

**Interactive comment on “Source apportionment of volatile organic compounds in the north-west Indo–Gangetic Plain using positive matrix factorisation model” by Pallavi et al.**

**Anonymous Referee #2**

Received and published: 9 August 2019

The article titled “Source apportionment of volatile organic compounds in the northwest Indo–Gangetic Plain using positive matrix factorisation model” by Pallavi et al., is generally well written and contains some useful information. VOC source apportionment studies are sparse in India, and this study could encourage more such studies in future which is required to understand the VOCs impact on air quality. However, on many occasions, authors seem to over-interpret the results and have drawn some rather farfetched conclusions. I would recommend publication provided my concerns are being addressed satisfactorily.

Abstract:

Discussion paper

**Reviewer comment:** Numbers can be presented in a better way for ease of reading. Authors can put the percentage contribution of different factors/parameters in the parenthesis beside them.

**Author response:** We thank the anonymous reviewer for this valuable suggestion. The anonymous reviewer #1, suggested to focus the abstract more on the big picture. In response to both comments we have reduced the numbers in the abstract and added the percentage contribution of different factor/parameters in parenthesis beside them. It now reads as follows:

**Changes in the manuscript:**

“In this study we undertook quantitative source apportionment for 32 volatile organic compounds (VOCs) measured at a suburban site in the densely populated North-West Indo-Gangetic Plain using the US EPA PMF 5.0 Model. Six sources were resolved by the PMF model. In descending order of their contribution to the total VOC burden these are “biofuel use and waste disposal” (23.2%), “wheat-residue burning” (22.4%), “cars” (16.2%), “mixed daytime sources” (15.7%), “industrial emissions and solvent use” (11.8%) and “two-wheelers” (8.6%).

Wheat residue burning is the largest contributor to the total ozone formation potential (32.4%). For the emerging contaminant isocyanic acid, photochemical formation from precursors (37%) and wheat residue burning (25%) were the largest contributors to human exposure. Wheat residue burning was also the single largest source of the photochemical precursors of isocyanic acid, namely, formamide, acetamide and propanamide, indicating that this source must be most urgently targeted to reduce human concentration exposure to isocyanic acid in the month of May. Our results highlight that for accurate air quality forecasting and modelling it is essential that emissions are attributed only to the months in which the activity actually occurs. This is important for emissions from crop residue burning (which occur in May and from Mid-October to the end of November).

The SOA formation potential is dominated by “cars” (36.9%) and “two-wheelers” (22.1%), which also jointly account for 47% of the human class I carcinogen benzene in the PMF model. This stands in stark contrast to various emission inventories which estimate the transport sector contribution to the benzene exposure as (~10%) and consider residential biofuel use, agricultural residue burning and industries to be more important benzene sources. Overall it appears that none of the emission inventories represent the regional emissions in an ideal manner. Our PMF solution suggests that transport sector emissions may be

underestimated by GAINSv5.0 and EDGARv4.3.2 and overestimated by REASv2.1, while the combined effect of residential biofuel use and waste disposal emissions as well as the VOC burden associated with solvent use and industrial sources may be overestimated by all emission inventories. Agricultural waste burning emissions of some of the detected compound groups (ketones, aldehydes and acids) are missing in the EDGARv4.3.2 inventory.”

**Reviewer comment: Methods:**

Sec 2.3: Line 20, why 20%? Please explain and incorporate in the manuscript as well.

**Author response:** We chose to assign 20% uncertainty to all masses to avoid a situation where the difference in the assigned uncertainty drives the PMF to dedicate a disproportionate number of a factors towards minimizing Q of a few compounds at the expense of others which may be equally useful as tracers of specific activities. The lower reported uncertainty of some compounds (8-12%) in Sinha et al. 2014 can be primarily attributed to the fact that the instrument has been calibrated with more than one independently sourced calibration gas bottles for that particular compound and the fact that the respective m/z has a good signal to noise ratio and high signals. For some other compounds the sensitivity had to be derived from theory, because no calibration gas is available, hence they carry a larger error.

We have followed the advice of Paatero et al. 2014, *Atmos. Meas. Tech.*, 7, 781–797 and performed sensitivity studies to better understand how errors and their handling can impact the PMF output in our specific case.

