SUPPLEMENTAL INFORMATION

for the paper

Interpretation and Sensitivity Analysis of Organic Components Extracted by Positive Matrix Factorization of Aerosol Mass Spectrometric Data

I. M. Ulbrich,[1,2] M. Canagaratna,[3] Q. Zhang,[4] D. Worsnop,[3] and J. L. Jimenez[1,2,*]

[1]Cooperative Institute for Research in the Environmental Sciences (CIRES), and [2]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO, 80309

[3]Aerodyne Research, Inc., Billerica, MA, 01821-3976

[4]Atmospheric Sciences Research Center, University at Albany, State University of New York, Albany, NY 12203

[*]Corresponding author: email jose.jimenez@colorado.edu

Weighting Factors for *m/z* 44-Related Peaks

The organic mass spectrum measured by the Q-AMS is estimated from the full measured mass spectrum but application of a "fragmentation table" (Allan et al., 2004). Several of the *m/z*'s in the AMS organic spectrum are related to *m/z* 44 by proportional constants in the fragmentation table such that the signals, x, are related by

$$x_{\mathrm{m/z}\ 18} = x_{\mathrm{m/z}\ 44} \qquad \text{(SI-1)}$$

$$x_{\mathrm{m/z}\ 17} = 0.27\ x_{\mathrm{m/z}\ 18} = 0.27\ x_{\mathrm{m/z}\ 44} \qquad \text{(SI-2)}$$

$$x_{\mathrm{m/z}\ 16} = 0.04\ x_{\mathrm{m/z}\ 18} = 0.04\ x_{\mathrm{m/z}\ 44} \qquad \text{(SI-3)}$$

and likewise their errors have the same relationships:

$$\sigma_{\mathrm{m/z}\ 18} = \sigma_{\mathrm{m/z}\ 44} \qquad \text{(SI-4)}$$

$$\sigma_{\mathrm{m/z}\ 17} = 0.27\ \sigma_{\mathrm{m/z}\ 44} \qquad \text{(SI-5)}$$

$$\sigma_{\mathrm{m/z}\ 16} = 0.04\ \sigma_{\mathrm{m/z}\ 44} \qquad \text{(SI-6)}$$

Including all of these *m/z*'s gives duplicates the signal at *m/z* 44 and gives excessive weight to this large signal in the PMF fit. The group of duplicate peaks (*m/z*'s 16, 17, and 18) could be omitted from the PMF input and then reconstructed afterwards, but it would be more convenient to include them in the fit but weigh them so that the Q-contribution of the group is the same as the Q-contribution if only *m/z* 44 had been included.

In order to satisfy

$$Q_{m/z\ 44\ \mathrm{only}} = Q_{\mathrm{all}\ m/z\ 44\text{-related ions}} \qquad \text{(SI-7)}$$

assume that the errors when all *m/z* 44-related peaks are included must be multiplied by a factor a such that

$$\sum_{j=1}^{m}\left(\frac{x_{m/z44}}{\sigma_{m/z44}}\right)^2 = \sum_{j=1}^{m}\left(\frac{x_{m/z44}}{a*\sigma_{m/z44}}\right)^2 + \sum_{j=1}^{m}\left(\frac{x_{m/z18}}{a*\sigma_{m/z18}}\right)^2 + \sum_{j=1}^{m}\left(\frac{x_{m/z17}}{a*\sigma_{m/z17}}\right)^2 + \sum_{j=1}^{m}\left(\frac{x_{m/z16}}{a*\sigma_{m/z16}}\right)^2 \qquad \text{(SI-8)}$$

Substituting Eqs. (SI-1) to (SI-6),

$$\sum_{j=1}^{m}\left(\frac{x_{m/z44}}{\sigma_{m/z44}}\right)^2 = \sum_{j=1}^{m}\left(\frac{x_{m/z44}}{a*\sigma_{m/z44}}\right)^2 + \sum_{j=1}^{m}\left(\frac{x_{m/z44}}{a*\sigma_{m/z44}}\right)^2 + \sum_{j=1}^{m}\left(\frac{0.27x_{m/z44}}{a*0.27\sigma_{m/z44}}\right)^2 + \sum_{j=1}^{m}\left(\frac{0.04x_{m/z44}}{a*0.04\sigma_{m/z44}}\right)^2 \quad \text{(SI-9)}$$

all of the proportionality constants cancel, and the equation can be reduced to

$$\sum_{j=1}^{m}\left(\frac{x_{m/z44}}{\sigma_{m/z44}}\right)^2 = \frac{4}{a^2}\sum_{j=1}^{m}\left(\frac{x_{m/z44}}{\sigma_{m/z44}}\right)^2 \quad \text{(SI-10)}$$

and $a = 2$ (i.e., a = the square root of the number of *m/z* 44-related ions).

Analyses of HR-ToF-AMS data (Aiken et al., 2008) have led to revisions of the fragmentation table such that *m/z* 28 is now assigned organic mass. When the revised fragmentation table is used, the scaling factor a for the error should be sqrt (5) = 2.24 because there are five ions (*m/z*'s 44, 28, 18, 17, 16) related to *m/z* 44.

Calculation of the Angle between a Vector and a Plane

The factor MS can be considered as vectors in a 270-dimensional space. Any two vectors in this space define a plane. Here we consider the plane defined by the HOA and OOA-1 MS vectors. Any 3rd vector in that 270-dimensional space will have an angle with respect to that plane. Any vector which is the linear combination of HOA and OOA-1 would lie exactly in that plane and have an angle of zero with respect to the plane. Any vector not in the HOA-OOA-1 plane is not a linear combination of HOA and OOA-1, and its "distance" from the HOA-OOA-1 plane can be quantified by the angle α between the plane and the vector.

In linear algebra terms, let **A** be a 270 x 2 matrix formed by combining the non-orthogonal, non-collinear column vectors ***HOA*** and ***OOA-1*** such that

$$\mathbf{A} = (\boldsymbol{HOA}, \boldsymbol{OOA\text{-}1}) \qquad \text{(SI-11)}$$

and rank(**A**) = 2. A projection matrix P may be constructed by

$$\mathbf{P} = \mathbf{A}(\mathbf{A}^{T}\mathbf{A})^{-1}\mathbf{A}^{T} \qquad \text{(SI-12)}$$

which gives the projection (***w***) of a third vector (***v***, which is OOA-2 in the quoted example) onto the plane by

$$\mathbf{P}\,\boldsymbol{v} = \boldsymbol{w} \qquad \text{(SI-13)}$$

A right triangle between ***v*** and the HOA-OOA-1 plane can be identified in which ***w*** is the base (and lies in the HOA-OOA-1 plane), ***v*** is the hypotenuse, and a vector ***z*** could be drawn perpendicular to the plane to define the height (Olver and Shakiban, 2006). The cosine of the angle α between ***V*** and the plane can be defined by

$$\cos\alpha = \frac{\|w\|}{\|v\|} \qquad \text{(SI-14)}$$

and α can be calculated with the inverse cosine function.

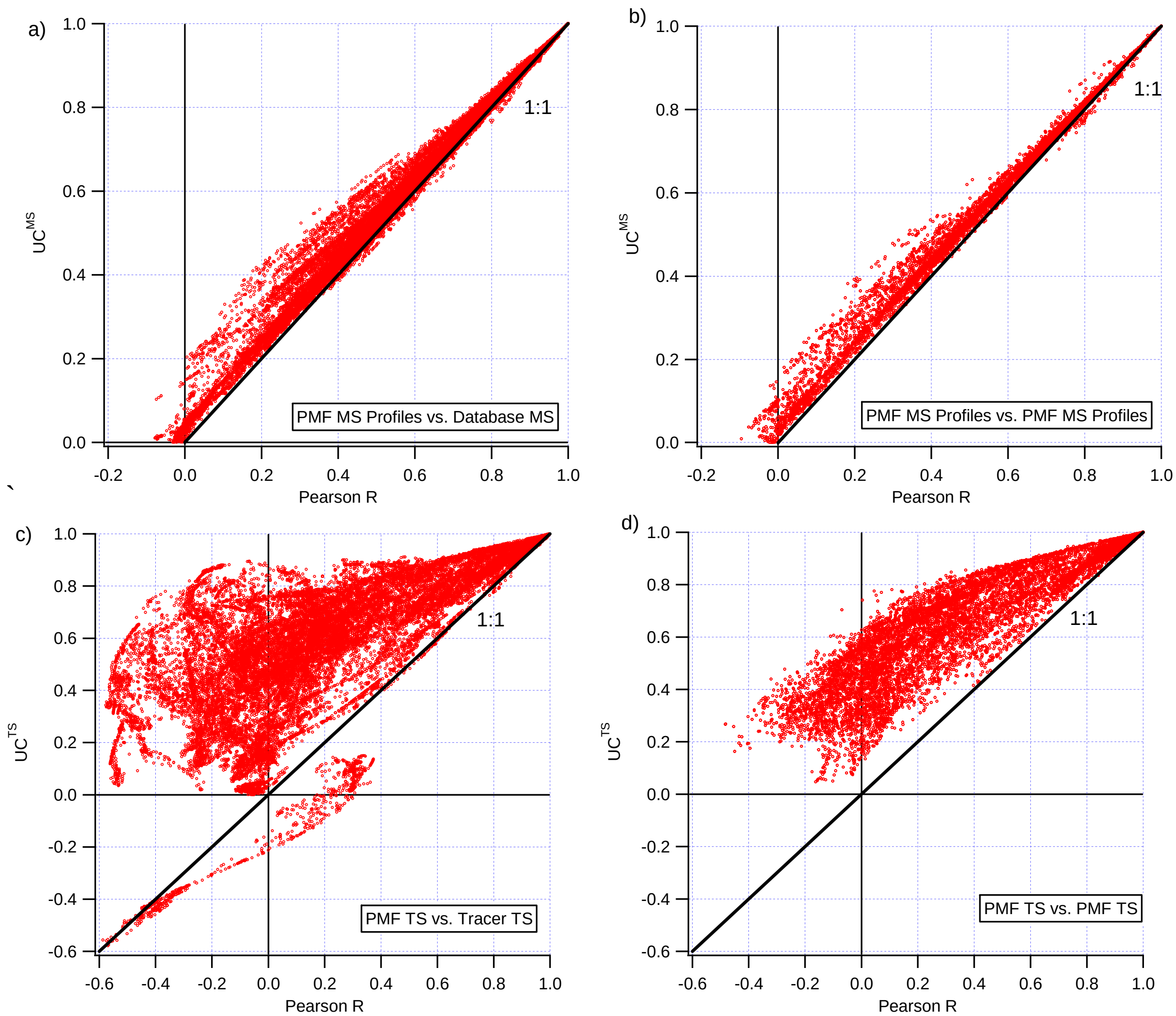

Figure S1. Comparisons of the uncentered correlation with Pearson R. All PMF factors are from solutions of the real Pittsburgh case with solutions from 1 to 7 factors at FPEAKS from -3.0 to +3.0. Sixty MS from the Mass Spectral Database are used for MS comparisons; 34 tracer TS are used for TS comparsions. a) MS from PMF solutions vs. database MS; b) MS from PMF solutions vs. MS from PMF solutions; c) TS from PMF solutions vs. tracer TS; d) TS from PMF solutions vs. TS from PMF solutions.

