Response to reviewer 1

We thank the Reviewer for carefully reviewing our manuscript and providing insightful comments. Below we address each comment point by point. For clarity we mark the reviewer comment in **blue**, our answers in **black**, and changes to the manuscript in **red**. Page and line numbers (in **black**) in our replies refer to the clean revised manuscript (without tracked changes), and **green** line/page numbers refer to this response.

Summary:

This work accomplishes a cross-comparison of several data-reduction analysis techniques. The analysis is performed on a chamber experiment. Car exhaust was directly sampled into an environmental chamber. A-pinene was also added to the chamber. The mixed car exhaust/pinene was then aged via OH-initiated photooxidation. Two instruments were used, a PTR-ToF MS and an AMS. Data were then analyzed using principal component analysis, positive matrix factorization, exploratory factor analysis, clustering, and non-negative matrix factorization. The resulting simplifications are compared in terms of the ability to reconstruct the original data set (residual), number of factors/groups required to explain data variability, time-series behavior of the factors, and chemical composition of the factors. The authors find that the preferred number of factors is roughly similar regardless of technique. Some factors (or groupings) are generally consistent (in terms of time-series behavior and composition) regardless of technique. Some techniques, particularly PCA and EFA were found to be difficult to interpret and not as useful. The clustering method was not useful for AMS data.

Major comments:

1) The pre-light mixing period is a substantial fraction of the whole experiment- is this interesting? The inclusion of this time period seems to have significant impact on the algorithm results.

One of our aims was to investigate the robustness of the different SDRTs with a data set having large changes in the concentration. The mixing period was important to include in the analysis as the step change when the photochemistry is started in the chamber is a good test scenario because we know the reason for the change and the exact time. Also, the direction of change is known. Please see our reply to specific comment ("page 7 line 16") in page 4.

2) Mostly this paper seems to show that various different algorithms distinguish similarly between primary and secondary VOCs. Is this a useful reduction of chamber data, to group all secondary VOCs into one or two blocs? From Table 1 it seems that the two matrix factorization techniques result in perhaps three oxidized factors, but this is hardly discussed in the text. It is not clear to me if the three oxidized factors are consistent between the two techniques. It would be helpful to have more interpretation of these factors, especially if it were supported by a more detailed connection to the chemistry of the system.
NMF also seems to result in some mixing of primary and secondary emissions (FN4) which is probably unphysical.

We want to point out that the PTR-MS mainly provides information on the precursors and not the heavily oxygenated compounds created in their oxidation. Thus, the data set is far from giving a full picture of the chemistry in the chamber during photo oxidation (see also our reply to specific comment "Page 24 line 13-14" in page 10). Separating primary emissions and grouping the reaction products into three groups without adding any additional information about the ongoing reactions is the first step for understanding the chemical processes in these experiments. For a more detailed study, gas phase data from different mass spectra that cover a wider range of oxidation products

(e.g. PTR-MS and I^- CIMS) need to be combined. But before this can be attempted, we wanted to investigate the performance of the available SDTRs which to our knowledge has not been performed in such detail yet. However, we added the factor contributions as separate spectra to the SI section S5 to clarify the similarity of the acquired results of the different SDRTs used (Figures S21 and S22) and added a few more comments there in addition to the more "mathematical" comparison of the factors by contrast angle.

S5, page 13, lines 16-20
[…] Figure S21 and S22 show the factor contributions in separate panels for each factor for the main results from gas phase data (PTR-MS). Despite the fundamental differences between these methods, the relative factor contributions clearly show similarity. Even PAM, that is relatively different when compared to other SDRTs as it only assigns one m/z into one factor, the same ions are enhanced when compared to other SDRTs. […]

3) The way the mass spectra (chemical composition of groupings) are presented is very difficult to interpret and compare. From Table 1 the authors have assigned a consistent identity to factors resulting from each technique. Can you show a direct comparison of "Factor 2" for example, perhaps by plotting one mass spectra against the other so that it is easy to see which VOCs are similarly enhanced, and which may be different?

We now added figures with the factor mass spectra in separate panels in the SI material section S5 for easier comparison (Figures S21 and S22). In addition, the contrast angle -procedure introduced in SI section S5 shows a direct mathematical way to compare the factor spectra between the different SDRTs.

