**Atmospheric
Chemistry
and Physics
Discussions**

Interactive
Comment

Interactive comment on "Simultaneous factor analysis of organic particle and gas mass spectra: AMS and PTR-MS measurements at an urban site" *by* J. G. Slowik et al.

J. G. Slowik et al.

jslowik@chem.utoronto.ca

Received and published: 29 August 2009

Response to Comments by P. Paatero
We greatly appreciate the thoughtfulness and attention to detail of the comments by P. Paatero. They have prompted a number of changes in the manuscript which we feel greatly improve the work. A point-by-point response is presented below. We have reordered the comments somewhat for clarity of discussion. For cross-referencing purposes, comments are numbered by reviewer and comment number, e.g. Comment 1.2 is the second comment in the first posted review. A parallel numbering scheme is used for figures and equations.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

Comment 1.1

It remains unclear what the authors did. They state that "the data from the two in-struments was combined into a single dataset, and PMF was applied to this unified dataset"

What does this mean? No mathematical equations or graphical illustrations are shown. Elsewhere in the ms, they write that the X matrix consists of time series of AMS and/or PTR mass spectra. At first, I thought that some columns of the combined X come from AMS while others come from PTR-MS. However, this cannot be true: in the sup-plement, they show how the G factor columns (time series for individual factors) are different for AMS and PTR. This would not be possible in the suggested setup where the same G matrix column applies both to AMS and to PTR-MS.

Response
Two points are raised in this comment: (1) the structure of the **X** matrix, and (2) the meaning of the different AMS and PTR-MS time series in Figs. S2 and S4. The two issues are addressed separately below.

(1) As suggested, the unified **X** matrix consists of time series of AMS and PTR-MS mass spectra. The **X** matrix contains 280 columns: 270 m/z from the AMS and 10 m/z from the PTR-MS. The structure of the unified **X** matrix is now illustrated in a new figure, shown below as Fig. 1.

(2) The issue raised by the reviewer in connection to Figs. S2 and S4 results from ambiguity in the definitions of the AMS and PTR-MS time series in these figures, which are also used in Fig. 11b. A similar issue is raised in Comment 4.10, and we repeat a portion of that response here. We clarify the time series definitions below and in the manuscript.

The AMS and PTR-MS time series shown in Figs. S2, S4, and 11b are not taken strictly from the columns of the **G** matrix. Because of the mixed units and difference

in scale, we feel a more meaningful presentation is obtained by plotting the factor time series for the unified dataset in the units of the respective instruments. Thus we do not directly report the g_{ih}, but rather $g_{ih}\left(\frac{\sum_j^{j=AMS} f_{hj}}{\sum_j f_{hj}}\right)$ and $g_{ih}\left(\frac{\sum_j^{j=PTR} f_{hj}}{\sum_j f_{hj}}\right)$, for the AMS and PTR-MS, respectively. (In our application, PMF2 is configured such that the output $\sum_j f_{hj} = 1$. Similarly, we do not directly report the f^{ih} in these figures, but rather the quantities $\frac{f_{hj}}{\sum_j^{j=AMS} f_{hj}}$ and $\frac{f_{hj}}{\sum_j^{j=PTR} f_{hj}}$ for the AMS and PTR-MS, respectively.

A slightly different reporting scheme is now utilized in the Supplement to overlay solutions to the individual and unified datasets, or to compare unified dataset solutions obtained at different seed, *fpeak*, or C_{PTR} (see Comments 1.14, 4.6, and 4.16). The object is to avoid the near-redundancy of separate plots of AMS and PTR-MS time series, which differ only by constant scaling factors, as evident above. We report $\frac{f_{hj}}{\sum_j^{j=AMS} f_{hj}}$ and $\frac{f_{hj}}{\sum_j^{j=AMS} f_{hj}}$ as above, but the time series are shown as $g_{ih}/\overline{g_{ih}}$ (i.e. normalized to the mean concentration). We report $\overline{g_{ih,AMS}}$ and $\overline{g_{ih,PTR}}$ for each trace so that the time series may be easily converted to the relevant units and the concentrations readily compared. These means are calculated as:

$$\overline{g_{ih,AMS}} = \overline{g_{ih}\left(\frac{\sum_j^{j=AMS} f_{hj}}{\sum_j f_{hj}}\right)} \tag{1}$$

$$\overline{g_{ih,AMS}} = \overline{g_{ih}\left(\frac{\sum_j^{j=PTR} f_{hj}}{\sum_j f_{hj}}\right)}$$

Mathematical explanations of these forms of presentation are presented in the revised manuscript.