In our specific case the fact that toluene has one of the smallest reported measurement errors (8.6 % in Sinha et al. 2014) in combination with the fact that there is a genuine and abundant source with a normalized source profile that is dominated by toluene (tailpipe exhaust of petrol fuelled 2-wheelers) can result into a serious modelling error. This problematic behaviour is observed for this particular dataset, because the second most abundant compound in the same tailpipe exhaust source profile (the xylenes) carries a larger uncertainty (11.8 %) and can be accommodated in other source profiles with a smaller penalty on Q. Most real world traffic contains a mixture of 4-wheelers and 2-wheelers and the ratio of these two vehicle classes in the traffic varies as a function of air mass origin and time of the day. At the same time the benzene/toluene ratio of all aged plumes varies with the photochemical age of the air mass. When all these factors are combined the situation becomes such that while running the model with differential errors the lowest Q for the equation

$$Q = \sum_{i=1}^n \sum_{j=1}^m \left[ \frac{x_{ij} - \sum_{k=1}^p g_{ik} f_{kj}}{u_{ij}} \right]^2$$

is obtained by creating a separate toluene factor and removing toluene from the factor profile of all combustion source profiles. In other words, in this specific dataset assigning correct but different random errors to different m/z triggers a serious modelling error which appears already in a 4 factor solution and is retained through any higher number of factors. Assigning equal random errors to all m/z prevents this modelling error from occurring. Hence, we assigned an error of 20% to all masses even though in reality only few strong m/z ratios (formaldehyde, propyne, styrene and phenol) carry such a large error. The reviewer is, however, correct that this choice of using the largest error for all compounds is somewhat

arbitrary and one could instead use the average uncertainty of all the strong compounds errors (i.e. assigning ~10% uncertainty to all compounds would also prevent the modelling error). The magnitude of the chosen error will impact the magnitude of Q (which will increase by a factor of ~4 when 10% instead of 20% uncertainty is assigned) but will not change the model output as long as equal uncertainty is assigned to all masses. However, considering the accuracy and precision error while initializing the PMF may not actually be the right approach at all, considering that the software treats the errors as random. One could argue that only the precision error should be considered while assigning errors in the software. However, as long as equal uncertainty is assigned to all strong m/z the assigned uncertainty will not change the model output and conclusions.

**Changes in the manuscript:** We have incorporated this reason and the section now reads “All 32 species were assigned a fixed 20% in the uncertainty, which represents the largest uncertainty estimated for strong compounds, to avoid a situation where the difference in the assigned uncertainty drives the PMF to dedicate a separate factor towards minimizing Q of a single compound with low uncertainty (toluene) by taking it out of all other source profiles and opening a separate factor profile containing just a single compound.”

**Reviewer comment:** Line 20, more than 50% of the measured species (18 of 32) are weak, isn't that going to influence the robustness & reliability of the PMF output?

**Author response:** This is definitely going to impact the robustness and reliability of the PMF output in a positive manner. A poor signal to noise ratio indicates that the measured values of a species throughout most of the time series are very close to the detection limit. All instruments have a higher precision error close to the detection limit. This is why the manual recommends assigning masses with low S/N ratio “weak” and we have followed this instruction for all compounds that have a poor S/N ratio and do not show any strong peaks. However, in our opinion S/N ratio should not be blindly used as a criterion to make masses weak. Let us consider the hypothetical scenario of a compound emitted only by a single source impacting the site. There can be a situation where such a source impacts the site only rarely (say less than 5% of the time) but when it does the plume brings a very high concentration of that tracer compound. In such a case that specific tracer could be extremely precious for constraining the rotational ambiguity of the PMF solution, even though its average S/N ratio would be very poor (because 95% of the values in the time series are noise around the detection limit). Hence one always needs to look at every species of the input dataset carefully to assess whether it should be made weak just because of its S/N ratio.

