Supplement of Atmos. Chem. Phys. Discuss., 15, 6077–6124, 2015
http://www.atmos-chem-phys-discuss.net/15/6077/2015/
doi:10.5194/acpd-15-6077-2015-supplement
© Author(s) 2015. CC Attribution 3.0 License.

Atmospheric Chemistry and Physics Discussions

Supplement of

Source apportionment of methane and nitrous oxide in California's San Joaquin Valley at CalNex 2010 via positive matrix factorization

A. Guha et al.

Correspondence to: A. Guha (abhinavguha@berkeley.edu)

S1. Determination of number of PMF factors

In PMF, the choice of modeled factors in the solution is made on the basis of a qualitative judgment and remains the most critical step in the interpretation of results (Ulbrich et al., 2009). A number of metrics aid in this decision making process. One of these is the *Q*-value which represents the total sum of the squares of scaled residuals. If the assumptions of bilinear model are appropriate and the errors in the input data have been properly estimated such that each reproduced data point is fit to within its estimated error value, then, Q/Q_{exp} should be ~ 1. Values of $Q/Q_{exp} >> 1$ indicate underestimation of the errors or inability of the PMF solution to explain a significant portion of the variability in factor profiles as the modeled sum of contributions of the chosen number of factors *p*. Hence, the estimated Q/Q_{exp} is explored as a function of the number of factors in order to determine the best modeled representation. Addition of factors (increasing *p*) adds more degrees of freedom to enable a better fitting of the data and decreases the value of Q/Q_{exp} and if the decrease is large enough, it implies that the additional factor has explained significantly more of the variation in the data and hence the added factor is real (Paatero and Tapper, 1993). The % decrease in Q/Q_{exp} values or slope of the curve at each step increase in *p* should be used as a criterion in determining the '*best*' number of factors in the solution. One should be careful and wise in not choosing a PMF solution solely based on Q/Q_{exp} values. Choosing too many factors in a PMF solution may make a real factor further dissociate into two or more non-existing sources. This phenomenon is known as *splitting* and discussed by Ulbrich et al. (2009). Hence rejecting a solution involving *splitting* behavior in factors should serve as a criterion while narrowing down on a PMF solution. Additional factors may also be non-unique with contributions from all major classes of compounds thus rendering the apportionment of the factor useless and should be used as a criterion to reject solutions. On the other hand, choosing too few factors will combine sources with different emission characteristics together to produce a single factor and hence yield a solution that will be difficult to interpret (Hopke, 2000). In the end, the ability to interpret a FP and issue it a name of a source category, based on *a priori* knowledge of the chemical compositional profile of the source, remains a qualitative but a necessary step in identification of the final PMF solution. As per P. Paatero (the creator of the PMF technique), this subjectivity is a part of the PMF process and should be reported in scientific publications (Ulbrich et al., 2009).

Figure S1.a shows the variation of Q/Q_{exp} values with increasing *p* for solutions including up to 10 factors at FPEAK = 0 (discussed in Section S.2). The Q/Q_{exp} values show a steep decrease from *p* = 1 to 5 (> 10 % drop at each step) but then gradually the decrease becomes steady and is less than 10 % at each

step ($p > 5$) indicating the *optimum* solution is at $p > 5$. PMF solutions for all cases in Figure S1.a (1 to 10 factors) were examined. A 7-factor solution was found to be the most suitable in explaining the variability in the data, yielding factor profiles which are unique and well-distinguishable from each other. The Q/Q_{exp} value at $p = 7$ (FPEAK = 0) is 4.3 which suggests that the errors are either somewhat underestimated, there are a fair number of *weak* data points (missing and BDL) and that the variability in the dataset cannot be modeled better than this due to physical parameters at the site. In this study, the slightly higher Q/Q_{exp} value can be attributed to limitations in the modeling ability which arises due to a lack of strong contrast in the time trends of species during the nighttime as all primary emissions accumulate in a shallow boundary layer and there is minimal chemical processing of the air parcels. The same was observed made by Bon et al. (2011) in their Mexico City study.