Comment 1.2:

Weighting the member matrices differently is one important contribution in this ms. As far as I can remember, such weighting has not been published in connection with PMF analyses. It is difficult to know if this part of the analysis has been carried through properly because the presentation of the weighting procedure is not adequate. There are many sentences of verbal explanation and only one equation (3). All new mathematical procedures must be presented, first of all, by giving all the equations. All different quantities must have their unique symbols. One source of confusion is that the symbol s_{ij} denotes both the original data value uncertainties and also the "weighted uncertainties". Any verbal explanations should be understood as aiding in understanding the equations, not vice versa. If no one among the authors is able to formulate the equations and define the needed symbols, then a mathematically/statistically oriented scientist should be invited to join the team.

Response:

The mathematical notation has been clarified, particularly with respect to discussion of (1) the instrument-weighted uncertainties, (2) downweighting of outliers, and (3) time-dependent contributions to Q, discussed separately below.

(1) Instrument-weighted uncertainties

We have clarified the notation of uncertainties to distinguish between the following:

$\mathbf{S}$: matrix of untreated uncertainties, containing matrix elements s_{ij}.

$\mathbf{S_{inst}}$: matrix of instrument-weighted uncertainties, containing matrix elements $s_{inst,ij}$.

$\mathbf{S}'$: matrix of uncertainties in which outliers have been downweighted using the "pseudo-robust method" (see point (2) below and the response to Comment 1.3), containing matrix elements s'_{ij}.

$\mathbf{S}'_{\mathbf{inst}}$: matrix of uncertainties in which both the pseudo-robust method and instrument weighting are utilized, containing matrix elements $s'_{inst,ij}$.

A mathematical description of the instrument-weighting method has been added. The instrument-weighted uncertainty matrix ($\mathbf{S_{inst}}$, containing matrix elements s_{ij}) is con-

structed as follows:

$$s_{inst,ij} = s_{inst,ij} \qquad \text{for } j = \text{AMS } m/z \tag{2}$$

$$s_{inst,ij} = s_{ij}/C_{PTR} \qquad \text{for } j = \text{PTR-MS } m/z$$

(2) Downweighting of outliers
The need for outlier treatment constituted a major focus of the review (see response to Comment 1.3). We define the following terms for discussion of outlier treatment:
Untreated-outlier method: no downweighting of outliers has been performed.
Pseudo-robust method: outliers downweighted as described in response to Comment 1.3.
Robust method: outliers iteratively downweighted through use of the PMF2 robust mode (applies only to individual instrument datasets).

Implementation of the pseudo-robust method alters the definition of $\Delta\overline{e_{sc}}$. Specifically, s'_{ij} replaces s_{ij}, to prevent outliers from dominating $\Delta\overline{e_{sc}}$, i.e.

$$\Delta\overline{e_{sc}} = \overline{\left(\frac{|e_{ij}|}{s'_{ij}}\right)}_{AMS} - \overline{\left(\frac{|e_{ij}|}{s'_{ij}}\right)}_{PTR} \tag{3}$$

(3) Time-dependent contributions to Q
The time-dependent contribution to Q, Q_{cont} is an important metric in evaluating solutions. We now define this concept mathematically as follows:

$$Q_{cont} = \sum_{j}^{m} \left(e_{ij}/s_{ij}\right)^2 \tag{4}$$

As an additional metric, we report the change in Q_{cont} between the p-factor solution

C4306

and the $(p+1)$-factor solution, ΔQ_{cont}:

$$\Delta Q_{cont} = \left[\sum_j^m (e_{ij}/s_{ij})^2\right]_p - \left[\sum_j^m (e_{ij}/s_{ij})^2\right]_{p+1} \qquad (5)$$

Comment 1.3:

Lack of proper treatment of outliers is the major fault in this manuscript. There are samples for which the fit is very bad. Clearly the used PMF model is not valid at all for these samples. This applies both to AMS and also to PTR-MS. The authors have chosen not to use robust mode. A combination of strong outliers and non-robust analysis makes the results so questionable that the whole modeling work must be redone.

When large residuals have been noticed, it is necessary to perform a preprocessing of the data set, especially in cases where a robustized LS method cannot be used. The simplest form of preprocessing is to omit from the data set the samples with extra large residuals. This should be done already when AMS or PTR-MS is fitted separately. Omitting the samples should be done in stages: first omit only samples having the largest residuals. Perform a new fit and again delete the samples with largest remaining residuals, until a satisfactory fit is obtained. When a sample is omitted, omit from both data sets so that the same set of acceptable samples remains in both member matrices. – It is important that this preprocessing is documented in the publication. Useful statistics should be reported about how many values were rejected, did they display understandable patterns, and about possible sources causing these extra large (or extra small?) concentrations. N.B. an outlier may also be caused by equipment failure, e.g. so that some concentrations are erroneously reported as zero.