There can be reverse cases of masses with a high S/N ratio (for which the average concentration is always far above the detection limit throughout the time series), which can negatively impact the PMF rotational ambiguity when not labelled as weak. This is the case for all masses with potential isobaric interferences. Let us consider an m/z where one of the compounds falling onto the mass to charge ratio is of pyrogenic origin and the other one a tracer for biogenic emissions or a product of daytime photochemistry and discuss how this will impact the PMF model output depending on whether the species is a strong or weak species. Any peak in the concentration observed can be due to either of the contributors i.e. due to a combustion source alone or due to biogenic emissions/daytime photochemistry alone or due to a mixture of both. The most serious impact of this on the model performance is that it can make resolving the rotational ambiguity difficult and can cause modelling errors. Resolving rotational ambiguity requires that the matrix contains a sufficient number of zero values where a source is totally absent. When two sources with different temporal profiles (night-time combustion and daytime biogenic emission or night-time combustion and

daytime photochemistry) contribute different compounds to the same m/z ratio, zero values are almost absent in that particular column of the matrix. When this column is made “weak” and given a higher uncertainty, other “strong” tracers with genuine zero values and strong peaks that can be attributed to a specific sources define source profiles and this reduces the rotational ambiguity of the model. The “weak” compounds with isobaric interferences are distributed among the source profiles available as per the solution that minimizes Q but they do not define any of the profiles. In our opinion, this is the most appropriate way to treat m/z ratios with potential isobaric interferences. As already described in the supplement and the main text we have made such masses weak in the PMF to improve the quality of the PMF output.

**Changes in the manuscript:** We have added a clarification that this makes the model more robust.

“Designating sources with isobaric interferences as weak is warranted because when two sources with different temporal profiles (night-time combustion and daytime biogenic emission or night-time combustion and daytime photochemistry) could potentially contribute different compounds to the same m/z ratio, zero values are almost absent in that particular column of the matrix and the tracer is affected by additional uncertainty not appropriately expressed by merely looking at the instrumental measurement error and the signal to noise ratio. When this column is made “weak” and given a higher uncertainty, other “strong” tracers, representing a single compound, define source profiles and this reduces the rotational ambiguity of the model. The “weak” compounds with isobaric interferences tend to be distributed among the source profiles available as per the solution that minimizes Q but they do not define any of the profiles.”

**Reviewer comment:** Line 24, Why the authors chose to remove missing values instead of replacing them with some other values as mentioned in the literature? Is this the standard practice? what is the % of missing values in the total sampled points?

**Author response:** Replacing missing values with the median while assigning it a greater uncertainty in the PMF helps a lot when the PMF is run with different tracers measured with different sets of instruments, each of which has a different set of missing values. The default setting of the EPA PMF model described in the literature was developed for such a scenario. To illustrate let us consider using a dataset with data from 10 different instruments each of which individually has less than 10% downtime in a situation where unfortunately problems rotate. In that scenario for > 50% of the data points a few variables would be missing. Using the exclude missing value option in such a case would mean throwing out more than half of the dataset as the model removes all lines (=points in time) with a missing value, even if only a single column has a missing value. In such cases lines with missing values still have a lot of data (because one instrument is down the other instruments are running) and only a small subset of species is missing for each point to be potentially excluded. Hence the default model setup suggests filling in missing values with the median of the time series while assigning a greater uncertainty to that point.

However, we are dealing with measurements of a single instrument and <5% of missing values in a month. Filling missing values does not improve the quality of the model output in our case. When the PTR-MS is undergoing calibration or ion source cleaning, there is no ambient data at all available for the gap. Hence the gap filling is unnecessary. It serves no purpose and would hardly affect the model output as all parameters would be filled in with their respective median for that particular point in time.

**Changes in the manuscript:** added ( <5% ) after “missing values”

**Reviewer comment:** Sec 3.3

Line 1-3,  $R = 0.4$  is not a good correlation, at best it can be termed as moderate. Please rewrite the explanation on why fire count is the best tracer for factor 2.

**Author response:** With best we simply meant that the  $R$  was better than that of other potential independent tracers such as  $\text{NO}_x$  (which correlated more with transport sector emissions) and  $\text{CO}$  (which correlated best with the more regular open burning activities such as biofuel use and waste disposal). However, we understand now, that this could be misunderstood and have revised the sentence.