Specific comments:

Page 3, line 3-4: I disagree with this statement, "PMF was originally developed for field measurement datasets where real changes in factors are expected to be much slower than e.g. the noise in the data." In the ambient environment, VOC composition and concentrations can actually change very quickly (on the order of seconds), especially when plumes of highly-concentrated primary emissions are intercepted. The abrupt changes in conditions during a chamber experiment therefore do not present a special challenge to PMF, compared to ambient measurements.

We agree with the reviewer that also in ambient conditions, the changes can be fast depending on the environment. Hence our statement was misleading. We wanted to highlight here with our statement, that most of the PMF applications, so far, have included ambient measurement data. We have now reworded our claim into:

page 3, lines 5-8
[…] The special conditions in lab experiments (sharp change at the beginning of experiments, e.g., switching on UV-lights) present an additional test scenario, as PMF has been mostly used for field measurement data sets where the main focus is often in the long-term trends and real changes in factors are often expected to be more subtle, than e.g. the variations in the noise in the data. […]

We also adjusted section 4.1.5 accordingly:
page 19, lines 17-24
[…] This is caused by the used error scheme, where errors are larger for the fast changes in the data (Fig. S4b). In ambient data not measured at instant proximity of strong emission sources, for which PMF is often used, this type of error is beneficial as there the fast changes are more likely to be noise or instrument malfunctions (excluding, for example, sudden primary emission plumes), and

we are more interested of the long-term changes instead. For laboratory data, where large changes are often caused by rapid changes in actual experimental conditions, e.g. due to injecting α-pinene or turning the UV lights on, the static type of error is most likely preferable. Usage of the static error scheme helps to avoid overcorrecting intentional (large) changes in experimental conditions and confusing them with real variation taking place during the experiment and typically being much less pronounced. […]

Page 3, lines 26-29: A few more details are needed here (instead of in the supplement):
What was the typical concentration of total VOC (or of a few key VOC e.g. aromatics)?
What concentration of α-pinene was added?
What was the VOC-NOx ratio, and how was it adjusted?
Why were these specific concentrations of vehicle exhaust VOC and α-pinene chosen?
Were vehicle exhaust and pinene added just at the beginning of the experiment, or were they continuously injected?
Was the chamber continuously refilled (and with clean air or with fresh emissions?) to replace air taken by the mass spectrometers, or did the volume of the chamber decrease over time?

We aimed to keep this description as brief as possible as the prime focus of this manuscript was the comparison of the SRDTs. Also, the experimental conditions for the whole measurement campaign are described and interpreted in detail in Kari et al. (2019). However, we adjusted section 2 to include the requested information, and also the mass resolution details requested by reviewer#2:

page 3 line 31 – page 4 line 4
[…] The exhaust was diluted using a two-stage dilution system and fed into the 29 m^3 collapsible environmental PTFE chamber ILMARI (Leskinen et al., 2015). For the experiment investigated in this study, α-pinene (~ 1 µL, corresponding to 5 ppbV) was injected into the chamber to resemble biogenic VOCs in typical suburban areas in Finland. Atmospherically relevant conditions were simulated by adding O_3 to convert extra NO from vehicle emissions to NO_2 and adding more NO_2 to the chamber if needed. With these additions, atmospherically relevant VOC-to-NOx (~ 7.4 ppbC/ppb) and NO_2-to-NO ratios were achieved to resemble the typical observed level in suburban areas (National Research Council, 1991). […]

page 4, lines 10-22
[…] Volatile organic compounds (VOCs) in the gas phase were monitored with a proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF MS 8000, Ionicon Analytik, Austria, hereafter referred to as PTR-MS). Typical concentration for few example VOCs in the mid-way of the experiment were 2 µm/m^3 for toluene, 0.2 µm/m^3 for TMB (trimethylbenzene) and 1.7 µm/m^3 for $C_4H_4O_3$. Detailed setup, calibration procedure and data analysis of the used high resolution PTR-MS have been explicitly presented in Kari et al., (2019b). In the campaign, the high mass resolution of the instrument (>5000) enabled the determination of the elemental compositions of measured VOCs. The instrumental setting was intended to minimize the fragmentation of most compounds, so the quantitation of the VOCs was possible. The chemical composition of the particle phase of the formed SOA was monitored with a soot particle high resolution aerosol mass spectrometer (SP-AMS, Aerodyne Research Inc., USA, hereafter referred to only as AMS, Onasch et al., 2012). In brief, the SP-AMS was operated at 5 min saving cycles, switching between the electron ionization (EI) mode and SP mode. In EI mode, the V-mode mass spectra were processed to determine the aerosol mass concentration and size distribution. The mass resolution in this mode reaches ~2000. The SP-mode mass spectra were used to obtain black carbon concentration. As the used chamber

was a collapsible bag, the volume of the chamber decreased over time due to the air taken by the instruments. […]

Section 3.1.2: EFA seems very similar to PMF. Could you please explain the major relevant difference(s) between EFA and PMF, and how they would affect the resulting dimensionality reduction?