Response:

Interactive
Comment

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

We agree that a more systematic treatment of the outliers is required, particularly for the unified dataset. Three methods were suggested by the reviewer: (1) outright rejection of outlying periods, on either 1 or 15-min time intervals; (2) calculation of averages using trimmed means (n highest and lowest values omitted from averaging); and (3) strong downweighting of outliers (i.e. by a factor of 10 or more). We have implemented an outlier downweighting method in the revised manuscript. The motivation for selecting this method and its implementation are discussed below, and a description of the downweighting method has been added to the paper. We define an outlier as data point with a scaled residual e_{ij}/s_{ij} that satisfies the condition:

$$|e_{ij}/s_{ij}| > \alpha \qquad (6)$$

In the present dataset, outliers mostly occur during periods of high particle and/or gas concentrations, and at m/z with consistently high signal-to-noise. Under such conditions, e_{ij}/s_{ij} may become large while e_{ij}/x_{ij} remains small. This is a result of issues such as minor variations in source profiles and the general approximation inherent in PMF that ambient data may be represented through a finite number of static factors. As a result, it is desirable to retain information from these periods, but to prevent them from unduly pulling the model fit. We therefore treat outliers with a downweighting procedure, rather than excluding data altogether (whether through exclusion of time periods or by trimmed means). (To illustrate the issues raised in this paragraph, the following plots have been added to the Supplement: (1) time series of the number of outliers, as defined by Eq. 1.6; (2) time series of raw signal, reconstructed signal, e_{ij}/s_{ij} for selected m/z).

Use of a discrete cut point for outlier identification coupled with strong, constant outlier downweighting introduces logical inconsistencies in the treatment of data. Here data points for which e_{ij}/s_{ij} is slightly larger than α are strongly downweighted and only slightly influence the solution, while e_{ij}/s_{ij} slightly less than α exert a strong influence. It is desirable to obtain a solution in which the relationship between scaled residuals is

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

preserved, but the outliers do not dominate the fit.

The outlier downweighting method utilized here, which we term the "pseudo-robust method" is modeled on the robust PMF analysis [Paatero, 1997]. In robust PMF, the PMF task is defined as:

$$\arg \min_{\mathbf{G},\mathbf{F}} = \sum_{j=1} m \sum_{i=1} n \frac{(\mathbf{X} - \mathbf{Y})^2}{h_{ij}^2 s_{ij}^2} \tag{7}$$

Here $\mathbf{Y}$ is the data matrix reconstructed from the PMF solution (i.e. $\mathbf{Y} = \mathbf{GF}$), and h_{ij} are downweighting factors applied to the outliers according to the criteria:

$$h_{ij}^2 = 1 \qquad \text{if } |e_{ij}/s_{ij}| \leq \alpha \tag{8}$$

$$h_{ij}^2 = \left| \frac{(e_{ij}/s_{ij})}{\alpha} \right| \qquad \text{otherwise}$$

In robust PMF, the h_{ij} are calculated for each iteration of the solution process. For pseudo-robust analysis, only a single calculation of the h_{ij} is performed. For each unique combination of p, C_{PTR}, *fpeak*, and seed, PMF is applied twice. The first application is to the $\mathbf{X}$ and $\mathbf{S}$ matrices, and no downweighting of outliers is performed. From these results, a new uncertainty matrix $\mathbf{S}'_{\mathbf{inst}}$, (containing matrix elements $s'_{inst,ij}$) is calculated as:

$$s'_{inst,ij} = s_{inst,ij} \qquad \text{if } |e_{ij}/s_{ij}| \leq \alpha \tag{9}$$

$$s'_{inst,ij} = \sqrt{\left| \frac{e_{ij} s_{inst,ij}}{\alpha} \right|} \qquad \text{otherwise}$$

A second PMF calculation is then performed on $\mathbf{X}$ and $\mathbf{S}'_{\mathbf{inst}}$, yielding $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{E}$ for analysis.

Here we use α = 4, which results in 4.8% of x_{ij} classified as outliers for the selected solution of C_{PTR} = 10 and p = 6. As discussed below in response to Comment 1.11, the pseudo-robust method results in a significant reduction in the influence of the outliers on the solution.

As expected, the pseudo-robust method does not affect solutions for the individual instrument datasets (note the negligible difference between untreated-outlier and robust-outlier analysis discussed in the original manuscript). However, the pseudo-robust method alters the solution for the unified dataset in that a balanced 6-factor solution is obtained at C_{PTR} = 10 (vs. p = 5 for untreated outliers). The dependence of $\Delta\overline{e_{sc}}$ on C_{PTR} and p is also altered. Characteristics of the new factors are briefly summarized below:

F1$_{UN}$: Charbroiling emissions (similar to original F1$_{UN}$).
F2$_{UN}$: Traffic emissions (similar to original F2$_{UN}$).
F3$_{UN}$: Aged SOA (similar to original F3$_{UN}$).
F4$_{UN}$: Local SOA (similar to original F4$_{UN}$).
F5$_{UN}$: Oxygenated POA: Includes contributions from several oxygenated VOCs, particularly formaldehyde and acetaldehyde, as well as aromatics (though toluene/benzene ratio is significantly lower than for the charbroiling or traffic factors). Does not correlate with any available tracer.
F6$_{UN}$: Local point source: Includes acetic acid and aromatics. Particle spectrum is dominated by hydrocarbon peaks, though total mass may be too low to be reliable. Occurs in a few discrete events, increasing coincident with low wind speeds. Events end when the wind speed increases.