**Changes in the manuscript:** “Figure 3 shows that the factor profile correlates most strongly with flaming wheat residue burning ( $R=0.9$ ). The average contribution of wheat residue burning to the total NMVOC mass at the receptor site and the daily fire counts over the NW-IGP show a moderate cross correlation of  $R=0.4$  with a lag of 2 days (Figure 5).”

**Reviewer comment:** Sec 3.7

Line 7, in  $\text{PM}_{2.5}$ , 2.5 should be subscript

**Author response:** Done this section has become Sec 3.1 in the restructured manuscript

**Reviewer comment:** Line 7, I don't think the way SOA being calculated enable the authors to make such strong quantitative assertion about the SOA contribution to  $\text{PM}_{2.5}$  in Mohali. At best, the adopted method can provide a qualitative and comparative assessment of SOA production efficiency among different PMF factors. I would suggest to remove or modify. line 6-8 to reflect this.

**Author response:** We understand that the SOA formation potential as calculated has its limitations and depends on the  $\text{NO}_x$  regime and may even show a non-linear dependence on VOC and  $\text{NO}_x$  concentration for some compounds (Xu et al., 2015, Atmospheric Environment 101, 217-225). However, we believe that providing a boundary condition may be useful. We have modified lines 6-8. We now explicitly mention that we applied the “SOA yields for the low  $\text{NO}_x$  regime” in the relevant section of the materials and methods, which the reviewer #1 asked to extend and in this section. We also now put the calculated SOA formation potential (i.e. the  $\sim 17 \mu\text{g}/\text{m}^3$ ) in brackets behind its first mention in the paragraph as we believe that despite all short comings this number provides an important perspective. In support, we have also qualified the estimate by citing more studies.

**Changes in the manuscript:**

“While the calculated SOA formation potential particularly from transport sector emissions (Ensberg et al., 2014) and aromatic compounds (Li et al., 2017, Li et al., 2018) is affected by large uncertainties and may depend in a non-linear fashion on  $\text{NO}_x$  and VOC concentrations (Xu et al, 2015) our calculated SOA formation potential seem to indicate that SOA formation could contribute significantly to the overall  $\text{PM}_{2.5}$  burden ( $104 \mu\text{g}/\text{m}^3$ ).

**Reviewer comment:** Sec 3.8

I am not sure about the utility or purpose of this section. Are authors trying to use this comparison as another tool for PMF results validation? Or to suggest which inventory is better? Every emission inventory is developed based on some underlying assumptions and approximations. I would rather be very surprised if a single site based study can reproduce or match the emission inventory values. It is quite expected that differences will be there and even a perfect match doesn't necessarily validate emission inventories or the PMF results, especially in a complex source environment as in India. Several assumptive statements were made to explain the mismatch/less match between PMF results and emission inventory values. So, based on this comparison one can't really assert which inventory is better or more representative than others. Authors should remove or rephrase the section to reflect those concerns.

**Author response:** This section is meant to identify which of the currently used emission inventory represents the regional sources best. This is the major motivation behind any source-receptor modelling study. The anonymous reviewer is correct that every emission inventory is developed based on some underlying assumptions and approximations. Some of these assumptions and approximations can be awfully wrong and the purpose of source receptor modelling studies is to point out such discrepancies. For example, several PMF based source receptor modelling studies in Europe found that the solvent source could be overestimated in most emission inventories while the transport sector may be underestimated (Gaimoz et al. 2011 *Environ. Chem.* **2011**, 8, 91–103., Niedojadlo et al. 2007 *Atmos. Environ.* **2007**, 41, 7108., Lanz, et al. 2008 *Atmos. Chem. Phys.* **2008**, 8, 2313.). Such discrepancies between inventories and source receptor modelling results which got replicated in several studies in different countries ultimately triggered a new series of road tunnel studies and on-road emission factor measurements to re-evaluate the assumptions and approximations used while building the transport sector emission inventories. These efforts not only resulted in a significant upward revision of transport sector emission estimates for NMVOCs while shifting from the EDGARv4.2 inventory for the year 2010 to EDGARv4.3.2 for the year 2012 but also exposed the diesel cheat software that switched off pollution control devices when the vehicles were driving on the road. Therefore, we believe that reality checks based on source receptor modelling of ambient data perform an important role. Their potential significance is even larger in a complex environment where activity data for informal sector industries and activities that officially don't happen (e.g. open waste burning) is hard to obtain while at the same time proper emission factor measurements for many sources are lacking. We, therefore, insist that this section is important to retain.