Besides the chosen 7-factor solution, other PMF solutions have been evaluated, and figures of factor profiles for a 6-factor PMF solution (FPEAK = 0) and an 8-factor PMF solution (FPEAK = 0) are provided in this supplement (Figures S2 and S3, respectively). On comparing the FP plots of various PMF solutions, we find that the gray colored factor in Figure S2 of the 6-factor solution does not resolve/separate the urban (green) and nighttime biogenics (navy blue) sources seen in the 7-factor solution (FigureS4). The chemical profile of this factor seems *mixed* with no major contribution from any specific source marker but instead has minor source contributions from almost all the tracers included in the PMF analysis and is thus indistinguishable. On the other hand, the agricultural soil management factor from the 7-factor solution (Figure S4) seems to be *split* into two separate factors in the 8-factor solution (gray and brown factors in Figure S3). Neither of the two split factors resembles any particular source category and do not provide any additional insight into the data. The diurnal profiles of the two split factors (not shown) look identical giving further evidence of the "factor splitting" phenomenon.

S2. Rotation of factors

The bilinear PMF analysis has rotational ambiguity and is not mathematically unique. The constraint of non-negativity reduces the rotational freedom in the system but does not generally produce a unique solution. There may be potentially infinite linear transformations, better known as "rotations", that can reduce the rotational freedom by introducing zero values in the factor mass profile (**F**) and time series (**G**) and can force the solution to produce an identical fit to the data (*Ulbrich et al.*, 2009), such that:

$$\mathbf{GF} = \mathbf{GTT^{-1}F}$$, where **T** = transformation matrix, $\mathbf{T^{-1}}$ = inverse of **T** (1)

In the PMF2 algorithm, the rotated factor product is allowed to differ slightly from the product of the **G** and **F** matrices ($\mathbf{GF} \approx \mathbf{GTT}^{-1}\mathbf{F}$) on account of the non-negative forcing of the matrices in order to produce "distorted" rotations which may lead to a slightly worse but acceptable fit to the data with similar but higher values of *Q* and potentially yield more physically realistic solutions (Paatero et al., 2002). After the case with the *best* number of factors has been established, a subset of the "distorted" linear transformations of the solution can be explored using the FPEAK parameter. Positive FPEAK values force the routine to add one **G** column vector to another and subtract the corresponding **F** row vectors from each other while negative FPEAK values explore the reverse scenario (Hopke, 2000; Paatero, 1997). Zero values in the **F** and **G** matrices (no rotations) will limit subtractions in the matrices owing to the non-negativity constraint and thus limit the scope of solutions. Only "rotations" for which the *Q*-value is not significantly greater than the central case (FPEAK = 0) are considered. Prior literature suggests not considering rotations for a FPEAK case in which the Q/Q_{exp} value shows an increase of 10 % or more above its minimum value (usually $Q_{FPEAK=0}$)(Paatero et al., 2002). The rotation procedure produces, for each FPEAK, new rotated matrices **GT** and $\mathbf{T}^{-1}\mathbf{F}$ that represents time series and factors respectively, that may appear to be closer to physically real source profiles than **G** and **F**.

A narrow FPEAK range is more appropriate in cases where Q/Q_{exp} value for (*p*-*1*)-factor solution (FPEAK = 0) is less than 10 % higher than the Q/Q_{exp} value for the corresponding case in the *p*-factor solution. This is true in the current case of 6 versus 7-factor solution (Figure S1.a). Figure S1.b plots the variation in Q/Q_{exp} values with respect to the FPEAK parameter for the 7-factor solution over a range of FPEAKS from -3 to +3 in increments of 0.2 units. Solutions with narrower FPEAK range that give an increase of 1 % over the minimum Q/Q_{exp} value have been investigated for acceptable PMF fits (Ulbrich, 2009). The FPEAK range that meets the 1 % criterion is -1.6 to 0.4 (Figure S1.b). The standard deviation over this FPEAK range is the estimated error in mass fraction of each tracer in each of the seven factors. We follow the guidelines in (Comero et al., 2009; Paatero et al., 2002) about behavior of Q/Q_{exp} with change in the FPEAK parameter and determine the physical plausibility of the all the factor profiles at each FPEAK within the shortlisted range and choose the *best* fit to the data at FPEAK = -1.0.