Comment 1.4:
This paper attempts to analyze results from two different instruments so that their results are merged into one large matrix. This is a very sensible and potentially useful approach. It has been tried repeatedly with PMF but several attempts have failed

and therefore remained unpublished. No formal analyses of the reasons of failures have been published. The likely reason is the following: A measured matrix of AMS, PTRMS, or other similar data may contain multiplicative errors or effects that affect uniformly entire rows of the matrix. All the values on any one row will become multiplied by an "erroneous coefficient" that is near unity but not quite equal to one. Such variation does not harm a customary PMF analysis. The composition factors remain unchanged and the erroneous variation is smoothly propagated to all time series factors. When matrices from two different instruments are combined, then the erroneous multiplicative variations in the two member matrices are generally different. This effect may distort the obtained composition factors. No resolution of this dilemma has been published so far. In practical work, the scientists should watch for signs of this problem in the residuals of the fit. If some rows of the PTR-MS part of the matrix mostly display positive residuals while other rows mostly contain negative ones, then this problem is indicated. In the present ms, the authors cannot possibly "solve" this problem but they should mention it and perhaps report about their findings regarding the residuals in their analysis.

Response:
This behavior was not observed in the selected solution to the unified dataset, as shown below in Fig. 1.2. This issue is mentioned in the manuscript, and Fig. 1.2 has been added to the Supplement.

Comment 1.5:
Using the term "residual" in order to denote Q is wrong. The word residual denotes the (signed) difference (measured-fitted) for any data value x_{ij}. The symbol Q denotes the sum of squares of scaled residuals, summed over all data values. The sum may be simply called "Q value". Sums of squares of scaled residuals over parts of data matrix may be called "Q contributions". If in doubt, you may include these definitions in your

text. But do not call Q a "residual" because this word already has another essential use.

Response:

We have corrected the terminology and now refer to a summation over part of the data matrix as the time-dependent contribution to Q, Q_{cont}. When this summation is performed for the unified dataset using the m/z from a single instrument, it is denoted $Q_{cont,AMS}$ or $Q_{cont,PTR}$. Mathematical definitions for these quantities have been added to the paper and are presented in response to Comment 1.2.

Comment 1.6:

"Equal weighting": the authors call their method of weighting the member matrices "equal weighting". This term is most unfortunate because it can be understood in three or four different ways. A much better term would be "balanced weighting". Even this term does not uniquely suggest what quantity was balanced. However, it guides one to think in the right direction.

Response:

The suggested terminology has been adopted in the manuscript. Additionally, a more precise description of the weighting methodology and its interpretation are provided as discussed in connection with Comments 1.2, 3.6, 3.7, and 4.1.

Comment 1.7:

The authors have correctly noted that there are two problems: (1) selecting a good weighting parameter C_{PTR} and (2) analyzing the rotational problem (= determining the range of possible rotations of the factors). They notice that variations of weighting have stronger influence on solutions than variations of rotations. Hence, they decide that it

is enough to solve the first problem, and they settle for one particular rotation, viz. the rotation obtained with fpeak=0.0, for both the member matrices as analyzed separately, and also for the combined matrix. Also, they seem to ignore the possibility of multiple solutions. If presence of multiple solutions was in fact excluded by running PMF from a sufficient number of random initial solutions, this fact should be documented!

This decision (of totally ignoring the rotational problem and the possibility of multiple solutions) might be OK if this ms would only be a presentation of the weighted approach to the two-instrument PMF analysis. However, the ms also contains several pages of discussion of the computed factors. This is a contradiction. If valid discussion of sources is attempted, then it is a MUST to either (1) demonstrate that the three solutions are rotationally unique (at least the PTR-MS solution is not!), or (2) to discuss the family of all rotationally possible solutions, both for member matrices analyzed separately and also for the combined matrix.

Response:
The reviewer raises two issues here: (1) testing for the presence of multiple solutions, and (2) rotational ambiguity in solutions to the individual and unified datasets. These issues are discussed separately below.

(1) The possibility of multiple solutions to the individual and unified datasets was excluded by initiating the PMF solution from 100 random starting points (seed 1 to 100 in the PMF2 software). All convergent solutions were found to be consistent with that obtained at seed = 1. Summary figures have been added to the Supplement comparing the seed = 1 profiles and time series with the highest and lowest values at each m/z and time point obtained through this analysis.