The main difference between EFA and PMF is that in EFA, the factorization algorithm is applied to the correlation matrix constructed from the data, whereas in PMF, the original data matrix is factorized. This means, that in EFA, the created factors include variables, that have stronger correlation compared to variables from other factors.

In PMF, the error matrix is used as a tool to downweigh noisy (or weak) signals. This can also be used to put emphases on certain parts of the data set (e.g. the part with most change in signal). EFA, however, has other options to discard noisy or insignificant signals (see last paragraph in section 3.1.1).

We have now added a summary section (new section 4.4.) to address and discuss the similarities and differences between the presented SDRTs as requested here and by reviewer#2.

Page 7 line 16 (and elsewhere): At multiple points in the manuscript it is mentioned that the rapid changes associated with lights-on cause problems when implementing the various dimensionality reduction techniques. Would it make more sense to exclude data prior to t=0? Was there a reason this was not done?

For a detailed analysis of such a chamber experiment, the data before the onset of photo chemistry would indeed be omitted. But the main focus of our study was to compare the performance of the different SDRTs with different types of mass spectra data. As the reviewer points out above, rapid changes in composition can also occur in the atmosphere. Thus, it is important to test the response of all SDTRs to such a change. The induced step change when photochemistry is started in the chamber is a good test scenario as we know what caused the change and the exact time. Also, the direction of change is known (precursor getting consumed and products formed).

We added a few sentences to the manuscript to further justify our selection of the data:

section 1, page 3, lines 24-25
[…] Further, we examine the performance of the SDTRs when the data includes large and rapid changes in the composition. […]

section 2, page 5, lines 2-4
[…] In addition, as the main focus of our study was to compare the performance of the different SDRTs with different types of mass spectra, instead of detailed analysis of the chamber experiment, we have also included the pre-mixing period during the α-pinene injection (i.e., $t < 0$) into our analysis. […]

Page 10 lines 28-31: Since Figure 1 relies on a comparison of BIC and SRMR, it would be helpful here to provide more detail on how these two metrics are calculated, what the relevant differences are, and why one may be preferred over the other. What was the purpose of calculating both metrics and why were these particular metrics chosen?

These particular metrics were selected, as they measure slightly different properties of the model. BIC is a comparative measure of fit balancing between increased likelihood of the model by adding parameters and a penalty term for number of parameters. The SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. It is a positively biased measure and that bias is greater for small N and for low df (degrees of freedom) studies.

The definitions and equations of these metrics can be now found from the SI material (new section S3.2), and we added a few sentences to manuscript section 3.3, as requested (page 12, lines 12-15):

[…] These metrics measure slightly different properties of the model. BIC is a comparative measure of the fit, balancing between increased likelihood of the model and a penalty term for number of parameters. The SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. See S3.2 for more details. […]

Equations 11 and 12: There are several errors in these equations which are likely a copying error from Brunet et al. 2004. The authors should check that the actual implementation was done correctly. The equations should read:

$$H_{au} \leftarrow H_{au} \frac{\sum_i W_{ia} X_{iu}/(WH)_{iu}}{\sum_k W_{ka}}$$

$$W_{ia} \leftarrow W_{ia} \frac{\sum_u H_{au} X_{iu}/(WH)_{iu}}{\sum_v H_{av}}.$$

I also suggest here to use "k" as the row index for W (in the denominator term), to avoid confusion with "p" being the factor rank, and for consistency with Lee and Seung, 2001.

We thank the reviewer to pointing this out. But while the formula was incorrect in the manuscript text, all calculations we performed with the correct set of equations. Thus, none of the results for NMF need to be changed. The equation is now corrected, and the index "p" is changed to "k" as suggested. We added a sentence before equation (11) to clarify the connection between k and p:

page 9, line 20
[…] The value of *k* is equivalent to the selected factorization rank *p*. […]

Page 11 lines 29-33: Given that the update functions (11) and (12) are derived from the divergence cost function $D(X||WH)$ (Lee and Seung, 2001, Eq. 3), I suggest that this cost function is monitored as a function of p, analogously to Q/Qexp(p) for PMF. The termination condition for NMF wasn't described in Section 3.1.4, but presumably it is not dependent on p; if this is the case then the divergence of the end solution can be compared for each value of p. The residual sum of squares is not an appropriate metric, as this was not the cost function used for the NMF implementation.