S3. Uncertainty estimates of solution

Bootstrapping in PMF is a quantitative technique that addresses the difficult topic of evaluating the stability and statistical uncertainty in a candidate PMF solution (Norris et al., 2008; Ulbrich et al., 2009). In the bootstrapping procedure, the PET creates a new data set by randomly selecting non-overlapping blocks of consecutive samples. The new data set has the same dimensions as the original

data set. PMF is then applied to this new data set. In every run, each bootstrap factor is assigned to a base run factor by comparing the contributions of each factor and assigning it to the one with highest correlation. At the end of the user-specified number of iterations, bootstrapping statistics for all the runs are generated in the PET which include average and 1σ values for each fractional component and sample mass in the FP and TS, respectively. The results of bootstrapping inform the analyst of the robustness of the factor profiles chosen in the base run.

Bootstrapping was applied to the base run (7-factor solution, FPEAK = +0.6, SEED = 0) with 100 runs. The FP of the seven factor profiles with their bootstrapping averages and standard deviation range is plotted in Figure S4. The fractional contributions to a source factor from tracers that occur in relatively high proportions in the base run (indicated by colored bars) is quite similar to the averages over the 100 bootstrapping runs (dots) in all the seven factors. The plot also shows the uncertainty in each mass fraction represented by the standard deviation (1σ) of these averages (indicated by whiskers about the dots). For e.g. the uncertainty in the normalized fractional proportion of CH_4 in the dairy and livestock source factor is 29 % while the uncertainty in PMF-derived N_2O fraction of agricultural and soil management factor is 70 %. The overall averaged mass fraction of compounds in all factors from the bootstrapping runs is similar to the factors from the base run (Fig. S4) suggesting that the chemical profile of each factor is reproduced consistently in the bootstrapping runs. Within a factor, the uncertainties in individual mass fractions are lower for major constituents while minor constituents have larger uncertainties. The uncertainties of the tracers that occur in relatively minor proportions in each source factor can be high which is a known limitation as PMF is *weak* in its partitioning of the mixing ratio signals due to collocated sources and artifacts arising due to meteorology (like strong daytime mixing), and hence suffers from the '*mixing*' and '*splitting'* phenomena (see Section S1). We conclude that the bootstrapping results show a robust 7-factor PMF solution with reasonable uncertainties for tracers that are major contributors to a source factor. The uncertainties also confirm that PMF analysis does not yield a unique solution but rather presents a range of possible combinations of mass fractions of compounds, all with low Q/Q_{exp} ratios. The uncertainties generated in the factor profile and the time series from the bootstrapping runs are propagated to determine the uncertainties in the relative apportionment of the trace gas distribution by source type (in Figure 7).

References

Bon, D. M., Ulbrich, I. M., de Gouw, J. a., Warneke, C., Kuster, W. C., Alexander, M. L., Baker, a., Beyersdorf, a. J., Blake, D., Fall, R., Jimenez, J. L., Herndon, S. C., Huey, L. G., Knighton, W. B., Ortega, J., Springston, S. and Vargas, O.: Measurements of volatile organic compounds at a suburban ground site (T1) in Mexico City during the MILAGRO 2006 campaign: measurement comparison, emission ratios, and source attribution, Atmos. Chem. Phys., 11(6), 2399–2421, doi:10.5194/acp-11-2399-2011, 2011.

Comero, S., Capitani, L., and Gawlik, B. M.: Positive Matrix Factorization - An introduction to the chemometric evaluation of environmental monitoring data using PMF, JRC Scientific and Technical Reports, EUR 23946 EN-2009.

Hopke, P.: A guide to positive matrix factorization, Work. UNMIX PMF as Appl. to PM2, 1–16 [online] Available from: ftp://128.153.5.141/users/h/o/hopkepk/IAEA/PMF-Guidance.pdf (Accessed 13 March 2013), 2000.