(2) Rotational ambiguity was analyzed for the individual and unified datasets using the $fpeak$ parameter. Solutions for each dataset were obtained at $fpeak$ = -1.5 to 1.5. A comparison of solutions obtained at $Pfeak$ = -1.0, 0, and 1.0 is presented in the

Supplement for each dataset. For the AMS and unified datasets, the solutions are very similar independent of $fpeak$. However, rotational ambiguity is evident in the PTR-MS dataset. Here the solutions at non-zero $fpeak$ are qualitatively consistent with that at $fpeak = 0$ in that factor identification is unaffected. However, quantitative analysis of the factor composition and relative factor concentrations is even more questionable than suggested by analysis of the somewhat ambiguous solution at $fpeak = 0$. Interestingly, comparison of the individual and unified factor mass spectra as a function of $fpeak$ suggests that PTR-MS rotational ambiguity is significantly decreased in the unified dataset.

Comment 1.8:
In another earlier AMS manuscript, the error estimates of AMS data values were erroneously computed so that a convolution with a box-car function of width three was ignored in the error computations. Apparently this convolution happens in the "standard" AMS software. The present ms does not explain if this problem was avoided or not. Considering that there is the risk that this box-car convolution still happens in the AMS software, the authors must definitely clarify the situation regarding this problem. If this question remains unanswered, interpretation of AMS Q values is impossible.

Response:
The reviewer's comment applies to a dataset collected in Pittsburgh and analyzed using PMF [Ulbrich et al., 2009] and a custom 2-factor mass spectral deconvolution method [Zhang et al., 2005]. In the Atmos. Chem. Phys. Discuss. version of the PMF analysis [Ulbrich et al., 2008], the AMS data was smoothed by use of a box-car function of width three to reduce high-frequency noise, but this convolution was ignored in error calculations.

In the final version of the manuscript [Ulbrich et al., 2009], this error has been rectified and is clearly designated as one of several data pretreatments applied specifically to

the Pittsburgh dataset. We therefore feel that the issue has been addressed, and that focusing on the analysis procedures used in the current study yields the clearest manuscript.

Comment 1.9:
On third line after Eq(2), the ms says that in an ideal solution, absolute value of the residual e_{ij} is equal to s_{ij}. This statement is absolutely wrong. In an ideal solution, these absolute values range from zero to more than 2^*s_{ij}.

Response:
This statement has been removed from the manuscript.

Comment 1.10:
There are also problems in the interpretation of PTR-MS Q values. The ms says "The theoretical value of Q, denoted $Q_{expected}$, is therefore equal to the number of elements in the input matrix X." This definition is based on an approximation that is useful if the number of columns in X is much larger than the number of fitted factors. For the PTR MS data, this approximation is not satisfactory. A better approximation for the expected Q value is obtained from the expression

$Q_{expected}$ = (number of elements in X) - ((number of elements in G) + (number of elements in F))

Although this expression is not accurate either, it should nevertheless be used because it is the best that can be known for the expected value of Q unless quite complicated analysis of scaled residuals is performed.

Response:

C4315

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

The expression suggested by the reviewer is now used to calculate $Q_{expected}$ in the manuscript for all three datasets.

Comment 1.11:

The obtained Q values are an important diagnostic for assessing the validity of PMF analyses. In this ms, Q values are not properly reported. The insets in figures are not sufficient for assessing Q values. From figure 1, one can estimate that the obtained Q was too large by a factor of 3 ... 5. Why did this fact not alarm the authors?

Response:

The figure insets have been expanded and now constitute a separate "b" figure. We have also added Q-values calculated using uncertainties from the pseudo-robust method (i.e. using S' rather than S), denoted Q_{pseudo}. Q-values for the selected solutions to the three datasets are listed below (and tabulated in the revised manuscript, see response to Comment 1.12):

AMS dataset ($p = 5$):
$Q_{true}/Q_{expected}$ = 7.55
$Q_{pseudo}/Q_{expected}$ = 4.83

PTR-MS dataset ($p = 5$):
$Q_{true}/Q_{expected}$ = 3.54
$Q_{pseudo}/Q_{expected}$ = 3.44

Unified dataset ($p = 6$):
$Q_{true}/Q_{expected}$ = 15.86
$Q_{pseudo}/Q_{expected}$ = 3.24

All of the Q_{pseudo} values are somewhat higher than expected. In the case of the AMS and unified datasets, this is likely due to the outliers (see response to Comment 1.3). This is supported by the significant difference between Q_{true} and Q_{pseudo}. For all three datasets, it is possible that the high Q-values may indicate the presence of additional meaningful factors at higher p. However, such factors cannot be supported with the current dataset through metrics other than the Q-value. (Further, in the AMS dataset factor splitting (an indication of excessive factors) is already observed at $p = 6$). Finally, for all three solutions the decreases in Q_{pseudo} at higher p become more incremental, whereas large improvements were obtained by moving from $p - 1$ to p.

