Interactive comment on “Source characterization of Highly Oxidized Multifunctional Compounds in a Boreal Forest Environment using Positive Matrix Factorization” *by* Chao Yan et al.

P. Paatero (Referee)

Pentti.Paatero@helsinki.fi

Received and published: 20 June 2016

This manuscript describes PMF analysis of a large matrix of time-of-flight spectra of ions formed of atmospheric VOC molecules. The main part of the ms is interpretation of the obtained six factors. This main part of the work is not discussed in this review.

The ms puts much emphasis in deriving reliable uncertainty estimates for the elements of the measured mass spectra. This is intended in order to provide firm foundation for PMF modeling of these measurements, also for analysis of future similar measurements.

The comparison of obtained uncertainty estimates with obtained residual values is badly erroneous. Hence, conclusions about quality of fit are also erroneous. I recommend that this manuscript should be published in ACP after this comparison is performed correctly and all text based on this comparison is rewritten according to the corrected comparison. Also, I request that all the numerous problems discussed below are corrected (or present text is enhanced so that correctness of present text becomes evident).

The abstract says

PMF was performed with a revised error estimation derived from laboratory data, and this approach was validated by mathematical diagnostics of the PMF solutions. Unfortunately, this statement is erroneous. Mathematical diagnostics indicate that the carefully derived uncertainty estimates of data values are in striking conflict with the residuals obtained by PMF modeling.

The main part of this review is concerned with this discrepancy. Additionally, here are remarks regarding different erroneous or questionable details in the presentation. Although I will present some criticism regarding the estimation of uncertainty estimates, I believe that these estimates are sufficiently accurate so that when the computed residuals are seen to be much larger than the estimated uncertainties, then the discrepancy is real, not caused by errors in uncertainty estimates.

Essential mathematical diagnostics are only found in the Supplement. No hint of them is found in the main text.

Section 3 of Supplement is "Examining Q distribution of time and variables" It is good that this data is presented. However, its interpretation was not right, as shown in the following.

IT IS ASSUMED THAT MODEL ASSUMPTIONS OF PMF DO HOLD

If model assumptions do hold, then residuals are only due to data noise, so that assumed data uncertainties agree with observed distributions of residuals. If model assumptions do hold, then all profiles stay unchanged throughout the measurement period, and the assumed number of factors is right.

Notation: The dimensions of the matrix are m rows, n columns. There are p factors. Q sums over columns and rows are denoted by Q_j and Q_i, respectively. When we say "residuals" we mean scaled residuals, i.e. residuals divided by respective assumed data uncertainties.

THEORETICAL VARIATION OF Qrow AND Qcolumn VALUES The Supplement says:

"The mean value of Q on all variables was well below 4, the threshold in robust-mode PMF. This suggests that all variables are well described by the model." These sentences confuses single point Q_ij contributions with the overall Q sums ob- tained for an entire matrix, for an entire row, or for an entire column. It is possible (within model assumptions) that a few individual points get residuals >4, whereby the Q_ij contributions from such points exceed 16.

Estimates for Q contributions due to columns or rows are obtained from Statistical theory. Approximately, theory says that the expected value of Q_j from any column j is = m-p, and from any row i, Q_i = n-p. It also says that approximately the statistical distribution of Q is equal to chi-squared distribution whose degrees of freedom is m-p for Q_j and n-p for Q_i.

In the present work, p«n and p«m, thus we may approximate: Distributions are: chi2(n) for row Q’s and chi2(m) for column Q’s. As m and n are large, chi2 is well approximated by the normal distribution. Thus distributions of Q values are approximately:

for row Q’s, distribution of Q_i is N(n,sqrt(n))
for column Q’s, distribution of Q_j is N(m,sqrt(m))
where N(*,*) denotes normal distibution and sqrt(m) and sqrt(n) are the standard deviations of respective normal distributions. As an example, compute limits of these distributions for n=400, m=10000. Then the lower and upper 2-sigma limits for Q values, under model assumptions, are

360 to 440 for row Q’s
9800 to 10200 for column Q’s

It is seen that Q values come very close to their expected values m and n when model assumption are valid. Such "well-behaving" Q values are obtained in numerical simulations when the only simulated error is the random error in data values. If computed Q values deviate more in analyses of real data, then random noise in data values cannot be the explanation if assumed uncertainties are correct for data noise.