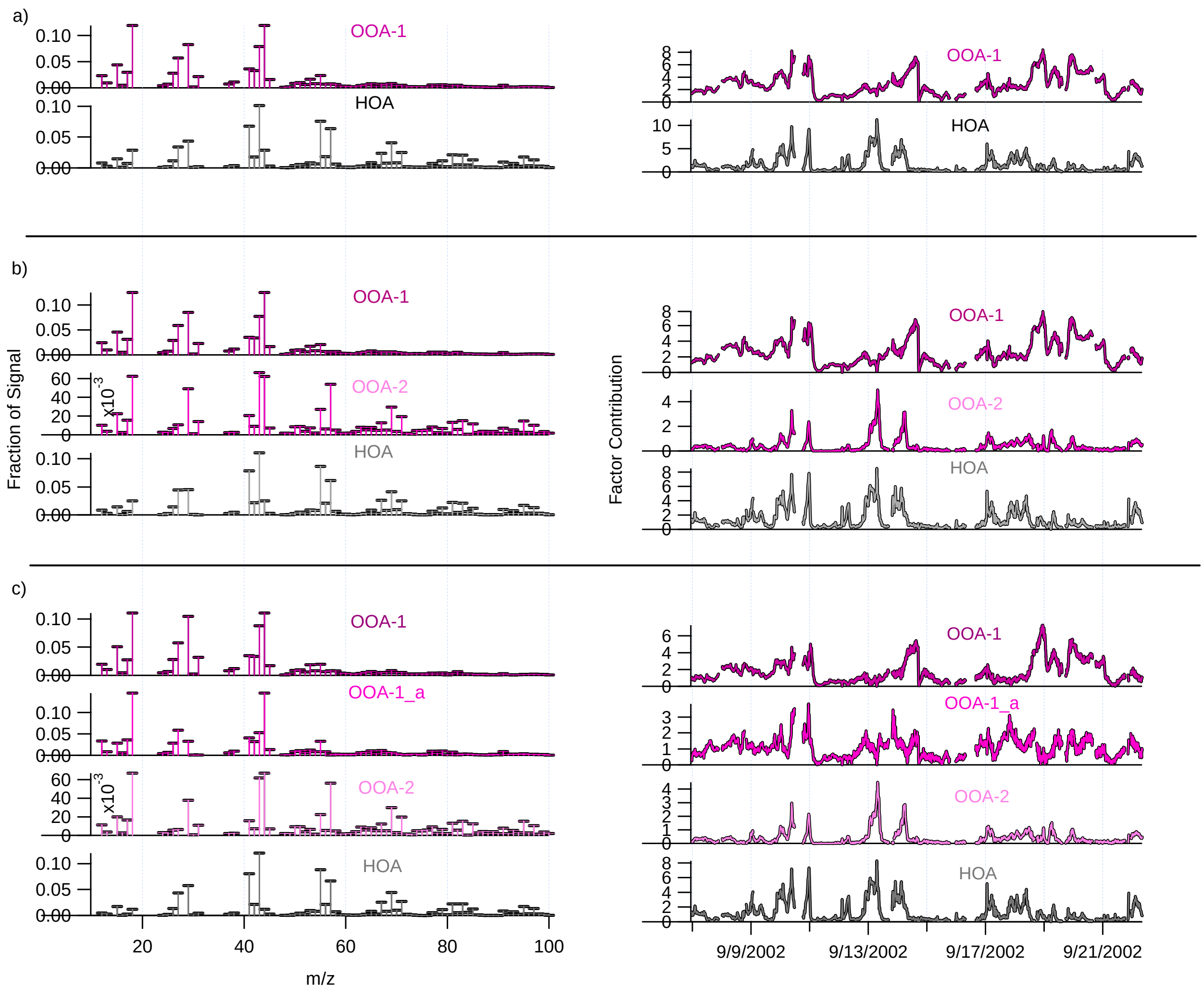

Figure S2. Solutions of the real Pittsburgh dataset in robust mode (thicker black markers) and the non-robust mode (thinner colored lines and markers, in front) with a) 2 factors, b) 3 factors, and c) 4 factors.

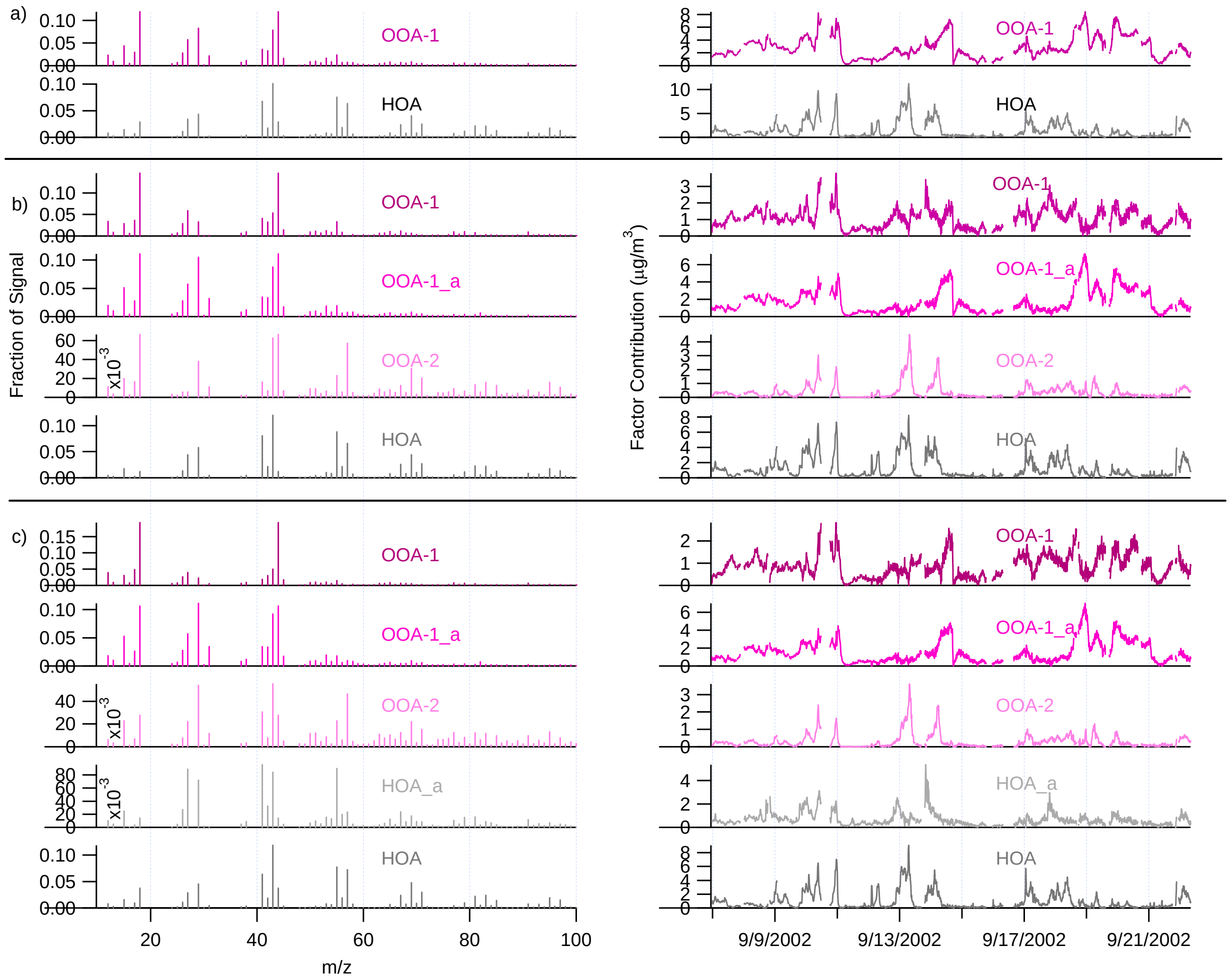

Figure S3. PMF solutions of the real Pittsburgh case with a) 2 factors, b) 4 factors, and c) 5 factors.