Norris, G., Vedantham, R., Wade, K. and Brown, S.: EPA positive matrix factorization (PMF) 3.0 fundamentals & user guide, Prep. US … [online] Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:EPA+Positive+Matrix+Factorization+(PMF)+3.0+Fundamentals+&+User+Guide#0 (Accessed 9 April 2013), 2008.

Paatero, P.: Least squares formulation of robust non-negative factor analysis, Chemom. Intell. Lab. Syst., 37(1), 23–35, doi:10.1016/S0169-7439(96)00044-5, 1997.

Paatero, P., Hopke, P. K., Song, X.-H. and Ramadan, Z.: Understanding and controlling rotations in factor analytic models, Chemom. Intell. Lab. Syst., 60(1-2), 253–264, doi:10.1016/S0169-7439(01)00200-3, 2002.

Paatero, P. and Tapper, U.: Analysis of different modes of factor analysis as least squares fit problems, Chemom. Intell. Lab. Syst., 18(2), 183–194, doi:10.1016/0169-7439(93)80055-M, 1993.

Ulbrich, I. M. et al.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmos. Chem. Phys., 9(9), 2891–2918, doi:10.5194/acp-9-2891-2009, 2009.

Figure S1

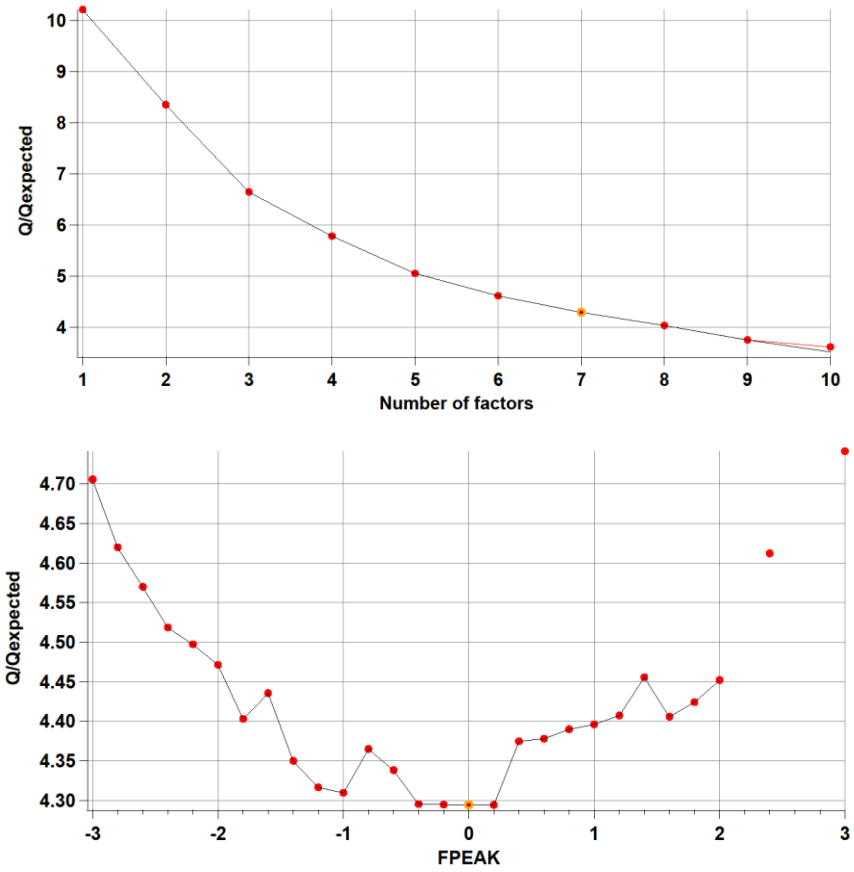


Figure S1. (a) Change in the quality of fit parameter (Q/Q_{exp}) with increasing number of factors at FPEAK = 0. The % change in the Q/Q_{exp} value is larger than 10 % at each successive step until p = 5. For p > 5, % change in Q/Q_{exp} value < 10 % for each successive step increase in p. **(b)** Change in the values of Q/Q_{exp} for the FPEAK range from -3 to +3. The Q/Q_{exp} values change by ~ 10 % from the minimum of 4.3 at FPEAK = 0 over this FPEAK range.