The update functions presented in the manuscript are indeed derived from the Eq. 3 presented in Lee and Seung (2001). We have now modified the manuscript and use the value of the cost function in the last iteration for each p instead of the RSS, as suggested.

We modified the last paragraph of section 3.3, corresponding figures (Figure 2 and Figure 11), and the results (sections 4.1.4 and 4.2.3) accordingly:

section 3.3
page xx, lines xx-xx

[…] In addition, we investigated the cost function that approximates the quality of factorization as a function of the factorization rank p. For the brunet-algorithm that we applied in this study, this cost function is the divergence between data matrix X and the approximation WH (see Eg. (3) in Lee and Seung (2001)). […]

section 4.1.4
page xx, lines xx-xx
[…] Figure 2a shows the divergence of the cost function D(X||WH) and CCC for factorization ranks from 2 to 10 for NMF. The CCC has a first decrease in the values at the rank 4 and the D(X||WH) shows an inflection point around the ranks 4-5. Figure 6 shows the factor time series and total contribution for the NMF with factorization rank 5. Five factors were selected, even though CCC suggest only 4 factors, as […]

section 4.2.3
page xx, lines xx-xx
[…] The D(X||WH) has an inflection point at factorization rank 4 and CCC shows the first decrease in the values with 4 factors, as shown in Fig. 11a. […]

Page 12 line 2: What is meant by "not achieved only by change?"

The purpose of the resampling here was to approximate and reduce the uncertainty in the factorization, based only on one dataset. As the figures show, the results vary significantly, especially when fast changes are taking place. This points out that all methods are rather sensitive for only small changes in the data and thus this kind of sensitivity test is necessary.
In order to avoid confusion by wording, the first sentences of section 3.4 (page 13, lines 20-23) were reformulated to:

[…] When analyzing the datasets, we realized that all of the factorization methods in this study are sensitive to even small changes in the data. In order to cross-validate the calculated factorization and approximate the uncertainty in the factors, 20 resamples of the measurement data were created with bootstrap-type sampling (Efron and Tisbshirani, 1986), i.e., sampling with replacement from the original data. […]

Page 13 line 5: Why not compare the absolute value of the residual?

For PMF and NMF, which follow the matrix equation X = WH + E, where E is the residual matrix, this type of comparison is applied in by calculating the total residuals as a function of time (i.e. summing all variables for each time point in E), see e.g. page 24 lines 11-12 for AMS. However, as described in section 3.4 (see page 14, lines 16-25), this type of residual calculation is not possible for EFA and PCA, as they factorize the correlation matrix and do not create W and H in a similar manner (i.e. simultaneously). Therefore, we can calculate the residuals for EFA/PCA only by reconstructing the correlation matrix (these values are shown in the manuscript, e.g. for EFA in sect. 4.1.1, page 15, line 24), but not the actual data. Thus, the actual residual values between PMF/NMF and PCA/EFA are not comparable. In addition, in PAM, which uses the distance matrix in the "decomposition", the situation is again very different.

Pages 14 and 15: For other researchers which would like to use this paper as a guide, it would be helpful to indicate the range of values that are acceptable. For example Page 14 line 4-5, what value of residual would be considered not acceptable? Page 14 line 32, what is the Kaiser limit and what

range of values are considered "close"? Page 15 line 11 are these considered large or small residuals?

As desirable as such absolute values would be, unfortunately it is not possible to set them in a general manner. Especially for PMF and NMF, there is a strong subjective element in the selection of p values and in the magnitude of acceptable residuals. It may depend on the type of data and the purpose of the analysis what is considered acceptable. The values may differ between analyzing a long ambient data set for the possible SOA sources or types (e.g. OOA vs HOA in an AMS data set) and studying a chamber data set trying to identify chemical processes. Please also note that most of these single value "goodness of fit" parameters are summarized over all observations and variables. It is known that for PMF the overall Q values may be low while Q values for individual variables are not (interactive comment from Paatero to Yan et al., 2016).