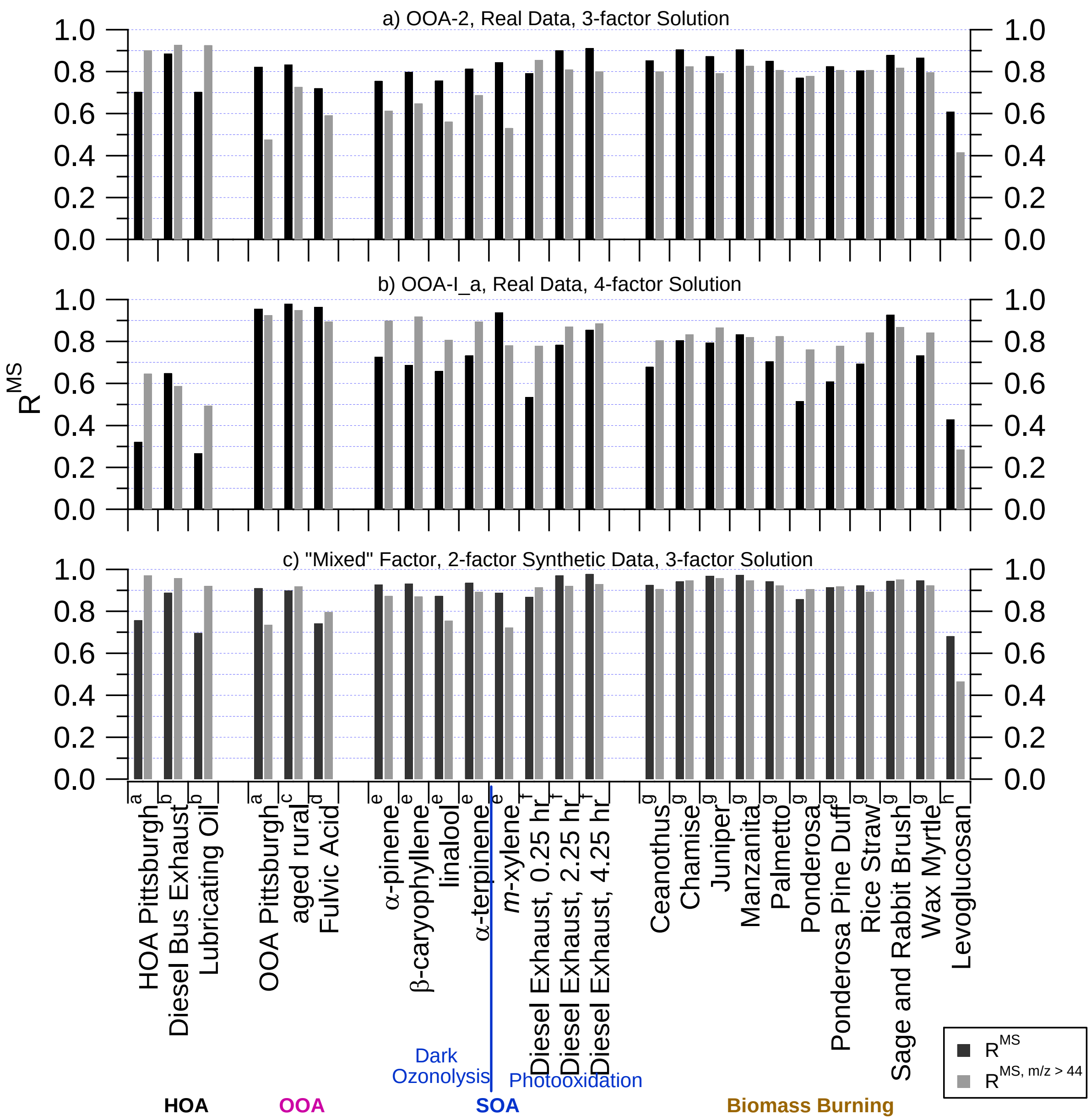

Figure S4. R^{MS} between representative spectra from the AMS Mass Spectral Database (http://cires.colorado.edu/jimenez-group/AMSsd) and a) the third factor mass spectrum from the 3-factor PMF solution of the real Pittsburgh dataset, b) the fourth factor mass spectrum from the 4-factor PMF solution of the real Pittsburgh dataset, and c) the “mixed” factor mass spectrum from the 3-factor PMF solution of 2-factor base case. Values are given in Table S1. Superscripts denote the source of the reference spectra as follows: (a) Zhang et al., 2005; b) Canagaratna et al., 2004; c) Alfarra et al., 2004; d) Alfarra, 2004; e) Bahreini et al., 2005; f) Sage et al., 2007; g) I.M. Ulbrich, J. Kroll, J.A. Huffman, T. Onash, A. Trimborn, J.L. Jimenez, unpublished spectra, FLAME-I, Missoula, MT, 2006; h) Schneider et al., 2006)

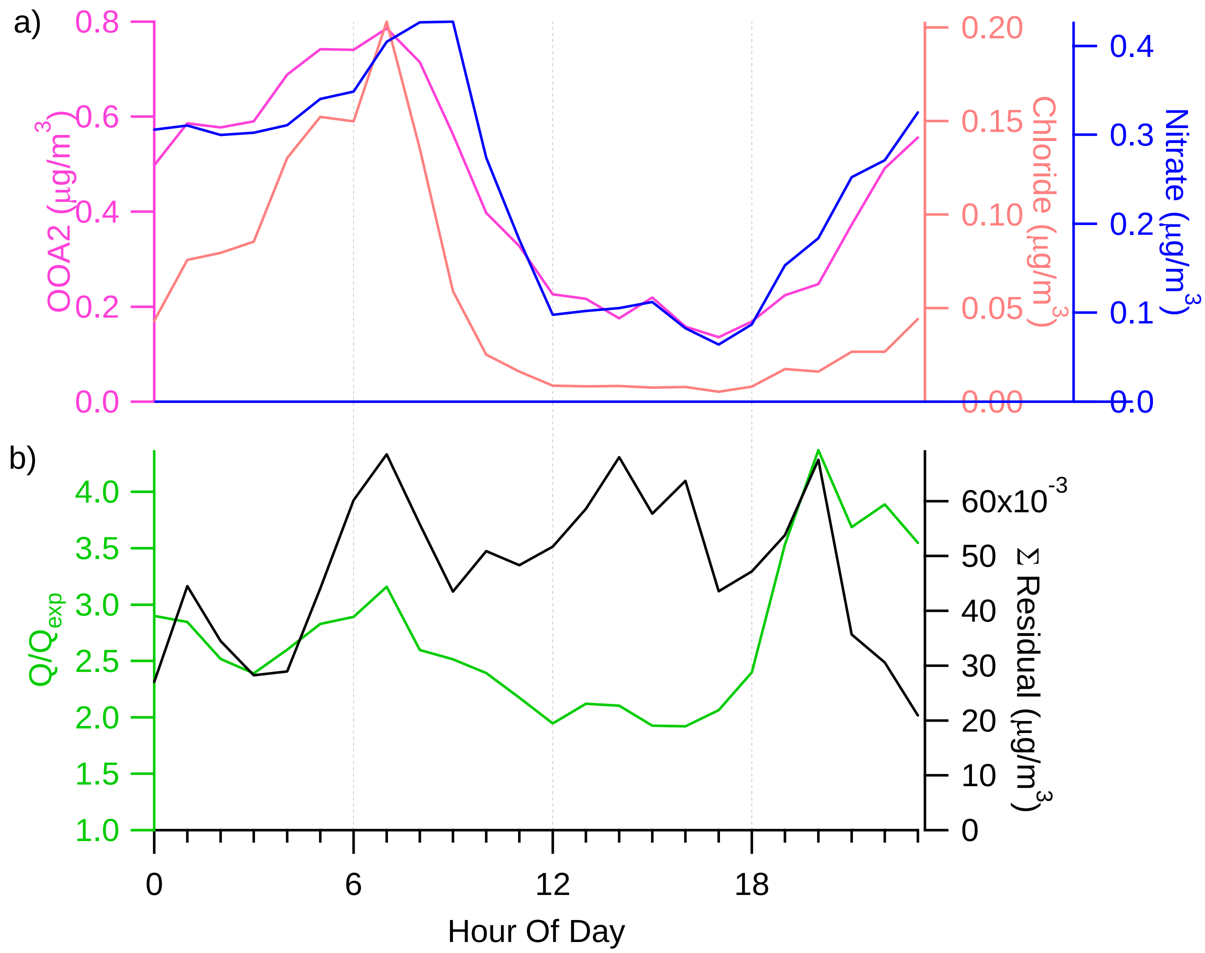

Figure S5. Diurnal profiles of a) OOA-2, Nitrate, and Chloride, and b) Q/Q_{exp} and the total residual from the 3-factor solution of the real Pittsburgh dataset.

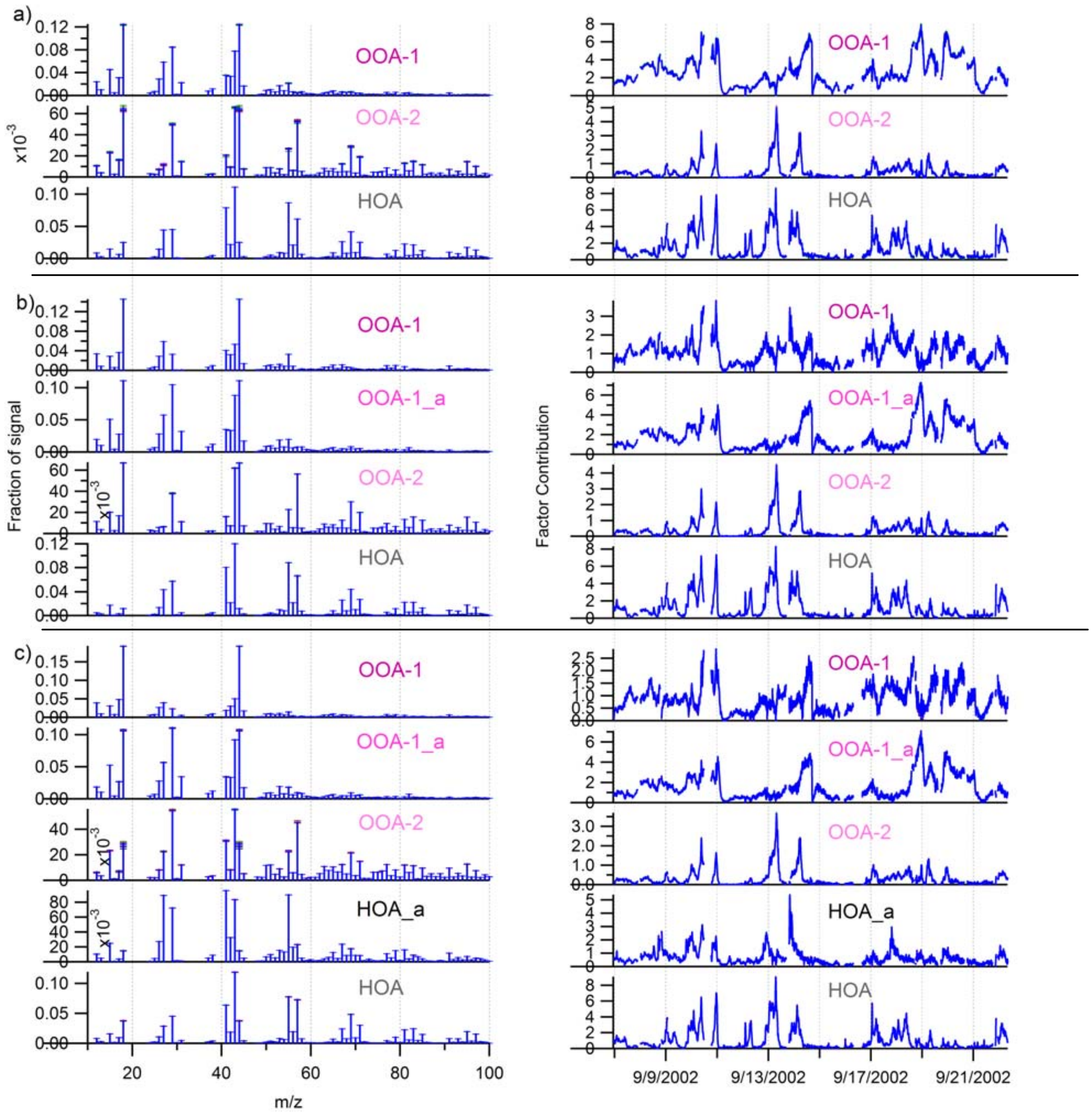

Figure S6. Solutions of the real Pittsburgh dataset from 64 random starts computing a) 3 factors, b) 4 factors, and c) 5 factors.