Figure S2

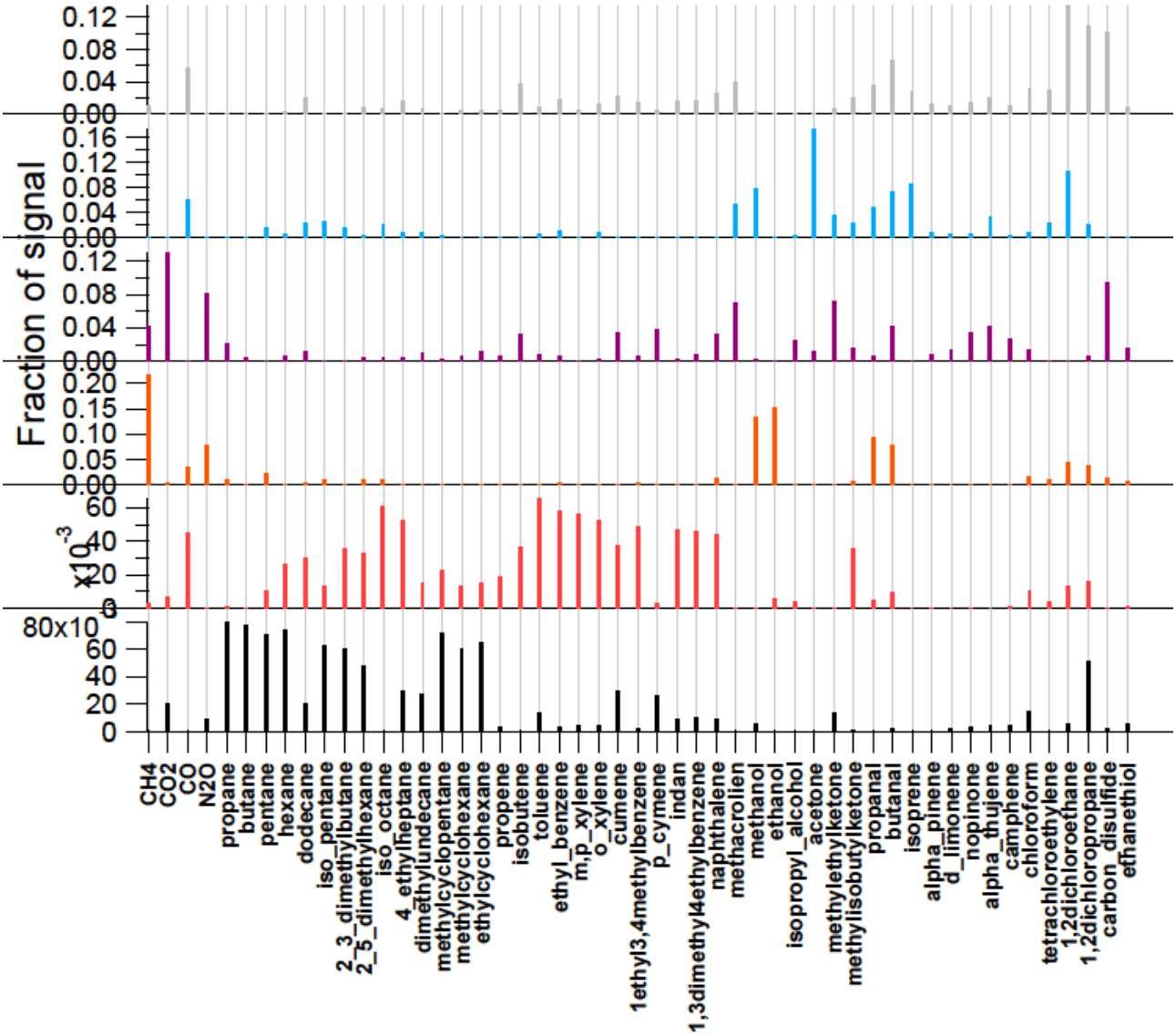


Figure S2. PMF 6-factor profile (FP). The source factors are: evaporative/fugitive (in black), vehicles (in red), dairy and livestock (in orange), agricultural + soil management (in purple), daytime biogenics + secondary organics (in light blue) and a *mixed* source factor (in grey) which is not unique and has contributions from more than one source.