The Kaiser limit/criterion refers to the ideal value of 1, and components having lower eigenvalues should be discarded (explained in sect. 3.3, page 12, lines 6-7). What is meant by being "rather close" to this limit, is that the relative decrease of the eigenvalues is getting smaller and smaller with increasing number of components. So even though we reach the Kaiser limit (our eigenvalues get smaller than 1) having as many as 9 components, it is not realistic to have this type of system with that many components. In addition, as mentioned in sect. 3.3 page 12, lines 17-20, this type of tests for the number of components/factors are not meant to be taken as strict rules, but rather give the first insights into the possible dimensions of the investigated system. We determined the most suitable number of dimensions for each data after investigating multiple solutions from various SDRTs.

As explained in sect. 3.4 (page 14, lines 27-29) the theoretical maximum value for the mean and IQR of the absolute residual correlations is 1 (indicating in principle no reconstruction at all) and the minimum 0 (perfect reconstruction). For example, if we would have a correlation coefficient of 0.6, and the absolute residual correlation is 0.0116 (as given the mean value in the line referred for PCA), we over- or underestimate the correlation coefficient by 1.93 % ((0.0116/0.6)*100). Thus, the smaller values we have, the better the reconstruction. Unfortunately, there are no general guidelines, and our main aim here was to compare if the reconstruction differs between EFA and PCA. To give the reader a more concrete way to understand the actual "size" of the errors for EFA/PCA in this study, we added the similar example to sect. 3.4:

page xx, lines xx-xx
[…] For example, for a variable-pair having a correlation coefficient of 0.7, a mean absolute correlation residual of 0.02, and an IQR of 0.04, this would mean the model over- or underestimates the correlation by 2.86% (= (0.02/0.7) * 100). An IQR of 0.15 would mean that 50 % of all variable-pairs with correlation of 0.7 are within 5.7% (= (0.04/0.7) * 100) of the original value of 0.7. […]

Page 14 line 1 and page 15 line 2: Can you show please how it is determined that the additional component is not a new component with different properties but rather a mixture of previous components?

Figure 1 below shows the related factors (FE5 from the 5-factor case and FE2&FE4 from 4-factor case) for EFA. From the time series we can see that the FE5 has a small peak at t = 0, suggesting it has picked up some properties from FE2. For example, m/z 89.06 in FE4 in the 4-factor case (Fig. 2b, contribution of 0.74 in FE4, others < 0.16), is divided more "equally" between the factors when using 5 factors (contribution of 0.17, 0.04, 0.11, 0.42, 0.26 for factors FE1-FE5), thus not being as

well separated. Similarly, some other m/z are more scattered between the factors in the 5-factor case, whereas with 4 factors they are more or less assigned to one, or possibly to 2, factor(s) only. This indicates the addition of the 5th factor does not introduce new properties but rather splits the existing ones into subsets. The reasoning for PCA is similar.

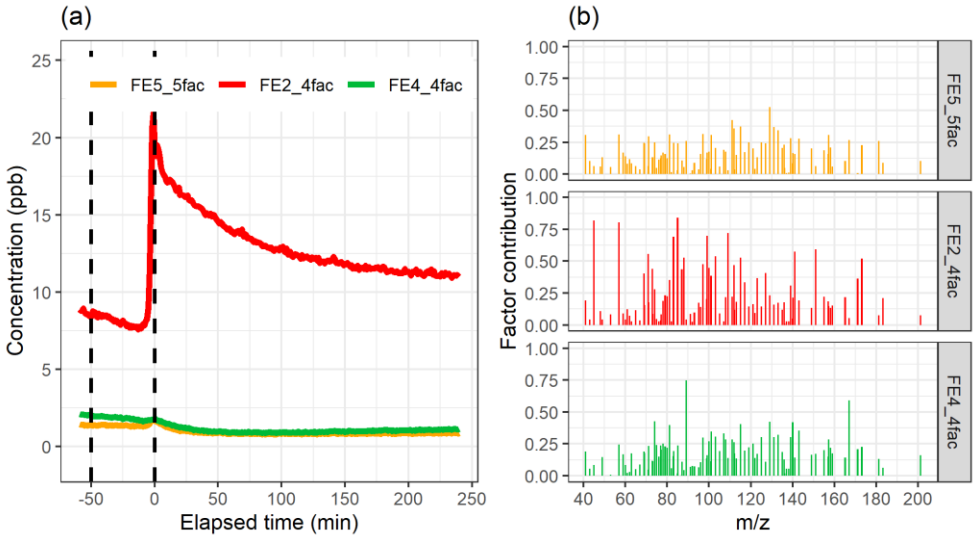

Figure 1 Factors 2 and 4 from the 4-factor case and factor 5 from the 5-factor case. (a) shows the time series and (b) the factor contribution.