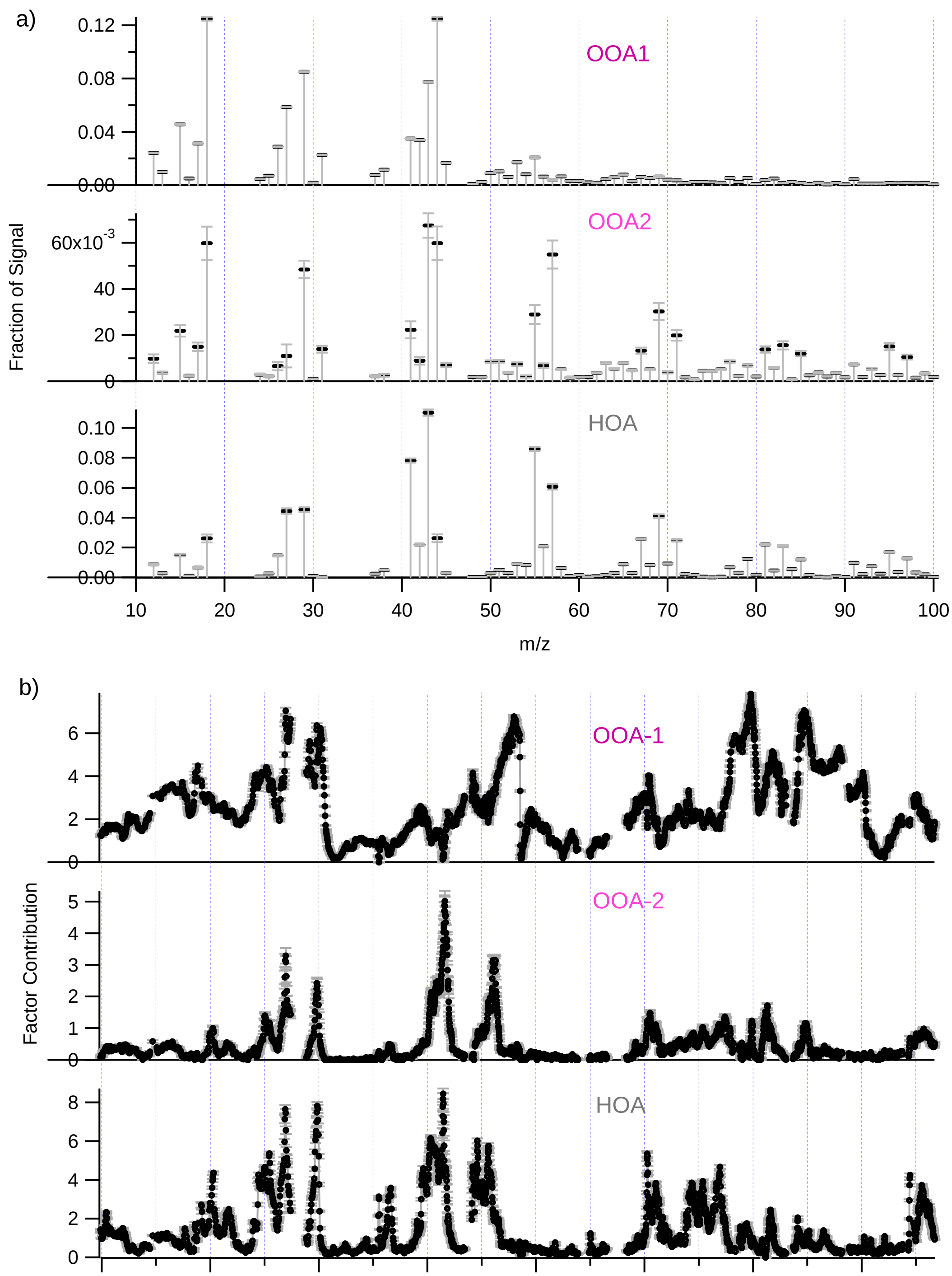

Figure S7. Results from bootstrapping analysis of the 3-factor solution of the real Pittsburgh case. Average (black) with 1-σ error bars (grey) are shown for factor a) MS and b) TS. The solutions from multiple FPEAKS (Fig. 10) show a greater range in MS than the 1-σ variation bars, while the TS show a similar range to the 1-σ variation bars.

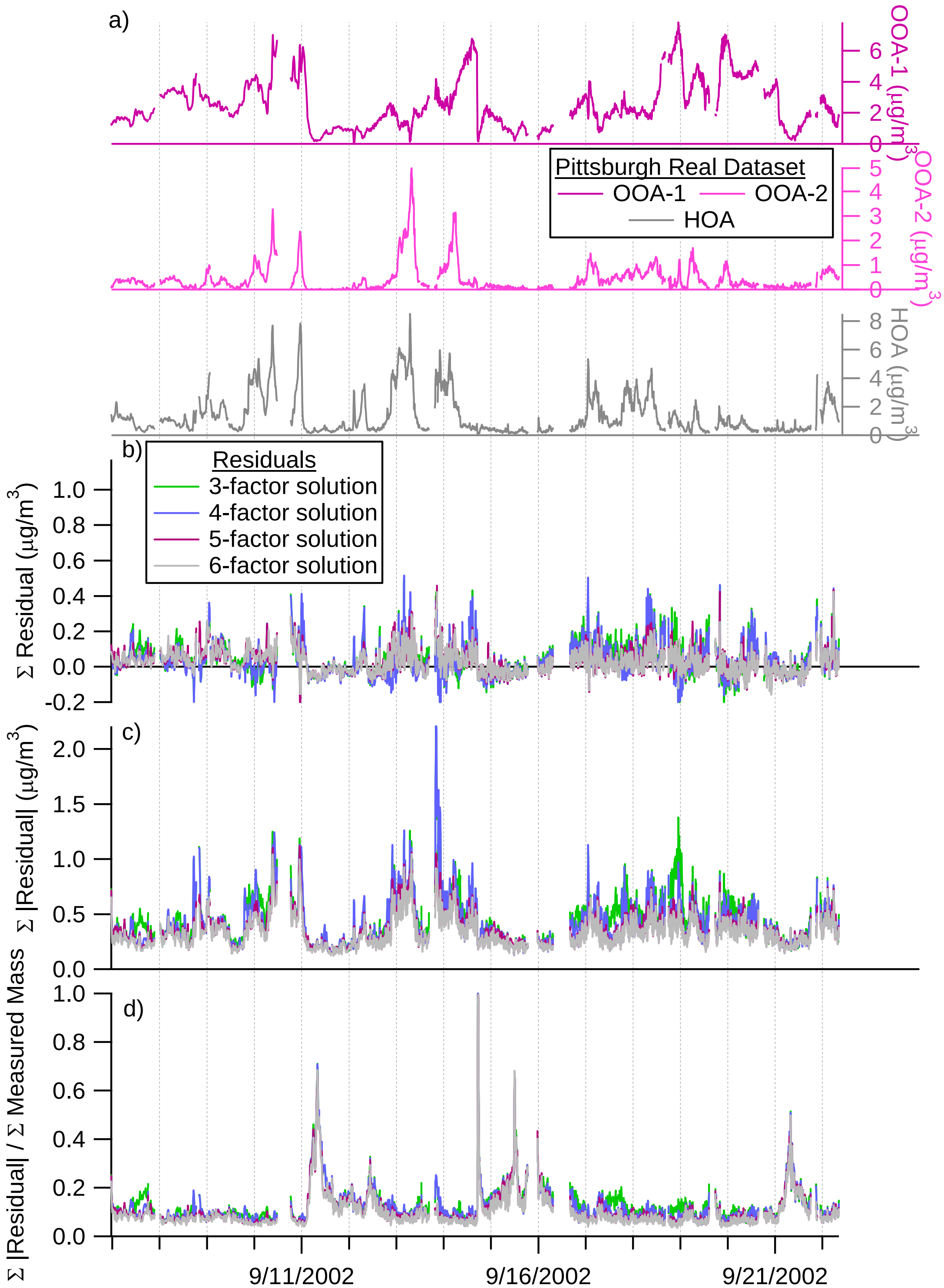

Figure S8. a) OOA-1, OOA-2, and HOA time series from the 3-factor solution of the real Pittsburgh dataset. b) Total Residual, c) Total absolute residual, and d) absolute residual normalized by total signal for the 3- to 6-factor solutions of the real Pittsburgh case.