Comment 1.12:
The ms hides all the important facts in the long verbose text. For better examples of presenting your results, please examine a few earlier PMF papers. Facts should be collected in tables. This ms does not have even a single table. Why? If I wish to know how many measurements were used, I have to search for this trivial fact by reading through the manuscript. Collect various statistics regarding both the original data and also the computed results in a few tables! Efficient use of tables makes the paper shorter and easier to read.

Response:
Figures 7 and 12 (fraction of m/z or AMS total mass apportioned to each factor) have been replaced by tables, and an additional table has been added for the AMS dataset. These tables contain all of the following that apply to the given dataset, as a function of factor: AMS mass fraction, PTR-MS m/z mass fraction, toluene/benzene ratio, ratio of AMS m/z 44 to total organics, and estimated AMS O/C ratio [Aiken et al., 2008]. For the mass fractions, two complementary calculations of this fraction and the associated uncertainties are presented, as discussed in response to Comment 3.11. Average fractional composition is calculated by (1) calculating the fraction at each time stamp,

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

then calculating the average fraction; and (2) calculating the average concentration across the study, then determining the fraction.

We have also added a table summarizing analysis parameters: matrix dimensions, α, and the range of p, $fpeak$, and seed values analyzed.

Another new table summarizes output values for the selected solution to each dataset: C_{PTR} (unified only), p, $Q/Q_{expected}$, $Q_{pseudo}/Q_{expected}$, and the percent of points in X classified as outliers (see response to Comment 1.3).

Comment 1.13:
It should be noted that the presence of significant outliers may strongly diminish the amount of rotational ambiguity. This might sound a desirable effect but it is not. The outliers may force the solution to one specific unique rotation, so that the scientist thinks that this is THE solution. However, this is a fallacy. The apparent uniqueness of rotation would be achieved because the model attempts to fit such data values (outliers) that are in inherent conflict with the model. Only after removing the outliers the rotational ambiguity of the model-conforming data can be observed.

Response:
Outliers are now treated using the pseudo-robust method, as discussed in response to Comment 1.3. Rotational ambiguity is discussed in response to Comment 1.7, solutions at non-zero $fpeak$ are presented in the Supplement.

Comment 1.14:
The results of the current modeling seem to indicate that when relative weighting of AMS vs. PTR-MS is changed by a "small" amount (such as a factor of two) the results change a lot. This behavior means that there is a conflict between AMS and PTR-MS. The weights act as a "partial" referee in this contest. If there would be no conflict, the

Interactive
Comment

adjustments of weighting would have only minor influence on the results. This aspect of the results should be discussed in the paper. However, at the current stage, the conflict can also come (I might even say, probably comes) from the outliers: the outliers in one member are not in harmony with the other member of the unified dataset. In order that the possible "real" conflicts can be identified, it is necessary to get rid of the conflict caused by the outliers. This is yet another reason why it is vital to process the outliers in a preprocessing step, as already discussed.

Response:
Outliers are now treated using the pseudo-robust method discussed in response to Comment 1.3.

We compare the 6-factor solution obtained at C_{PTR} = 10, yielding $\Delta\overline{e_{sc}}$ = 0.052, with the 6-factor solutions yielding $\Delta\overline{e_{sc}} \sim \pm0.25$, $\Delta\overline{e_{sc}} \sim \pm0.50$, and $\Delta\overline{e_{sc}} \sim \pm1.0$. The factor profiles and time series are overlaid as discussed in response to Comments 1.1, 4.10, and 4.16. Results are analyzed as a function of their sensitivity to $\Delta\overline{e_{sc}}$. At $\Delta\overline{e_{sc}} \sim \pm0.25$ the basic factors are preserved, with the major changes occurring in weakly attached features of a factor (e.g. paint fume contamination in the traffic factor).

Comment 1.15:
It would be good if the authors could discuss the basic assumptions underlying the work: when do the aerosol and gas concentrations vary together so that they form one factor in this model, and when do they vary so that they cannot enter in the same factor. If one source type emits both aerosol and gas, then covariation might be expected. However, if a source emits gas that then transforms into aerosols in the atmosphere, then the situation is more complicated. –If you can discuss this, please do. If not, that is OK, too.

Response:

A related issue was raised in Comment 2.1, and portions of the response are duplicated here.

In a fundamental sense, incorporation of particle and gas data into a single dataset for analysis is no different from analyzing gas and particle data separately. Differences in reactive lifetimes and variable emissions sources complicate PMF analyses generally, and are not unique to the unified dataset. For example, reactive lifetimes of co-emitted aromatic VOCs vary from minutes to hours, and the dependence of VOC and particle composition on engine load is well-documented. In practice, some sources and lifetimes are "similar enough" to be approximated as identical by PMF (e.g. traffic emissions, long-range transport of oxygenated aerosol and VOCs). The criteria for being "similar enough" vary from dataset to dataset, though some recurring trends are evident, e.g. the similarity between OOA-1, OOA-2, HOA, and BBOA PMF factors across datasets.