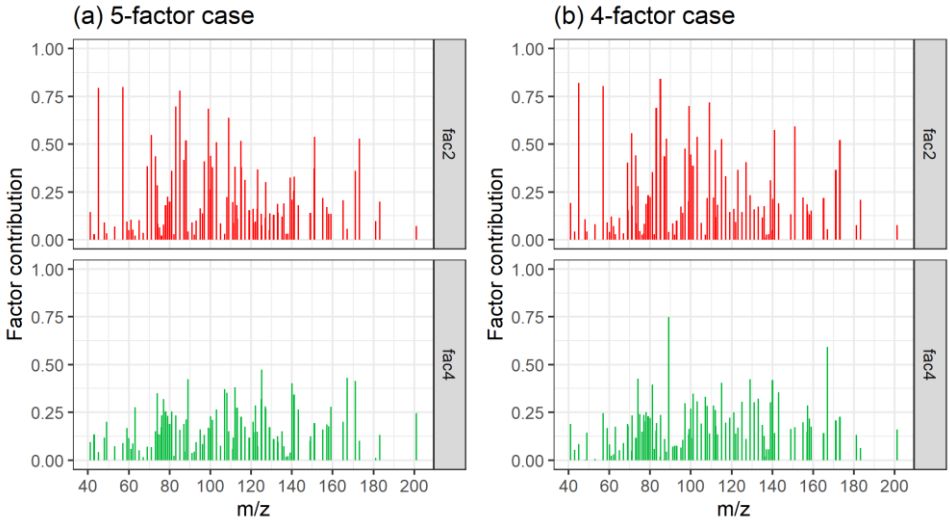

Figure 2 Comparison of the factor contributions for factors 2 and 4 from the (a) 5-factor case and (b) 4-factor case.

Page 17 Lines 16-17, 23-30: The signal following error is essentially introducing a smoothness constraint, which doesn't seem appropriate given that you know there are sharp changes due to experimental conditions. Is it recalculated for each data resampling? Why not resample the error along with the data? Is it possible to split the data into time periods whose start and stop are defined by sample injection, lights on, etc. so that the running standard deviation does not include these sharp changes? Additionally, lines 26-27: Ambient data often has fast changes that are due to real variability. This is one of the reasons why fast online techniques such as PTR-MS are used for ambient measurements, because they allow the observation of these changes.

One of our aims in this study was to see how PMF performs with very different error estimates. Thus, we introduced the static error, that gives equal weight for every time point for each variable in question, and the signal following error, which has larger errors for rapid changes. To split the

data set into sections requires a priori information, which is usually not available especially for ambient data sets. For our laboratory measurements, we naturally have this information, but as we specifically wanted an error estimate, that could possibly account for the large changes, we did not split the data for the error calculation. Instead, we also had the static error giving equal weights for each time point. We reworded the section 4.1.5 to clarify, please see reply to specific ("Page 3, line 3-4") comment in page 2.

For each data resampling we recalculated the error, thus the larger errors were in the "correct" places for each resample too.

Page 22 line 20: This indicates that the error estimation should be revised, or that these compounds should be downweighted. PMF and NMF are extremely similar techniques with the crucial difference being only the inclusion of an error matrix, so it does not seem likely that the difference in performance is due to the size of the dataset. The extremely small values of NMF residuals also seem suspect. The authors should check that residuals for PMF and NMF were calculated in exactly the same way so as to enable the direct quantitative comparison.

One of our aims was to investigate how each of the SDRTs perform with datasets possibly including noisy background ions, thus we did not downweighed them. As we also do not have similar downweighing option for NMF, we omitted this procedure for PMF too to give each method an equal "starting point". In addition, the error estimates we used did have relatively large errors for these signals (noisy, small concentration) and therefore already giving PMF the possibility to not to give so much weight for this type of signal in the first place. As the PMF and NMF results were agreeing nicely with PTR-MS data, we do believe the issue here is either the data size and/or the used brunet-NMF is indeed computationally more efficient/accurate, compared to the version of PMF used here, when handling small data. In addition, PMF performed without any issues for a combined dataset from the same measurement campaign (Kari et al., 2019), indicating the issue in PMF here is indeed related to the small size of the data set.