The validation of PMF results in all prior studies has been performed by cross correlating one or several columns of one of the two matrices produced during the factor decomposition, namely the factor contribution matrix, with independent variable in the form of the time series of compounds that were not used to drive the model. We performed this cross verification step for all six identified factors using the species NO<sub>y</sub> (cars & 2-wheelers), SO<sub>2</sub> (industrial emissions), CO (domestic fuel usage and waste disposal), fire counts (wheat residue burning) and O<sub>3</sub> (mixed daytime factor). However, our study, to the best of our knowledge, is the first one to add an additional verification step in the form of grab samples collected at the source which were used to independently verify the factor profiles (i.e. the second matrix) that the PMF model created during the matrix decomposition. This validation was performed using samples collected at the source for five of the six factor profiles (wheat residue burning, domestic fuel usage and waste disposal, industrial emissions and solvent use, car tailpipe emissions, and 2-wheeler tailpipe emissions). It appears that this validation procedure was not described clearly enough, hence we have now added a section describing the procedure to the materials and methods section. We added a section 2.4 Validation of the PMF output. Some of the text in this section has been shifted from section 3.1 to this section. We also added a new reference since the source signature of brick kilns has recently been published and has now been included.

We have also switched to a new version of EDGAR (v4.3.2) which has recently become available and have added more depth to the comparison by looking at individual compound classes of the speciated emission inventory rather than just at the total VOC mass. We have also removed some of the quantitative statements.

### **Changes in the manuscript:**

#### **“2.4 Validation of the PMF output**

The PMF generates two matrices from the intrinsic variability in the dataset. A factor contribution matrix and a factor profile matrix.

Traditionally the PMF output has been validated by cross-correlating the factor contribution matrix with independent tracers which were not used to initialize the model, but are considered useful tracers for the respective source (Brown et al. 2007, Leuchner et al. 2011, Bon et al. 2011, Gaimoz et al. 2011, Sarkar et al. 2017). We perform this validation step for all six source factors resolved by the PMF model. These were identified as “biofuel use and waste disposal”, “wheat-residue burning”, “four-wheelers”, “two-wheelers”, “industrial emissions and solvent use” and “mixed daytime sources”, respectively. The factor contribution for 4-wheelers ( $R=0.7$ ) and 2-wheelers ( $R=0.6$ ) correlated best with the independent tracer  $\text{NO}_y$  which is considered to be a vehicular exhaust marker (Ramanathan et al., 1985). The factor contribution of the domestic fuel usage and waste disposal factor correlated best with the independent tracer  $\text{CO}$  ( $R=0.9$ ), a proxy for inefficient combustion, while the factor contribution of the industrial emission factor correlated best with the independent tracer  $\text{SO}_2$ . ( $R=0.6$ ). The wheat residue burning factor days showed a moderate cross correlation with MODIS fire counts with an  $R=0.4$  and a lag of 2 days. Ozone ( $R=0.8$ ) was the best independent tracer for the mixed daytime factor.

However, our study goes one step further than all previous studies in validating the PMF output. For 5 out of 6 factors we validated the factor profiles generated by the PMF model (Figure 3) against grab samples collected at the source. Factor profiles were cross-correlated with the fingerprints of source samples collected from a number of potential sources including wheat residue fires (Chandra et al., 2017; Kumar et al., 2018), ambient air samples from a busy traffic junction (Chandra et al., 2017) and an industrial area (this study), tail-pipe exhaust of various vehicles (this study), waste burning (Sharma et al., 2019), leaf litter burning (this study) and domestic biofuel use (Stockwell et al., 2016) and brick kilns (Zhong et al., 2019) to identify the sources. Figure 3 shows the factor profiles obtained from the PMF run (in dark blue), the percentage of each species explained by the respective PMF factor (red squares) and the source profiles of those sources which best matched the factor profile (in various colors as indicated in the legend). The factor profile of residential fuel usage and waste disposal correlated most strongly with the measured VOC source speciation profiles of domestic cooking ( $R=0.8$ ), leaf-litter burning ( $R=0.7$ ) and smoldering garbage fires ( $R=0.6$ ), the wheat residue burning factor with flaming wheat residue burning ( $R=0.9$ ), the 4-wheeler factor with petrol-fueled cars ( $R=0.5$ ) and urban traffic junction grab samples ( $R=0.8$ ) and the 2-wheeler factor with the tailpipe exhaust of petrol-fuelled 4-stroke two-wheelers ( $R=0.6$ ). The industrial emissions correlated moderately with the source profile of brick kilns ( $R=0.5$ ) and ambient air samples collected in an industrial area (0.6). For mixed daytime no source profile sampling is possible. “