COMPARISONS WITH Q VALUES SHOWN IN SUPPLEMENT

There appears to be a conflict between parts a and b of Fig. S8. The overall averages of Q in parts a and b should be identical because they are averages of the same matrix of individual Q_ij values. In the following discussion, we only consider part b of Fig. S8. which seems to agree with the value of Q/Qexp for 6 factors that is given for the overall Q in Fig 7 of the manuscript proper.

Distribution of Qrow values (fig S8/b) ranges from approximately 0.2n to 2n, some rows even exceed these limits. This variation is very much wider than the expected width of chi2 distribution for Qrow.

Thus it is concluded that model assumptions did not hold for this PMF modeling. The estimated noise in measured values does not explain the observed variation of Qrow from row to row. Small values of Qrow may have a simple explanation: censoring of values <DL (see below) has eliminated some variation from the data, so that Q contributions from BDL values are much less than expected. Thus Qrow values of low-intensity rows will be (much) smaller than expected.

Note that downweighting low-intensity ("weak" and "bad") columns will also decrease both Qrow and Qcolumn values below their expected values. On the other hand, there are significant numbers of rows with Qrow > n + 4 sqrt(n).

It is difficult to know the reasons for these large Qrow values. Possible reasons are e.g.:

- variation of component profiles from sample to sample.

- small slow variation of mass calibration and/or resolution from sample to sample.

- small variation of critical parameters of the ionization process, e.g. of temperatures

Careful study of residuals would be the first step in finding the reason(s) for increased Qrow and Qcol values.

It is essential to admit that "something" happens in the atmosphere, or in the measurement process, that is not currently understood.

One must not try to "explain away" this "something" by arguing that yes, this was already seen by others, there is nothing new in such variation of row and column Q values.

On the contrary, this is indeed not an exceptional case. It is a phenomenon that should be studied so that it is understood. If there is component profile variation, understanding this variation might be a significant step in understanding chemical processes in the atmosphere.

It is important to modify the manuscript so that this "something" is clearly presented. It is not necessary nor possible to determine in this manuscript the reasons behind the observed Q variation.

It might perhaps be good to discuss or mention possible reasons. Figure S8 is crucial in demonstrating this Q variation. It might be good to move Fig S8 to the main text. After all, determination of data uncertainties is one of the main contents of the paper. Importance of Fig S8 is based on carefully determined data uncertainties, thus it is not logical to hide Fig S8 in the Supplement.

It would be good to point out that the observed good overall Qexp values are misleading in this case. Some Qrow values are much too large while others are much too small, so that these two effects largely cancel each other in the value of overall Qexp. See also remark lines 306,307, below.

We would like to thank Pentti Paatero for giving the constructive, helpful and detailed comments and suggestions, especially for the discussion on the diagnostics of PMF solutions. These comments and suggestions greatly improved this manuscript.

According to the comments, we modified the manuscript in the following aspects:

1) Correct all problematic statements on the diagnostics of PMF solutions, and move the figure showing Q distribution to the main text, so that the “something” is clearly presented.
2) Improve the Supplementary Information by correcting typos and mistakes, and polish the language;

In the following, we respond to the referee’s comments item by item.

DETAILED DISCUSSION OF THE MAIN PART

Eq(1) is unclear. What quantities are represented by [X] and by the numerator and denominator. The text says

"the numerator on the right hand side is the sum of all detected ions" This probably means: ... is the sum of detected ion concentrations?

Further: "the denominator is the sum of all reagent ion signals" What does this mean? Note that there are no square brackets in the denominator.

We simplified this equation to

$$[HOM] = \frac{HOM(NO_3^-)}{\sum_{i=0}^{2}(HNO_3)_i(NO_3^-)} \times C$$

For each HOM molecule, its concentration ([HOM]) equals to its respective detected signal normalized by the summed reagent ion signal, multiplied by the calibration coefficient *C*. We also clarified the related text.