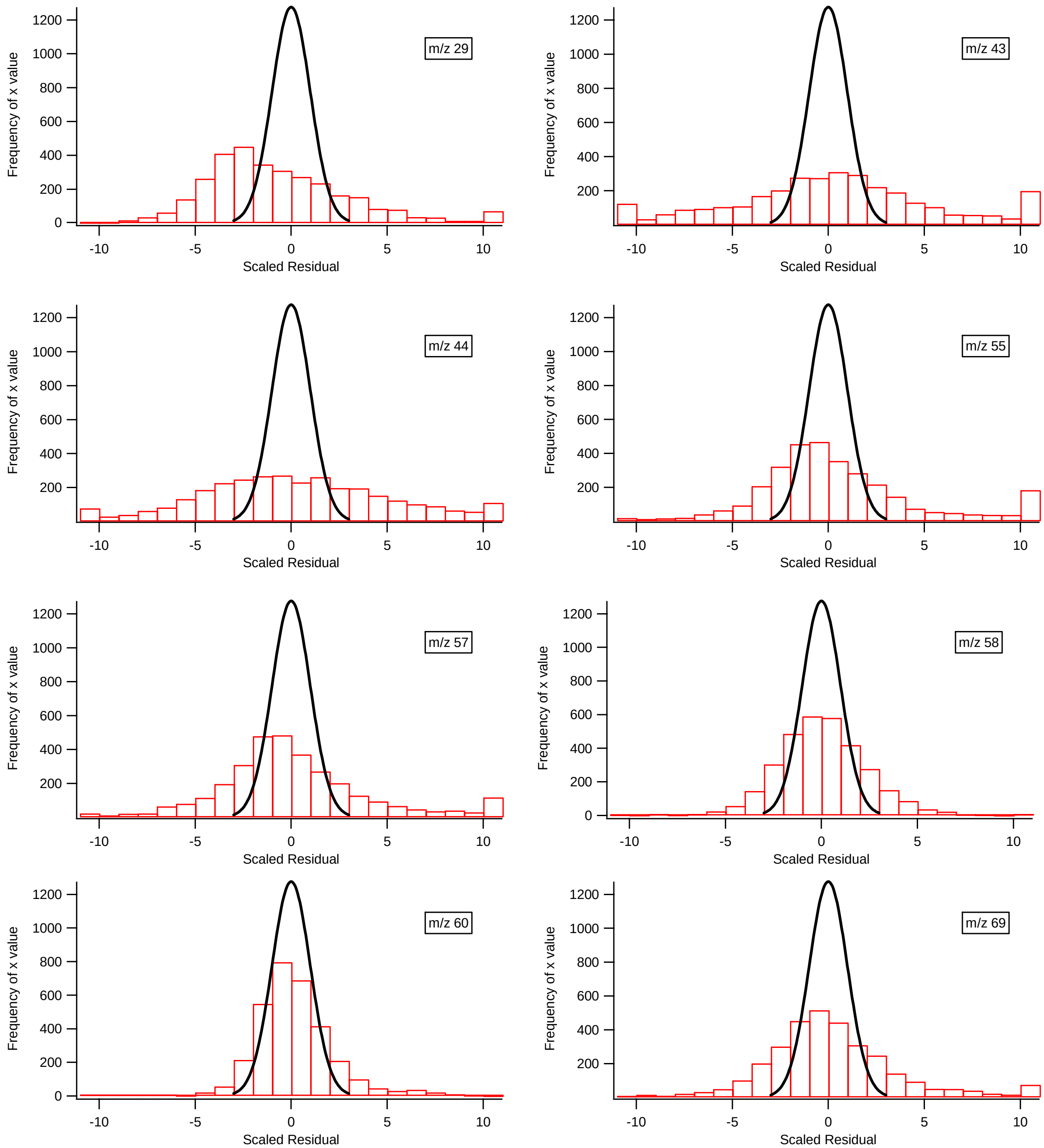

Figure S9. Histograms (red) of scaled residuals of selected m/z's from the 3-factor solution of the real Pittsburgh case. The first bin includes all scaled residuals < -10; the last bin includes all scaled residuals > 10. A Gaussian distribution (mean = 0, σ = 1, multiplied by 3199 to have the same area as the histogram) is shown in black to guide the eye.

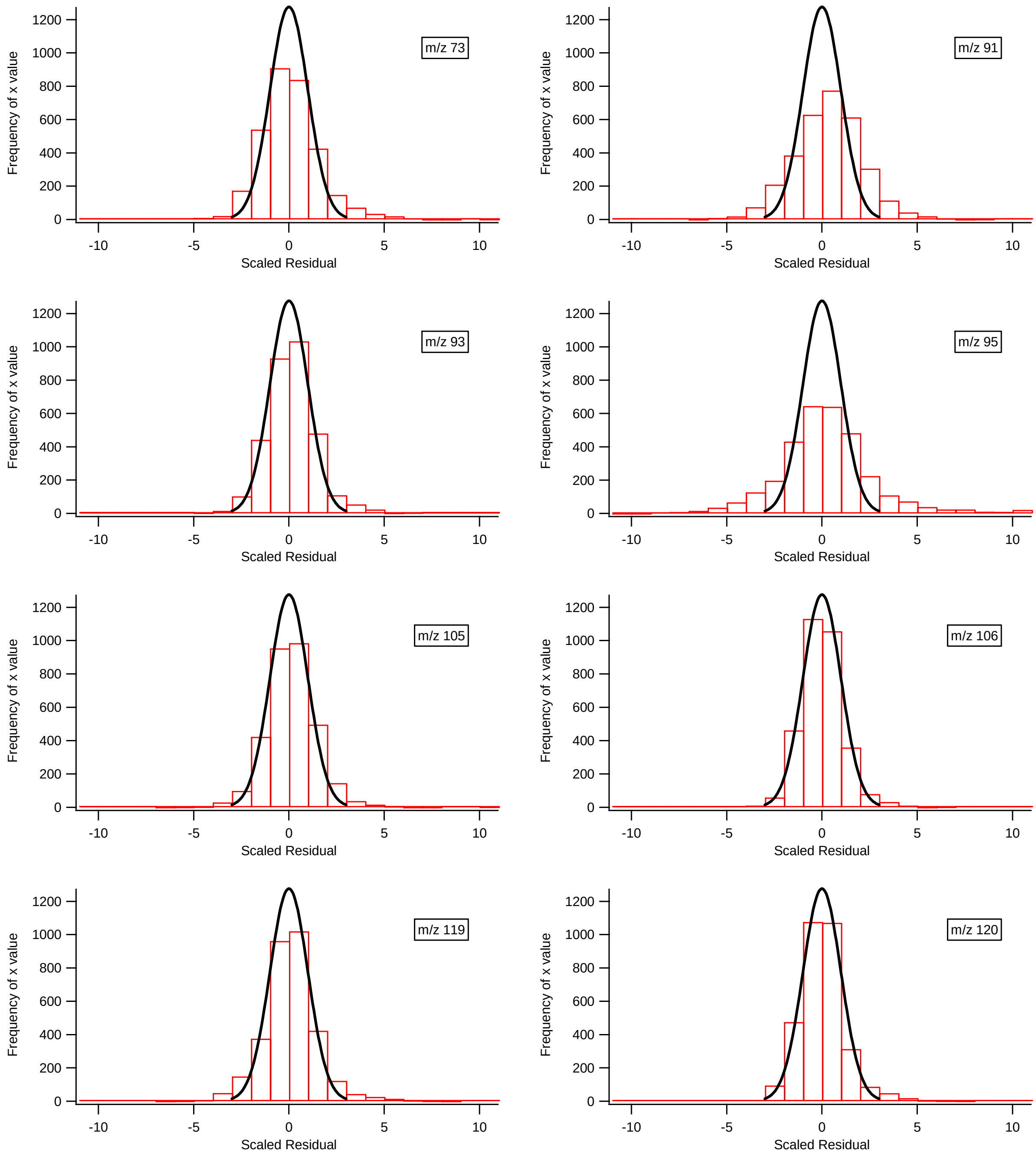

Figure S9 cont.

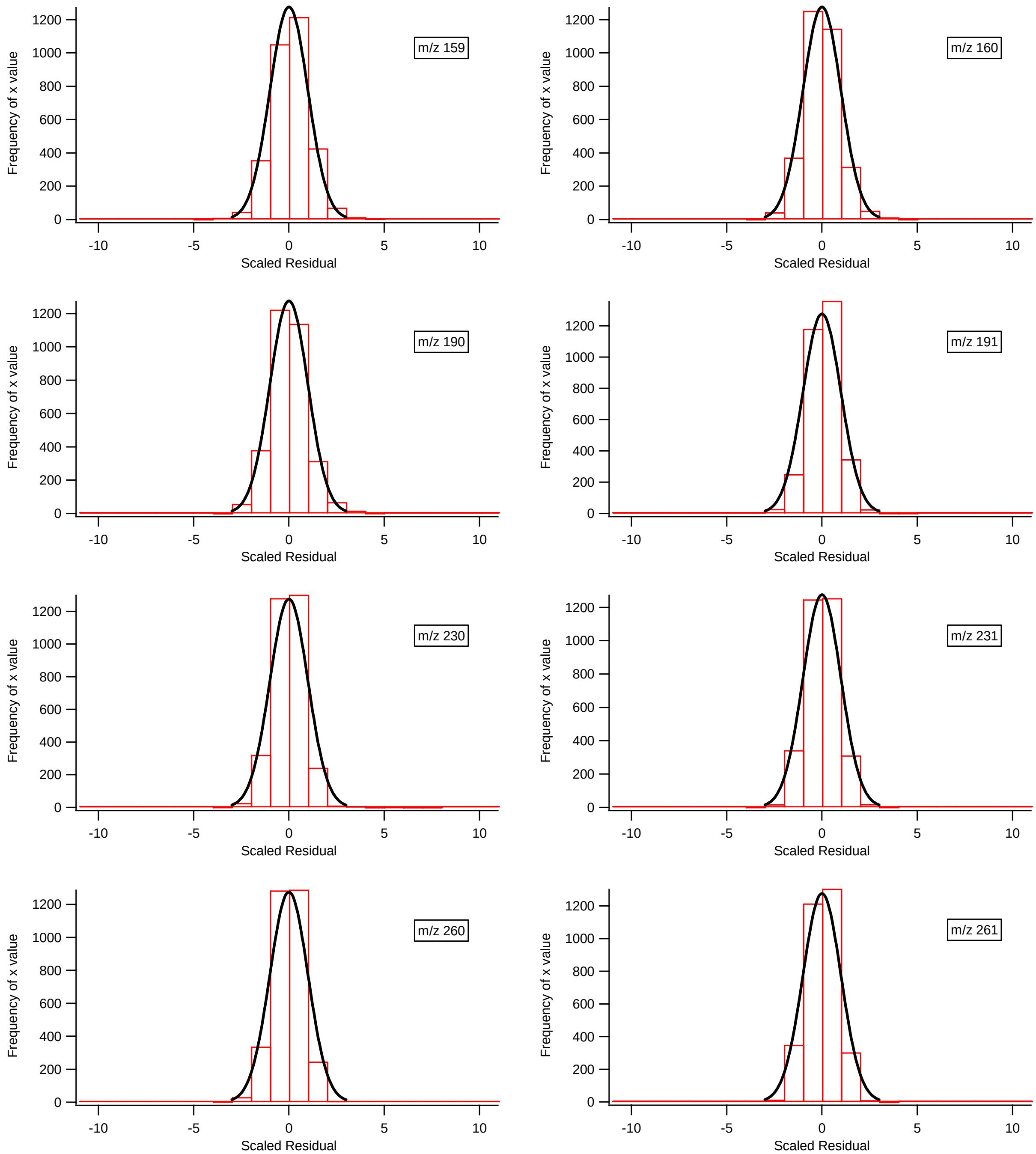

Figure S9 cont.