### **The revised section 3.8 now reads:**

“Figure 8 shows pie charts depicting the contribution of different sectors to the total VOC mass burden for the emission inventories and our PMF output. Biofuel use and waste disposal were responsible for 28.1% of the mass in our PMF but 39%, 44% and 42% of the mass in EDGARv4.3.2, GAINS and REASv2.1 respectively. The contribution of crop residue burning (27.1%) to the VOC mass in the month of May would be highly underestimated by both GAINS (7%) and EDGARv4.3.2 (4.76%) if the annual emissions are attributed equally to all months of the year. However, if both emission inventories would attribute their annual crop residue burning emissions over the region only to the 2.5 months when crop residue burning actually occurs (middle of October to end of November and May), these emission inventories could be reconciled with the PMF solution, as emissions in May would amount to

26.5% and 23% GAINS and EDGARv4.3.2, respectively as shown in Figure 8. At the same time the percentage share of domestic fuel use and waste disposal would drop to 32% and 35% in EDGARv4.3.2 and GAINS, respectively and the contribution of industrial emissions and solvent use would drop to 18% in GAINS and 30% in EDGAR, respectively. Our PMF solution indicates that industrial emissions and solvent usage (14.3%) are currently overestimated in all emission inventories but are closest to GAINS (540 Gg y<sup>-1</sup>, 18%) for industrial emissions and solvent use. For domestic biofuel use and waste disposal EDGARv4.3.2 (968 Gg y<sup>-1</sup>, 32%) appears to agree best with our PMF solution. For wheat residue burning GAINS agrees well with our PMF output, while the agricultural waste burning emissions of some of the detected compound groups (ketones, aldehydes and acids) appear to be missing in the EDGARv4.3.2 inventory. Our PMF solution for road transport sector emissions (30.5%) lies in between the estimates of GAINS (558 Gg y<sup>-1</sup>, 24%) and REAS (1230 Gg y<sup>-1</sup>, 36.2%), possibly, because not all pre-2000 super-emitters for which the 20-year vehicle lifetime has been exceeded have been retired as planned. Overall it appears that none of the emission inventories is ideal at the present. Our PMF solution suggests that transport sector emissions may be underestimated by GAINS and EDGARv4.3.2, while the combined effect of residential biofuel use and waste disposal emissions as well as the VOC burden associated with solvent use may be overestimated by all emission inventories. Similar results have been reported previously. Sarkar and co-workers (Sarkar et al., 2017) reported an underestimation of transport sector emissions for the REAS and EDGAR emission inventory for the Kathmandu valley in Nepal and an overestimation of the residential biofuel use and waste disposal source in all emission inventories, while Gaimoz and co-workers (Gaimoz et al., 2011) reported an overestimation of the VOC emissions from solvent use in Paris.”

**Reviewer comment:** Figures:

I want to see Q/Q<sub>exp</sub> plot in SI.

done

**Reviewer comment:** Fig. 7: Why the evening peaks in Car & Two-wheeler contributions are significantly more pronounced than morning hours?

**Author response:** Because more traffic activity happens in shorter timespan in the evening and in addition the boundary layer is more shallow during that evening rush hour. In the mornings various activities associated with vehicular movement are spread out over a longer time period (e.g. schools tend to start earlier (8 am) than offices (9 am) and most markets open only at 11 am). In addition, the sun rises before 6 am and hence peak morning traffic occurs after the daytime boundary layer has been well established. This results in greater dilution and lower mixing ratios. We have included the explanation in the revised manuscript

**Changes in the manuscript:**