On lines 126, 132, 176, and possibly elsewhere, PMF is called "an algorithm". This is wrong. PMF is a model, it defines the equations that should be fulfilled by the computed factor elements. Algorithm is a procedure for finding the values for factor elements so that they fulfill the model. There are currently at least 4 different algorithms for fitting or "solving" this model PMF. Please use correct terminology! Admittedly, the majority of chemically oriented papers do not pay attention to this distinction. In fact, it would be good to specify which PMF algorithm was used: the original algorithm in PMF2, or a PMF script executed by program ME-2? There are slight differences between these programs, especially if there is rotational ambiguity in the model. (There is probably very little rotational ambiguity in this work, so that the distinction PMF2 vs. ME-2 does not matter now. However, it is good manners to specify the used tools.)

Agreed. We corrected the terminology. Sofi5.2 is based on program ME-2, we specified this in line 132-134.

Lines 167 - 169 say "... Therefore, the data matrix used in this work is in unit-mass resolution, and peak fitting was performed afterwards to identify the elemental formula of peaks..."

This is an important decision and probably quite suitable for these data sets. There would also be other ways of formulating the matrix. It might be useful to learn whether the authors experimented with different ways, and what considerations lead them to select the unit-mass resolution. However, if the authors plan to examine this question later in more detail in another paper, then it is ok to not discuss this question now.

While the use of UMR data here does not make full use of the acquired HR spectra, the determination of errors for HR fits becomes much more complex. Especially in this case, when our signals are low, we would first need to average to at least 1-hour data in order to get most peaks smooth enough for reliable fitting. In addition, even assuming all peaks are smooth enough, the presicion of peak fitting on overlapping peaks shows complicated dependence on mass calibration, resolving power, and peak intensity, as studied by Cubison and Jimenez (2015). Thus we chose UMR data as this already gave us much new insights into the HOM formation pathways.

Cubison, M. J. and Jimenez, J. L.: Statistical precision of the intensities retrieved from constrained fitting of overlapping peaks in high-resolution mass spectra, Atmos. Meas. Tech., (8)2333-2345, 2015.

Lines 184-186 say " I is the signal strength (ions/second) of the ion, ts is the integration time in seconds, and a is a factor accounting for the fact that a single ion will generate a Gaussian-shaped pulse in the detector, rather than a single peak. " Here is confusion (or sloppy wording). I am not sure how to understand this topic.

First, I believe that the words "generate a Gaussian-shaped pulse in the detector" are a mistake. The intention is probably to say that "individual ions produce pulses whose pulse height distribution is of Gaussian shape."

Second, why does the pulse height distribution matter at all? If ions are not actually counted but count rate is determined by integrating the current that is due to accumulated pulses, then the statement would be understandable: the variation of charge produced by each ion does indeed contribute to the uncertainty of integrated current. In contrast, if ion pulses are actually counted (I believe this is the case), then the variation of pulse height from ion to ion does not directly contribute to uncertainty, except if the variation is large enough so that a fraction of ions are not counted at all. Please clarify or correct.

This sentence meant to say that the signal generated by a single ion in the micro-channel plate (MCP) detector is not constant but follows Gaussian distribution, which also contributes to the overall imprecision. The referee is right, since the CI-APi-TOF used in this work was using a time-to-digital converter as the data acquisition card, the Gaussian distribution of signal height should not affect the accounting statistics, as all signals were amplified to be large enough to cross the threshold. The factor *a* was empirically determined by fitting the lab data, and it might arise from some unaccounted uncertainty in the data, such as shifting of mass calibration, baseline correction. We have corrected the statement in the main text. Keeping the empirical factor *a* also makes the equation (Eq.7) applicable regardless of the type of the data acquisition card.

Section 2.3.2. It would be good to state clearly that the data matrix consists of counts-per-second values, obtained as 5-minute averages. This fact can be inferred from the present text but why not help the reader by stating it explicitly.

Agreed, we now stated it explicitly. We slightly modified the sentence to “These input data matrix consists of counts-per-second (cps) values, averaged from raw data using 5-minute time resolution, and a total number of 9084 mass spectra were then obtained”.

The paragraph beginning on line 209 claims that Paatero et al (2003) recommended censoring variables that are below DL by fixing the values at DL/3 and uncertainty at DL. This claim is entirely fictitious and wrong. There is not a single word suggesting censoring BDL values in the 2003 paper.