However, we do agree that to achieve the best chemical interpretation (which was not the main aim of this study) for the factors from PMF, the downweighing is advisable. To justify our decision, we now also applied PMF with the downweighing, and no change in the results was observed. We now state this in the manuscript, and in the SI. In addition, as requested by reviewer#2, we have now included the 2-factor PMF solution for particle phase data to the manuscript and moved the 4-factor solution to the SI. Please see the reply to comment "Figure 13" in page 11.

We have added more justification for omitting the downweighing into section 3.1.3 and S2.3:

section 3.1.3, page 8 line 27 – page 9 line 2

[…] . In contrast to the suggested best practice (Paatero and Hopke, 2003), we did not downweighed any ions in our data sets. This approach was used in order to give each SDRT an equal starting point for the analysis, as e.g. for NMF or PCA similar downweighing is not possible, as we do not have any error estimates to calculate the signal-to-noise ratios in similar manner. However, to avoid misguiding the reader to omit recommended data pre-processing practice for PMF, we also tested PMF with downweighing. This, as expected, did not change our results significantly, but we acknowledge it should be indeed applied if aiming for a more detailed chemical interpretation of the PMF factors. […]

section S2.3, page 5, lines 1-5

[…] However, no removal or downweighing of the variables was applied in the results presented in the main manuscript, as for the weak ions the error values were already in the range of the ion signal itself, and for the bad variables the error was usually way above the signal throughout the ion time series. In addition, the number of weak or bad ions for both AMS and PTR-MS data were rather small. To justify our decision, we did run the PMF analysis with the downweighing. This, as expected, did not change our results or interpretation, and thus those results are omitted. […]

We re-checked the NMF and PMF residuals, and they both are calculated exactly the same way (i.e., multiplying the factorization matrices W and H (or G and F in PMF) to obtain the reconstructed data matrix, and then calculating the difference between the data and the reconstruction for each point in time). The very small residuals in NMF suggest the optimization with the brunet algorithm in NMF is performing relatively better than the PMF algorithm. However, better mathematical performance does not necessarily mean easier/better interpretation of the physical/chemical properties. Brunet-NMF was also selected, as the residuals were smallest when comparing to other available algorithms for NMF. Please note that there also exist other algorithms for PMF, depending on the interface that is used to apply PMF, and those might have different performance (see. e.g. Ramadan et al., 2003).

Page 24 line 13-14: Is this correct? I read this paper as well. I thought they had PTRMS and AMS. Additionally, PTR-TOF is a subset of TOF-CIMS, or?

By definition, every PTR-MS is a chemical ionization instrument. So, a PTR-TOF could be called a H_3O^+ TOF-CIMS. But one has to be careful as these labels are often used to differentiate between different instrument designs. The main difference lies in the design of the sample inlet and the actual ionization region. The PTR3 (used by Koss et al. 2020) has a much improved overall sensitivity over the earlier PTR-TOF due to changes to the electric field guiding the ions through the ionization region. Additionally, the sample inlet was re-designed to better transmit lower volatility compounds (i.e. mostly higher oxygenated) which were previously lost as they are too "sticky". Riva et al. (2019) compared the performance of different chemical ionization instruments. The VOCUS PTR is here similar to the H_3O^+ PTR3. C10 compounds with 3 or more oxygen atoms are barely visible in a PTR-TOF but can be detected with a VOCUS-PTR (Figure 6 in Riva et al. 2019). The Aerodyne I^- CIMS (which has yet another design for the inlet and ion reaction region) mostly detects C10 compounds with 2 to 8 oxygen atoms. Thus, we can conclude that the PTR-TOF used in our study mostly detects the precursors and early oxidation products (with generally low oxygen content) while the instruments used by Koss et al. (PTR3 and I^- CIMS) observe these compounds and also a large fraction of later generation oxidation products and also highly oxygenated compounds (HOM).