We note that, although gas and particle data has been combined into a single dataset, it is not required that every factor contains both VOC and particle data. This provides one means of testing assumptions relating to gas/particle covariance. In the present study, significant VOC signal is observed in all 6 factors resolved in the unified dataset. However, strong particle signal is observed in only 4 of the 6. While a particle mass spectrum is resolved in the remaining two, the particle signal is sufficiently low to be potentially unreliable. Factors including both VOC and particle signal result from both direct local emissions (charbroiling, traffic) and chemical processing (aged and local SOA formation). These examples can be grouped into three scenarios in which mixed gas/particle factors would be expected.

In the first scenario, VOCs and particles are co-emitted, and the source-to-receptor time is short relative to the atmospheric lifetime of the emitted species. This is the case for the charbroiling and traffic factors, where sources are in the immediate vicinity of the sampling duct. The combined gas/particle nature of the charbroiling and traffic

factors supports the reviewers suggestion that covariation would be expected from local sources emitting both gases and particles. This is particularly true for nearby sources, as differences in reactivity could potentially decrease covariation at longer distances.

In the second scenario, VOCs and particles are generated through related chemical reactions in the atmosphere, and have comparable lifetimes. This is the case for the aged SOA factor. Based on the correlation of this factor with sulfate and nitrate and analysis of back trajectories, it is likely that the particulate mass comprising this factor accumulated on the particle by photochemical reaction and persists in the atmosphere for days to weeks. This is similar to the atmospheric lifetimes of acetone and acetic acid, photochemical reaction products included in the factor.

Finally, in the third scenario, the local SOA factor provides a case where SOA and VOCs (primarily acetaldehyde) are produced through related chemical reactions but presumably have different lifetimes (days vs hours). In this case it is reasonable to expect a mixed factor when the production-to-receptor time is short (analogous to scenario one). However, the PMF treatment of the factor beyond this timeframe is more complicated, as a particle-only dataset might be expected to preserve a single factor, while a gas/particle dataset could push towards representing this situation with two factors, distinguished by the VOC tracer lifetime. Unfortunately, this question cannot be investigated in the present dataset, because the local SOA factor is only be obtained in the gas/particle dataset (and not the AMS dataset).

The two VOC-dominated factors both appear to result from local sources, although the lack of sharp features in the oxygenated POA factor may indicate a diffuse source. This suggests the lack of particle signal is driven by the emission source profile. Other possible scenarios in which VOC or particle-only factors might be expected include factors driven by gas/particle partitioning or by VOCs with very short lifetimes (e.g. biogenic molecules such as monoterpenes). However, no factor of either type was observed in the present study.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

References:

Aiken, A.C., DeCarlo, P.F., Kroll, J.H., Worsnop, D.R., Huffman, J.A., Docherty, K.S., Ulbrich, I.M., Mohr, C., Kimmel, J.R., Sueper, D., Sun, Y., Zhang, Q., Trimborn, A., Northway, M., Ziemann, P.J., Canagaratna, M.R., Onasch, T.B., Alfarra, M.R., Prevot, A.S.H., Dommen, J., Duplissy, J., Metzger, A., Baltensperger, U., and Jimenez, J.L.: O/C and OM/OC Ratios of Primary, Secondary, and Ambient Organic Aerosols with High-Resolution Time-of-Flight Aerosol Mass Spectrometry, Environ. Sci. Technol., 42, 4478-4485, doi:10.1021/es703009q, 2008.

Paatero, P.: Least squares formulation of robust non-negative factor analysis, Chemom. Intell. Lab. Syst., 37, 23-35, 1997.

Ulbrich, I.M., Canagaratna, M.R., Zhang, Q., Worsnop, D.R., and Jimenez, J.L.: Interpretation of organic components from positive matrix factorization of aerosol mass spectrometric data, Atmos. Chem. Phys. Discuss., 8, 6729-6791, 2008.

Ulbrich, I.M., Canagaratna, M.R., Zhang, Q., Worsnop, D.R., and Jimenez, J.L.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmos. Chem. Phys., 9, 2891-2918, 2009.

Zhang, Q., Alfarra, M.R., Worsnop, D.R., Allan, J.D., Coe, H., Canagaratna, M.R., and Jimenez, J.L.: Deconvolution and Quantification of Hydrocarbon-like and Oxygenated Organic Aerosols Based on Aerosol Mass Spectrometry, Environ. Sci. Technol., 39, 4938-4952, 2005.

Interactive comment on Atmos. Chem. Phys. Discuss., 9, 6739, 2009.