In contrast, certain PMF-related papers advise against this practice. There is no demonstrated benefit from this practice, provided that low S/N (i.e. "weak" and "bad") variables (entire columns) are downweighted as recommended in that 2003 paper. On the other hand, my personal experience in reviewing has revealed several cases where such censoring created one or two ghost factors, i.e. numerical artefacts caused by censoring. Also, censoring often creates bias in the results.

It is possible (likely?) that in the present case, censoring BDL values did not cause noticeable harm in main results because there were so many strong variables. The only likely harm might be that some details were lost from those columns where itensities are lowest. Thus it is not reasonable to suggest that the work should be redone using the original measured BDL values and uncertainties. On the other hand, the present formulation of the paper would be interpreted by your readers as a rule saying that BDL values -must- be censored. In order to help prudent practices prevail among atmospheric scientists, I request the following addition in the ms:

After explaining that BDL values were replaced by DL/3 in this work, you shall insert a remark, something like the following:

After this work was completed, we became aware that this practice of replacing BDL values by fixed values is harmful and provides no advantage at all, although in this specific case, the main results were possibly not harmed. Thus we emphasize that in future studies, our example should not be followed. Instead, values < DL and their uncertainties should remain unchanged in data and error matrices.

This was a wrong citation. The approach we used in this work is similar to what has been suggested by Polissar et al. (1998). We corrected the citation. We also realized that there was another mistake that we actually replaced corresponding error with 2DL ($6\sigma_{noise}$) instead of DL ($3\sigma_{noise}$). This is to make those data as “bad signals”. For the variables that contains many censored data points, very likely they will be further downweighted by 10-fold. We also corrected this in the manuscript and the supplement.

We checked the effect of censoring data in our case by running PMF with uncensored data. We evaluated the similarity of the results by checking the correlation between solutions using censored data and those using uncensored data. The results are provided in Table S1. From 2 factors to 6 factors, PMF solutions are almost identical (UC>0.99). Similar to what is expected, censoring data did not affect the results in our case. Even so, we do agree that censoring data is probably not a good practice. We will add some discussion in supplementary information S3, where we recommend readers not censoring data in future works.

Signal-to-noise estimates (S/N) are discussed in the paragraph beginning on line 209. Please state which formulation of S/N was used. There are two published formulations:

(1) the recommended S/N definition, distributed with EPA PMF v5 and published in the Supplement of

Brown, S. G., Eberly, S., Paatero, P. & Norris, G. A., Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. Science of the Total Environment. 518-519, p. 626-635, 2015

(2) the earlier problematic S/N definition, suggested in the quoted 2003 paper and used in all earlier EPA PMF versions.

The numerical values produced by these two methods differ from each other, thus the readers need to know which method was used by you: one of these, or your own method (define).

The Sofi5.2 program uses the earlier S/N definition suggested in Paatero et al. (2003). We now mention this in the main text.

The manuscript mentions that some columns were downweighted (DW) by 2 or by 10. State how many columns were DW by 2 and by 10. Numbers of DW columns also influence the Qexp values. When you reported Q/Qexp, did you use correct Qexp values that take this influence into account? If not, state this clearly! If yes, state that, too!

The influence of DW columns on Q_{exp} value has been taken into account in Q/Q_{exp} values. We have downweighted 173 variables as weak signals and 152 variables as bad signals. Fig. S9 in the revised supplement shows the distribution of signal-to-noise ratio, and we added this information in the main text.

Lines 306, 307 say "From two to seven factors, Q/Qexp decreases stepwise from 2.44 to 0.76. The closeness to unity indicates that the estimated error is appropriate for the model."

This good agreement of Qexp with its theoretical value is misleading (see above, too). Because of downweighting and/or censoring, low concentrations contributed to Qexp much less than expected. This is OK once it is recognized. However, there are also high concentration values that contribute to Qexp much more than expected, so that the overall Qexp appears acceptable. Thus it is not right to claim that the estimated error is appropriate for this PMF modeling with 6 or 7 factors.

Agreed. We re-wrote this section and corrected all wrong statements concerning Q/Q_{exp}.