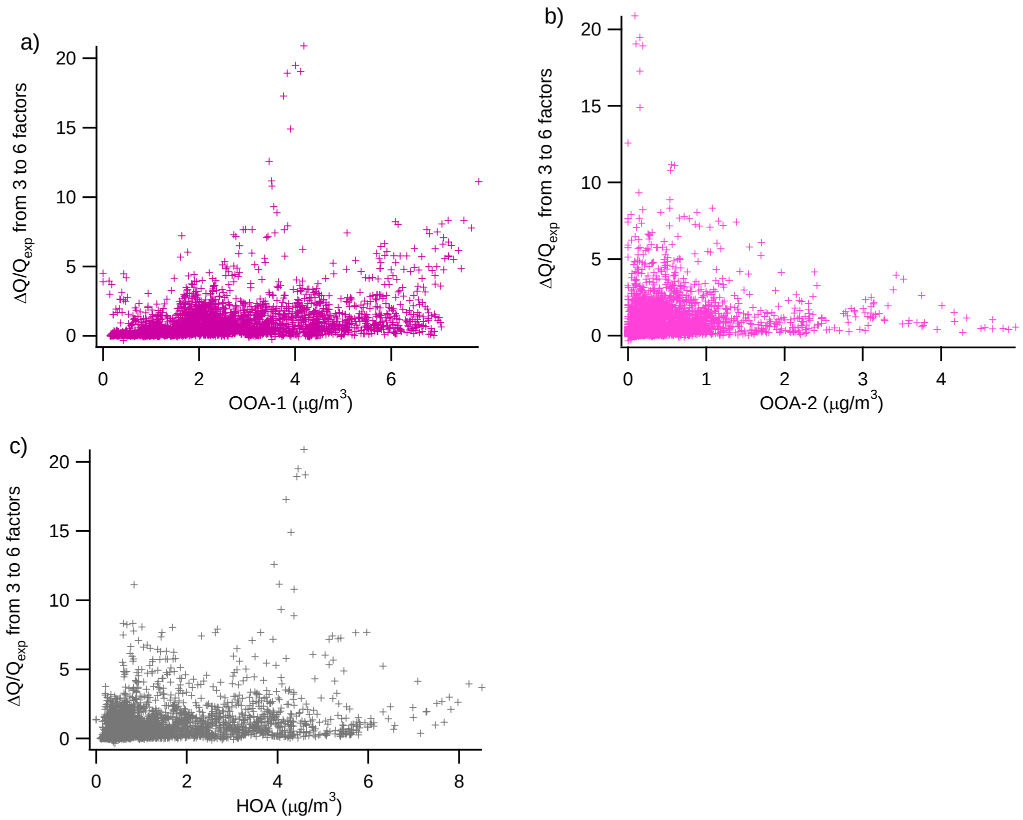

Figure S10. Change in Q/Q_{exp} in the real Pittsburgh dataset from 3 to 6 factors vs. a) OOA-1 contribution, b) OOA-2 contribution, and c) HOA contribution from the 3-factor solution.

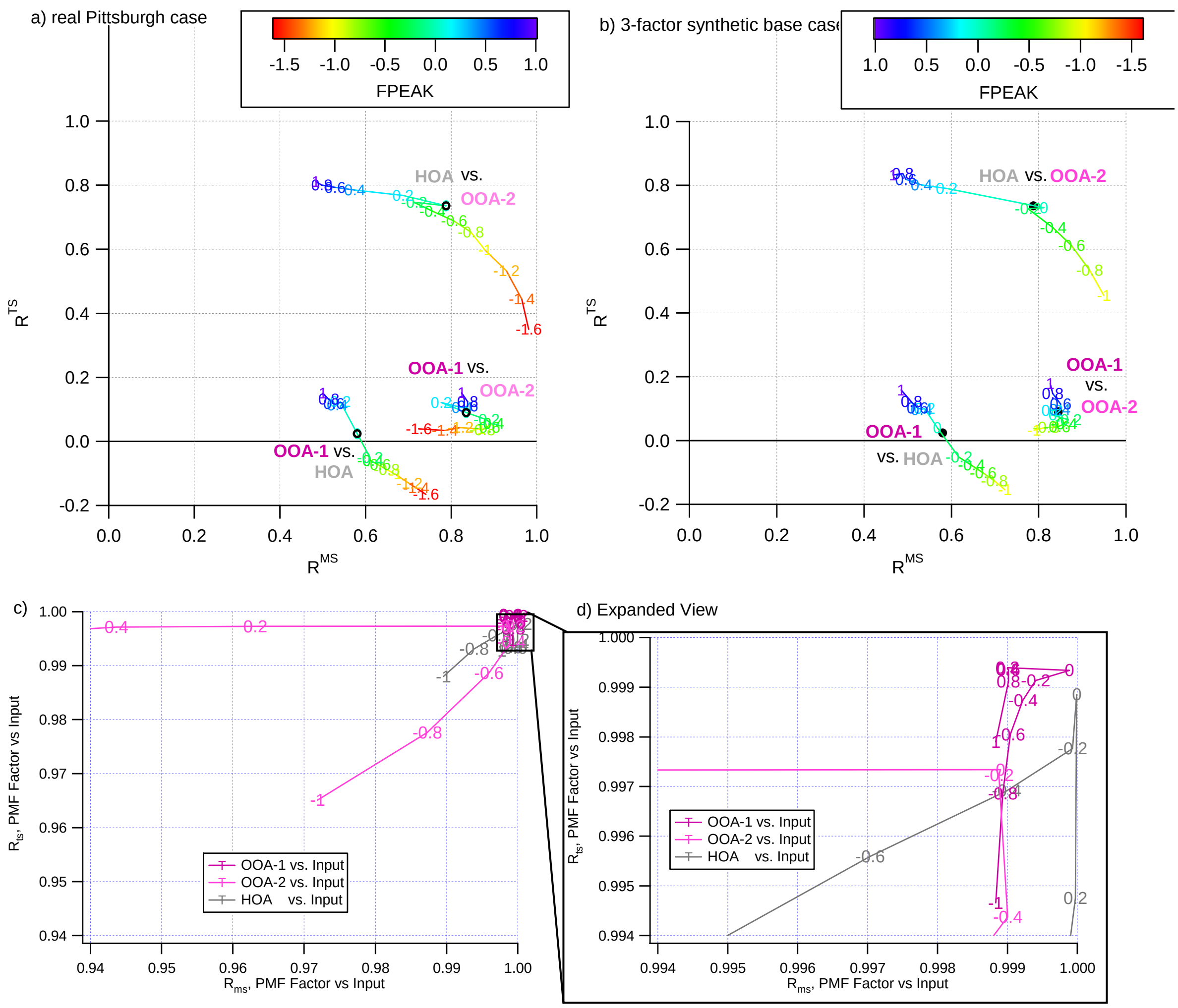

Fig. S11. Correlations of PMF factors to each other as they change with FPEAK for a) the real Pittsburgh case and b) the three-factor synthetic base case. The red labels denote the correlations of the input factors. Traces are colored by FPEAK and numbers denote the FPEAK of each solution. Black dots in b) indicate the correlation of the factors in the input. c) Correlation of the PMF factors to the input factors for the three-factor synthetic case. d) Expansion of c).

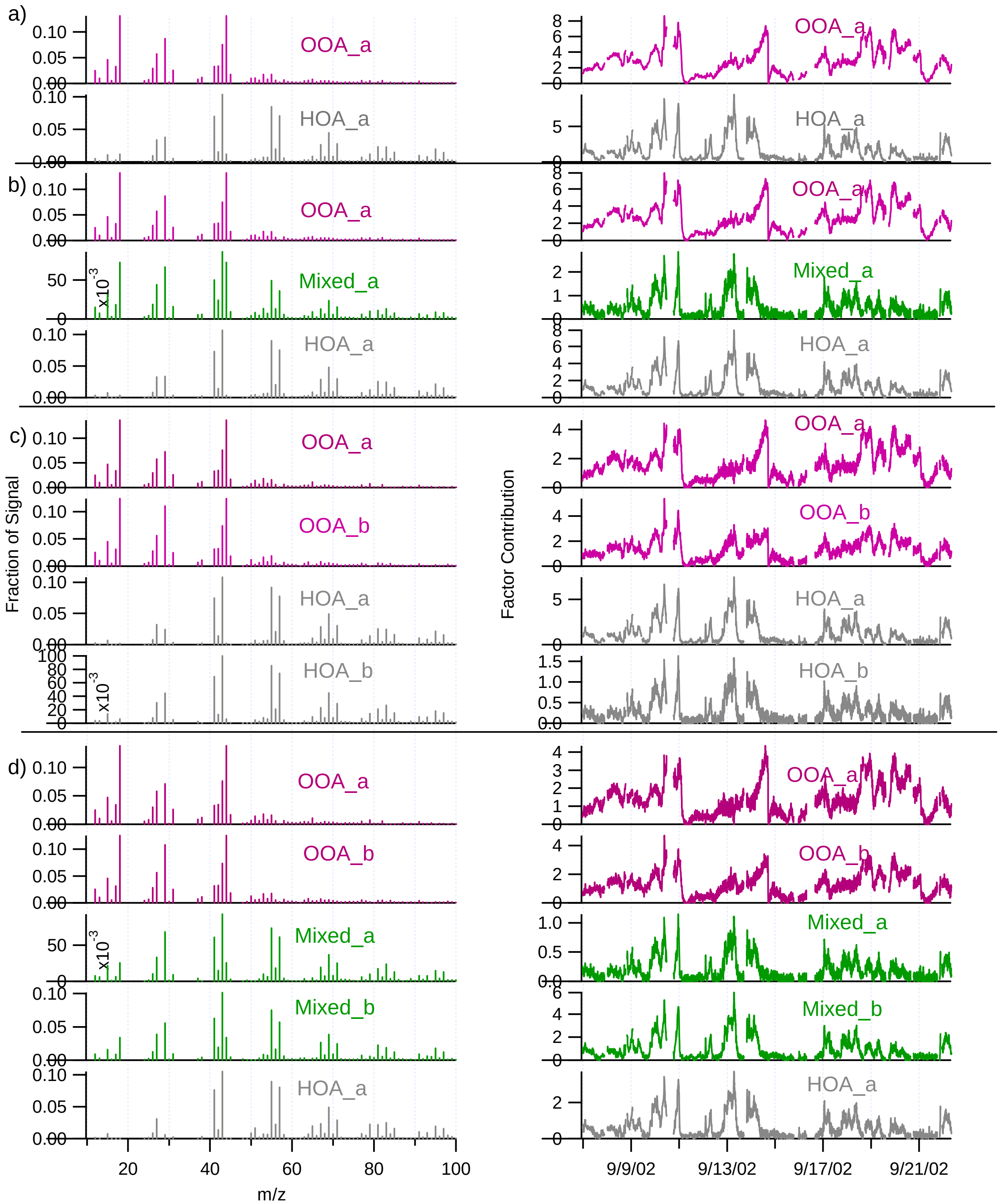

Figure S12. PMF solutions of the 2-factor synthetic base case with a) 2 factors, b) 3 factors, c) 4 factors, and d) 5 factors.