Full Screen / Esc

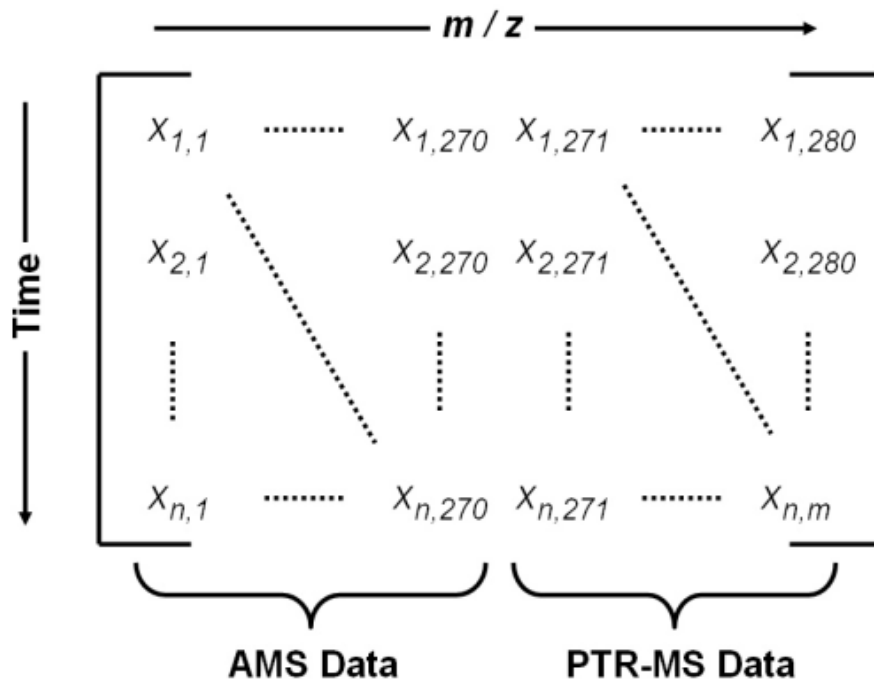Printer-friendly Version

Interactive Discussion

Discussion Paper

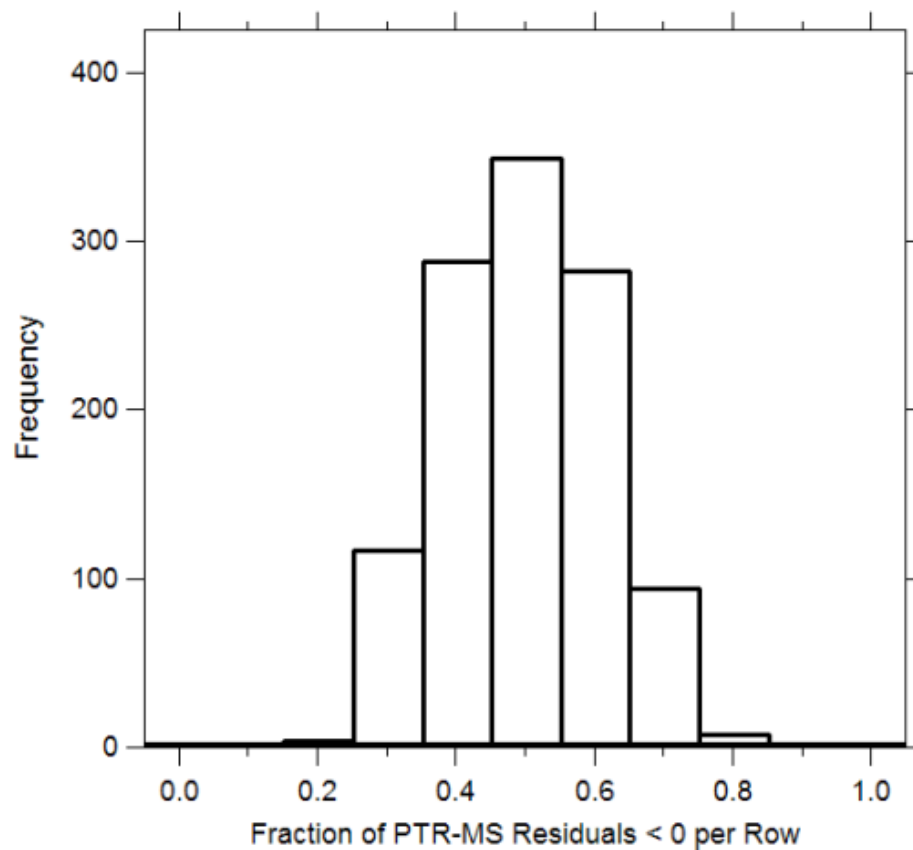**Fig. 1.** Schematic of the unified AMS/PTR-MS data matrix.

Fig. 2. Frequency of negative PTR-MS residuals per row (i.e. time point) of residual matrix E, for the solution at p = 6, C_PTR = 10 in the unified dataset.