DETAILED DISCUSSION OF THE SUPPLEMENT

There are too many typos and broken sentences, poor language, and even mistakes in the equations. It must be emphasized that all text, even the Supplement, should be carefully checked by one or two of senior authors before the manuscript is submitted. Language must be improved in the whole of Supplement text. This is essential in order that the text may be understood!

lines 9, 10, 35

These lines refer to "a" in Eq(6). There is no "a" in Eq(6)? Confusion? Is Eq(7) intended? This confusion may be present elsewhere, too

The referee is correct, Eq. 6 should be Eq.7. Other similar mistakes were also corrected, and the language of the Supplement was considerably improved.

line 33 says

"also confirms the validity of the pre-assumption" which pre-assumption? Do not pose extra difficulties for the reader, please be explicit!

This sentence was indeed confusing. We removed this sentence and rephrased the whole paragraph.

line 64 says

"by dividing the "noise estimate" (i.e. signal minus trend) data" what does "trend" mean here. I cannot even guess.

We now changed the sentence to "…by dividing the "noise estimate" (i.e. signal minus moving median) data into bins according to their signal (cps)" to be more clear.

lines 64 to 70

This paragraph is almost impossible to understand. Equations are the language of mathematics. Please use equations as the main method of defining what was done. Verbal explanations may only be used as a help for understanding the equations.

The paragraph has been revised and a figure added to illustrate the signal binning procedure.

lines 82 to 89

This paragraph confuses superposition and convolution. First do the math properly, then rewrite the paragraph.

Quite right. As this paragraph had redundant verbal explanation of Eq. 6 to Eq. 8 and Eq. S1, and only otherwise stated the obvious, we have omitted it for clarity.

Fig S1, caption contains: "All the flows were set identical throughout the experiments" better to write: "All the flows were kept unchanged throughout the experiments"
or "All the flows were constant throughout the experiments"

Agreed, we changed it to “All the flows were constant throughout the experiments”

Eq(S1)

This equation contains a parameter "a". The "Allan equation" Eq(7) in the main text also contains a parameter "a". However, the equations are different, and the "a" values are thus different, too. In Eq(7), the "a" is dimensionless and approximately =1. In Eq(S1), the "a" has dimension and its value depends on integration time.

This confuses the reader significantly and quite unnecessarily. Please rewrite the supplement so that the same equation is used in both texts.

We agree, this was confusing. We now re-labeled *a* in the supplement equations and text to *c*, and added a sentence to explain its relation to the Allan equation.

lines 103-104 say

"Parameter "a" is similar the "a" in the Allan et al. (2003) equation".

This statement is badly misleading, as already noted, above. If Eq (S1) is not changed as I suggested, then this statement must be changed to its opposite, warning the reader that the two symbols "a" are not the same.

Corrected, same as above.

Eq(S2): There is a problem with this equation. I suspect that an equals sign is missing before the square root.
The referee is correct. This was amended.

Eq(S3)

I do not understand this equation at all. What are the X symbols? What is the equation trying to say?

This equation was supposed to contain the final numerical values for c, e and their errors. It is now corrected in Eq.S2.

Fig S7.

These confidence limits for sigma values do not make sense. There must be some problem in their evaluation. Possibly, an invalid estimation principle was used.

We checked the error values and there was indeed a problem - with plotting, absolute conf int. (min/max) values were plotted as error bar height (from centre), making them way too large. Now corrected.

Fig S8

I assume that s8/a represents the 6-factor solution (why is it not stated?). Then Figs a and b are based on same Q_ij values. However S8/b and S8/a of Q/Qexp seem to be in conflict: there are many more values >1 in S8/b than in S8/a. Is there a natural explanation (show the explanation if there is one) or is there an error in generating the figures?

Both sub figures are intended for evaluating 6-factor solution, we now explicitly mention this in the figure caption. The original Fig.S8a suffers a mistake in generating the figure. Also in original Fig.S8a&b, we used the estimated error before downweighting to calculate the Q_{ij}, and this causes a discrepance in comparison to the overall Q/Q_{exp} given by Sofi using the error after downweighting. We corrected both mistakes. As suggested in the general comments, we now moved the Q distribution figure to the main text as Fig. 7b&c.

Fig S9
The caption says "... the purple one is the residue ..." The correct term in numerical context is "residual". In chemistry, "residue" might be used for remains of substances.

This is now corrected.