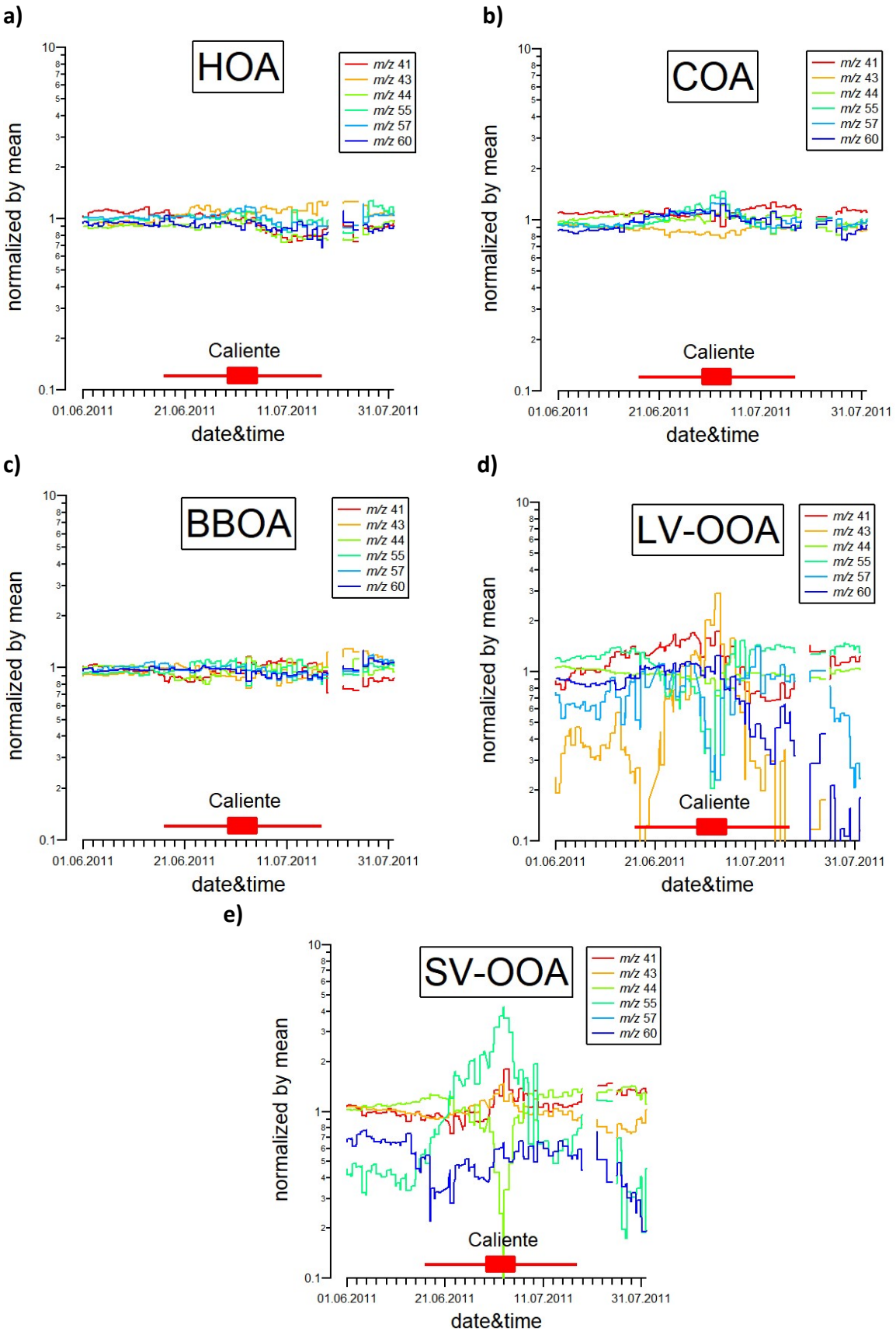


Figure S4. Daily averaged fractions of important AMS/ACSM variables for the Caliente episode. Each variable is normalized by its mean to better stress its temporal variation. The Caliente episode is represented by the red rectangle, whereas the line depicts the possible Caliente influence on the neighboring points due the 14-days rolling window.

D. *a*-value selection

Figure S5 contains the daily averaged *a*-values for the constrained factor profiles. The a-values for HOA, COA and BBOA are on average 18.5 %, 14.0 % and 21.8 %, respectively. The 90th percentile equals the prescribed upper limit of 0.4 more often for BBOA and less frequently for HOA and COA, suggesting that on average the HOA and COA profiles are less limited by the constrained *a*-values than BBOA.

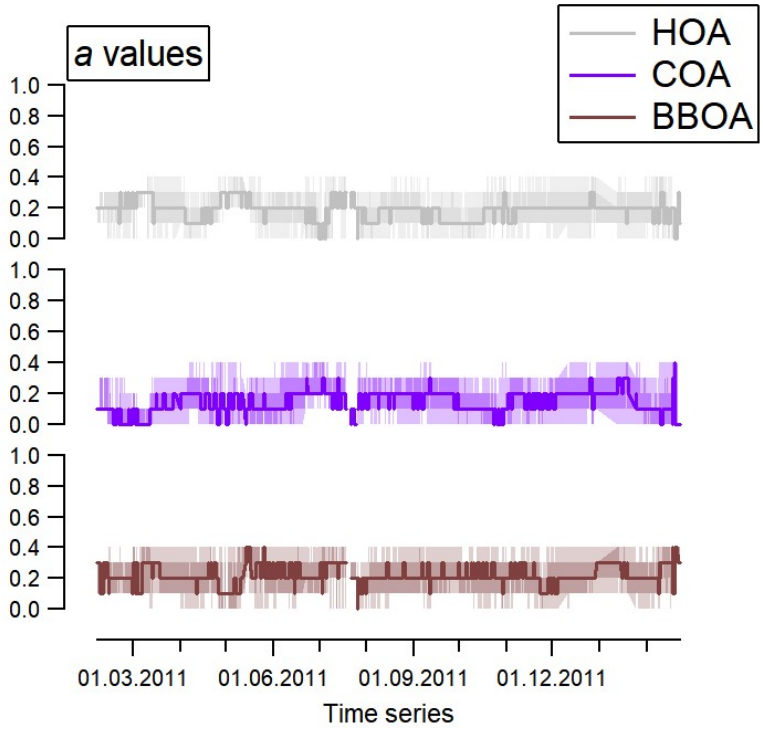


Figure S5. Daily averaged median *a*-values for the constrained factor profiles. The lines represent the medians, the dark shaded areas span the interquartile ranges, whereas the light shaded areas span the ranges from 10th to 90th %.