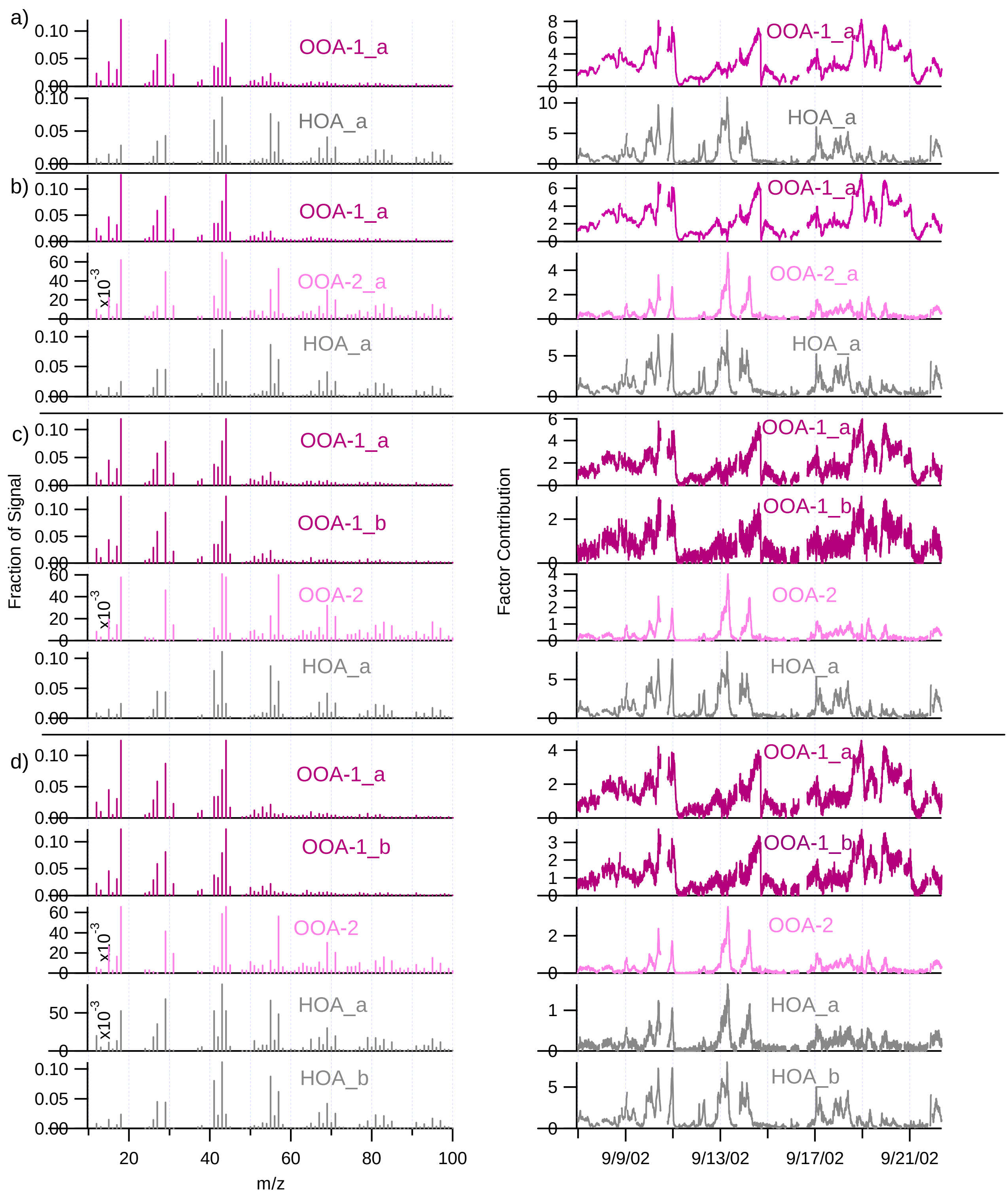

Figure S13. PMF solutions of the 3-factor synthetic base case with a) 2 factors, b) 3 factors, c) 4 factors, and d) 5 factors.

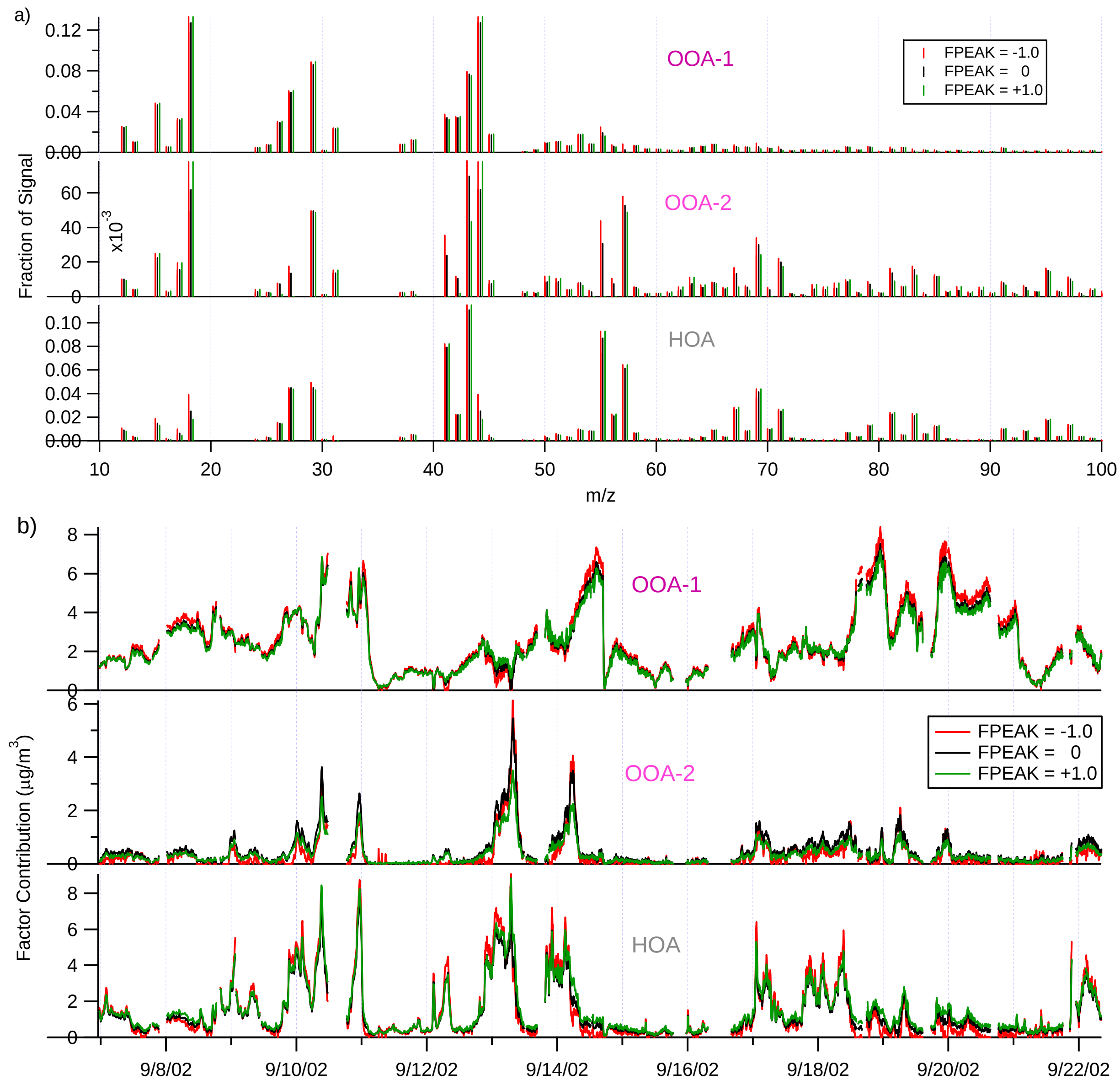

Figure S14. 3-factor solutions of the 3-factor synthetic base case for selected "good" FPEAK values.

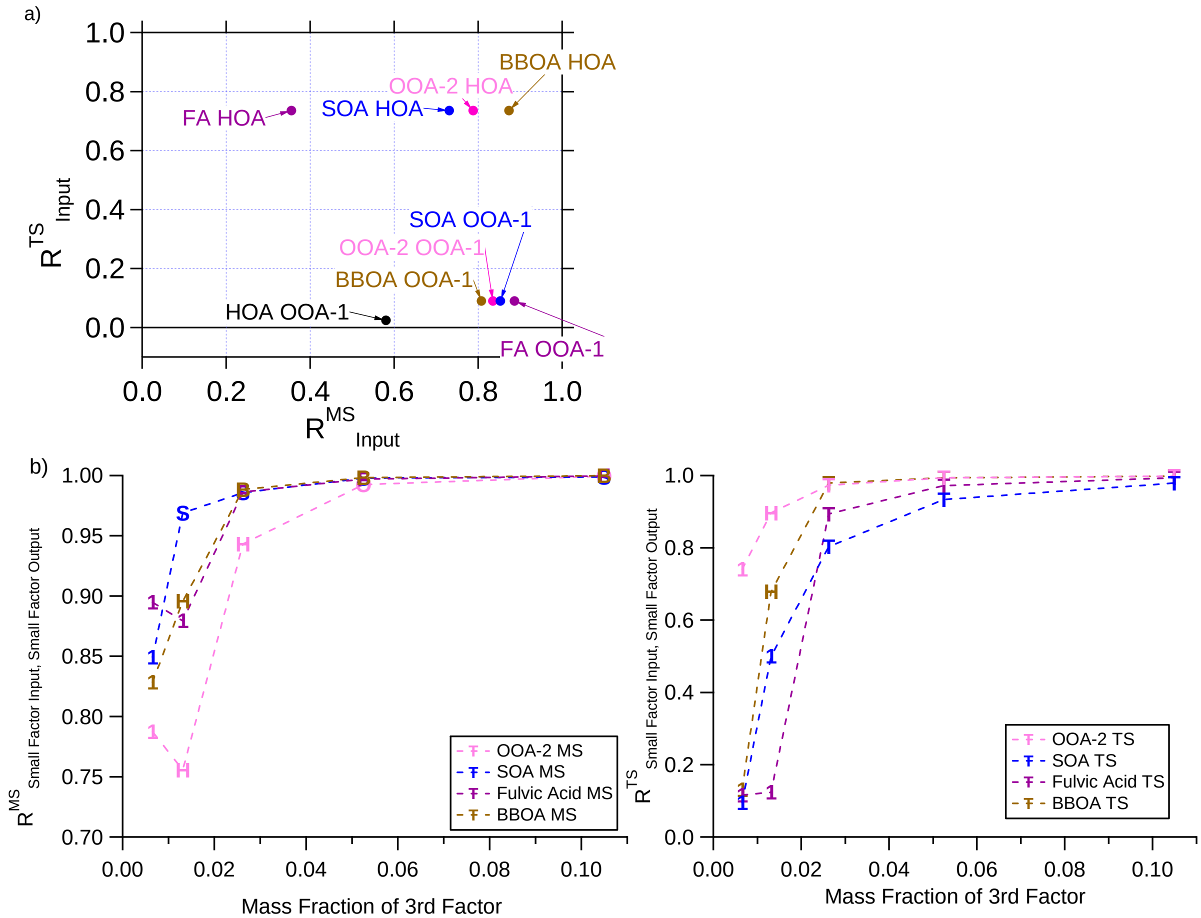

Fig. S15. a) Correlations between 3-factor real output and all synthetic input cases. b) Retrieval correlations between PMF and input TS and MS versus mass fraction of different third factors in synthetic cases. Markers denote the resemblance of the factors to the input MS or TS. For the MS, O, S, F, and B refer to OOA-2, SOA, FA, and BBOA, respectively. For the TS, T refers to the input TS. For both cases, H and I refer to HOA and OOA-1, respectively.