To clarify these differences in the instruments we modified our sentence to:

page 27, lines 25-27
[…] In addition, Koss et al. (2020) used gas-phase data from I^--CIMS and PTR3 with NH_4^+ as a reagent ion, which are more sensitive to later generation oxidation products compared to the PTR-MS which we have used here. […]

Figures 3,4: In the plot caption or legend it would be helpful to have a brief description of the interpretation of each factor, e.g. "pinene", "car exhaust", "background". Additionally the display in plot (b) of factor contribution as a function of m/z doesn't add much to the paper; I wouldn't expect the factor contribution to depend on m/z in any particularly meaningful way. Since (I believe) your

PTRMS has multiple peaks resolved at each nominal mass, showing a unit-mass stick spectrum here is also not especially meaningful. If you want to show mass spectra, I strongly suggest to break panel b into 4 separate spectra, one for each factor, so that they can be examined separately.

Note that in these figures we did indeed show the high-resolution data and not the UMR signals. The factor contribution plots (Figures from 3 to 7) were not shown to suggest any correlation of the factors to m/z. Rather, we looked for a way of displaying the factor/compound "compositions" in a way allowing direct comparison of the different methods. Normalizing the factor contribution for each ion was chosen to avoid the very different weighting methods in NMF and PMF, and to be able to compare how much each ion contributes to each factor between all SDRTs. The exact ion mass was used as a unique identifier for all ions as we did not have sum formulas or compound names assigned to all ions. We now added figures with the factor mass spectra in separate panels in the SI material for easier comparison (section S5, renamed to "Contrast angle and factor spectra comparison" Figures S20, S21), and refer to this in the beginning of section 4.1.6. We added description (those in table 1 & 2) to the figures in the manuscript (when applicable), as requested.

section 4.1.6, page 20, lines 1-2
[…] Table 1 summarizes the acquired results from different SDRTs for the gas phase composition data measured with PTR-MS, and Figs. S20 and S21 in SI section S5 show separate factor contributions for each of the SDRTs. […]

Figure 9: Where does isoprene come from in this experiment? Is this more likely to be a hydrocarbon from vehicle exhaust? Cycloalkanes in fossil fuel are known to create PTR ions at C5H9+, see e.g. Yuan et al. Chem Rev. 2017 doi.org/10.1021/acs.chemrev.7b00325.
We used the label "isoprene" for the ion C5H9+ following general naming conventions in the PTR community. But the reviewer is correct in assuming that here the signal at C5H9+ is connected to vehicle emissions and not biogenic emissions. The factor/component contributions of this ion are very similar to trimethylbenzene (TMB). Also, in PAM the C5H9+ is assigned to the cluster containing among others furan and TMB. Even with the short-comings of the clustering approach this strongly indicates the main correlation of C5H9+. We changed the label in Figure 9 and Figure 10 to C5H9+ to avoid the association with biogenic emissions that the label "isoprene" may cause.

Figure 13: The factor time series for the most part do not look realistic. What physical process could lead to the non-smooth behavior and multiple maxima?

Following the suggestion by reviewer#2, we adopted a 2-factor solution from the PMF results to the manuscript instead of the 4-factor solution (which is now moved into SI material). Please see the revised manuscript sections 4.2.4, 4.2.5 and the figures referred there.

Minor/Technical corrections:
Page 2 line 22, "alike" -> "like"
Page 3 line 22, "was" -> "were"
Page 13 line 24, "described in section 0" -> "described in section 3"
Page 15 line 19, "gab" - > "gap"
Page 19 line 27, "much" -> "many"

The suggested corrections were applied to the manuscript

REFERENCES

Kari, E., et al.: Dual effect of anthropogenic emissions on the formation of biogenic SOA, Atmos. Chem. Phys., 19, 15651-15671, 2019

Koss, A. R., et al..: Dimensionality-reduction techniques for complex mass spectrometric datasets: application to laboratory atmospheric organic oxidation experiments, Atmos. Chem. Phys., 20, 1021-1041, 2020.

Paatero, P.: Interactive comment on "Source characterization of Higly Oxidized Multifunctional Compounds in a Boreal Forest Environment using Positive Matrix Factorization" by Chao Yan et al., Atmos. Chem. Phys. Discuss., 2016

Riva., M., et al.: Evaluating the performance of five different chemical ionization techniques for detecting gaseous oxygenated organic species, Atmos. Meas. Tech., 12, 2403–2421, 2019

Lee, D. D., and Seung, H. S.: Algorithms for non-negative matrix factorization, Adv. Neural. Inf. Process. Syst., 13, 556-562, 2001.

Ramadan, Z., et al.: Comparison of Positive Matrix Factorization and Multilinear Engine for the source apportionment of particulate pollutants, Chemometr. Intell. Lab, 666, 15-28, 2003.