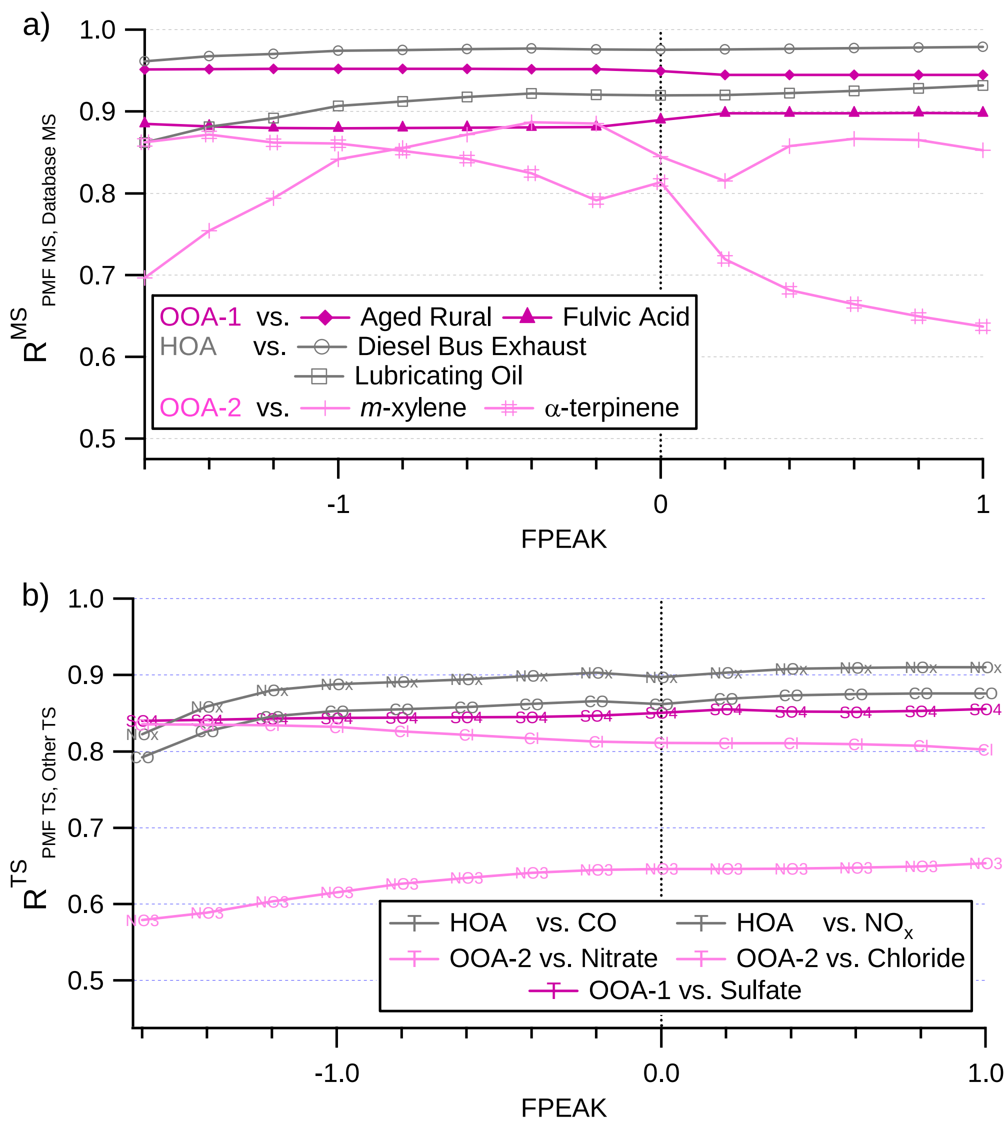

Fig. S16. Correlation versus FPEAK between PMF factor a) MS and selected reference MS, and b) TS and tracer TS.

Table S1. Correlations between PMF factor TS and tracer species.

	U^{TS}	R^{TS}
OOA-1 vs. Sulfate	0.95	0.85
OOA-2 vs. Nitrate	0.79	0.79
OOA-2 vs. Chloride	0.82	0.82
HOA vs. CO	0.93	0.93
HOA vs. NOx	0.95	0.95
HOA vs. EC	0.93	0.83

Table S2. Correlations between PMF factor and selected reference MS from the AMS spectral database.

Reference Spectrum	a) OOA-2, Real Data 3-factor solution				b) OOA-1a, Real Data 4-factor solution				c) "mixed" Factor, 2-factor Synthetic Data 3-factor solution			
	UC^{MS}	$UC^{MS\ m/z>44}$	R^{MS}	$R^{MS\ m/z>44}$	UC^{MS}	$UC^{MS\ m/z>44}$	R^{MS}	$R^{MS\ m/z>44}$	U^{MS}	$U^{MS\ m/z>44}$	RMS	RMS m/z>44
HOA Pittsburgh (Zhang et al.)	0.74	0.91	0.71	0.91	0.37	0.69	0.32	0.65	0.82	0.98	0.92	0.99
Diesel Bus Exhaust	0.90	0.93	0.89	0.93	0.68	0.63	0.65	0.59	0.92	0.97	0.96	0.98
Lubricating Oil	0.74	0.92	0.71	0.93	0.32	0.54	0.27	0.50	0.76	0.93	0.86	0.95
OOA Pittsburgh (Zhang et al.)	0.83	0.56	0.82	0.47	0.96	0.94	0.96	0.93	0.86	0.71	0.71	0.59
aged rural	0.84	0.78	0.83	0.73	0.98	0.96	0.98	0.95	0.87	0.91	0.73	0.85
Fulvic Acid	0.72	0.71	0.72	0.59	0.97	0.91	0.97	0.90	0.69	0.78	0.50	0.69
α-pinene	0.78	0.66	0.75	0.62	0.74	0.91	0.73	0.90	0.93	0.86	0.90	0.81
β-caryophyllene	0.82	0.71	0.80	0.65	0.71	0.93	0.69	0.92	0.94	0.86	0.92	0.81
linaloolSOA	0.77	0.62	0.75	0.56	0.68	0.83	0.66	0.81	0.87	0.74	0.81	0.64
α-terpinene	0.83	0.74	0.81	0.69	0.75	0.91	0.73	0.90	0.94	0.88	0.90	0.83
m-xylene	0.84	0.61	0.84	0.53	0.94	0.82	0.94	0.78	0.84	0.69	0.69	0.59
Diesel Exhaust, 0.25 hr	0.83	0.89	0.80	0.86	0.57	0.81	0.53	0.78	0.90	0.91	0.95	0.91
Diesel Exhaust, 2.25 hr	0.92	0.85	0.90	0.81	0.80	0.89	0.78	0.87	0.98	0.91	0.94	0.90
Diesel Exhaust, 4.25 hr	0.92	0.85	0.91	0.80	0.86	0.90	0.86	0.89	0.97	0.92	0.91	0.90
Ceanothus BBOA	0.88	0.85	0.85	0.80	0.70	0.84	0.68	0.81	0.94	0.90	0.94	0.89
Chamise BBOA	0.92	0.86	0.90	0.83	0.82	0.86	0.81	0.84	0.93	0.94	0.85	0.93
Palmetto BBOA	0.89	0.84	0.87	0.80	0.81	0.89	0.79	0.87	0.97	0.95	0.93	0.90
Juniper BBOA	0.92	0.87	0.90	0.83	0.84	0.85	0.83	0.82	0.97	0.94	0.94	0.94
Manzanita BBOA	0.88	0.85	0.85	0.81	0.72	0.85	0.70	0.83	0.95	0.91	0.92	0.94
Ponderosa Pine Duff BBOA	0.81	0.83	0.77	0.78	0.56	0.80	0.51	0.76	0.88	0.90	0.93	0.89
Ponderosa BBOA	0.85	0.85	0.83	0.81	0.64	0.82	0.61	0.78	0.93	0.91	0.91	0.88
Rice Straw BBOA	0.84	0.85	0.80	0.81	0.71	0.85	0.69	0.84	0.93	0.88	0.91	0.87
Sage and Rabbit Brush BBOA	0.88	0.86	0.87	0.82	0.93	0.89	0.93	0.87	0.91	0.94	0.79	0.92
Wax Myrtle BBOA	0.89	0.84	0.87	0.80	0.75	0.87	0.73	0.84	0.95	0.92	0.93	0.91
Levoglucosan	0.64	0.45	0.61	0.42	0.46	0.33	0.43	0.29	0.67	0.46	0.61	0.41

Table S3. Correlations between input factors for the 3-factor synthetic cases.

	OOA-2		BBOA		FA		SOA	
	U^{MS}	U^{TS}	U^{MS}	U^{TS}	U^{MS}	U^{TS}	U^{MS}	U^{TS}
HOA	0.81	0.84	0.88	0.84	0.39	0.84	0.76	0.84
OOA-1	0.84	0.55	0.81	0.55	0.89	0.55	0.86	0.55

OOA-1	
U^{MS}	U^{TS}
0.61	0.60

	OOA-2		BBOA		FA		SOA	
	R^{MS}	R^{TS}	R^{MS}	R^{TS}	R^{MS}	R^{TS}	R^{MS}	R^{TS}
HOA	0.79	0.74	0.87	0.74	0.36	0.74	0.73	0.74
OOA-1	0.83	0.09	0.81	0.09	0.89	0.09	0.85	0.09

OOA-1	
R^{MS}	R^{TS}
0.58	0.02