Figure S3

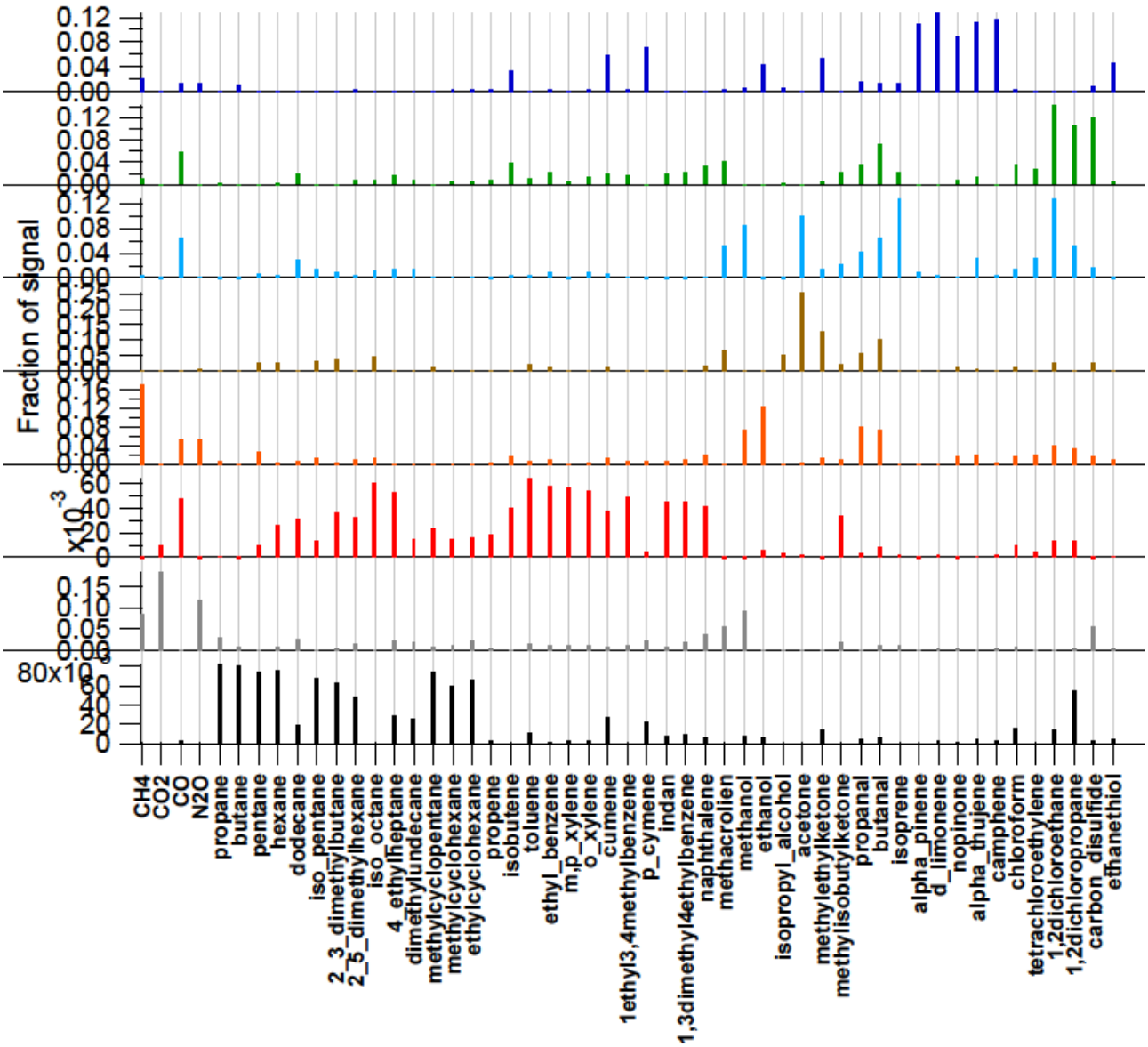
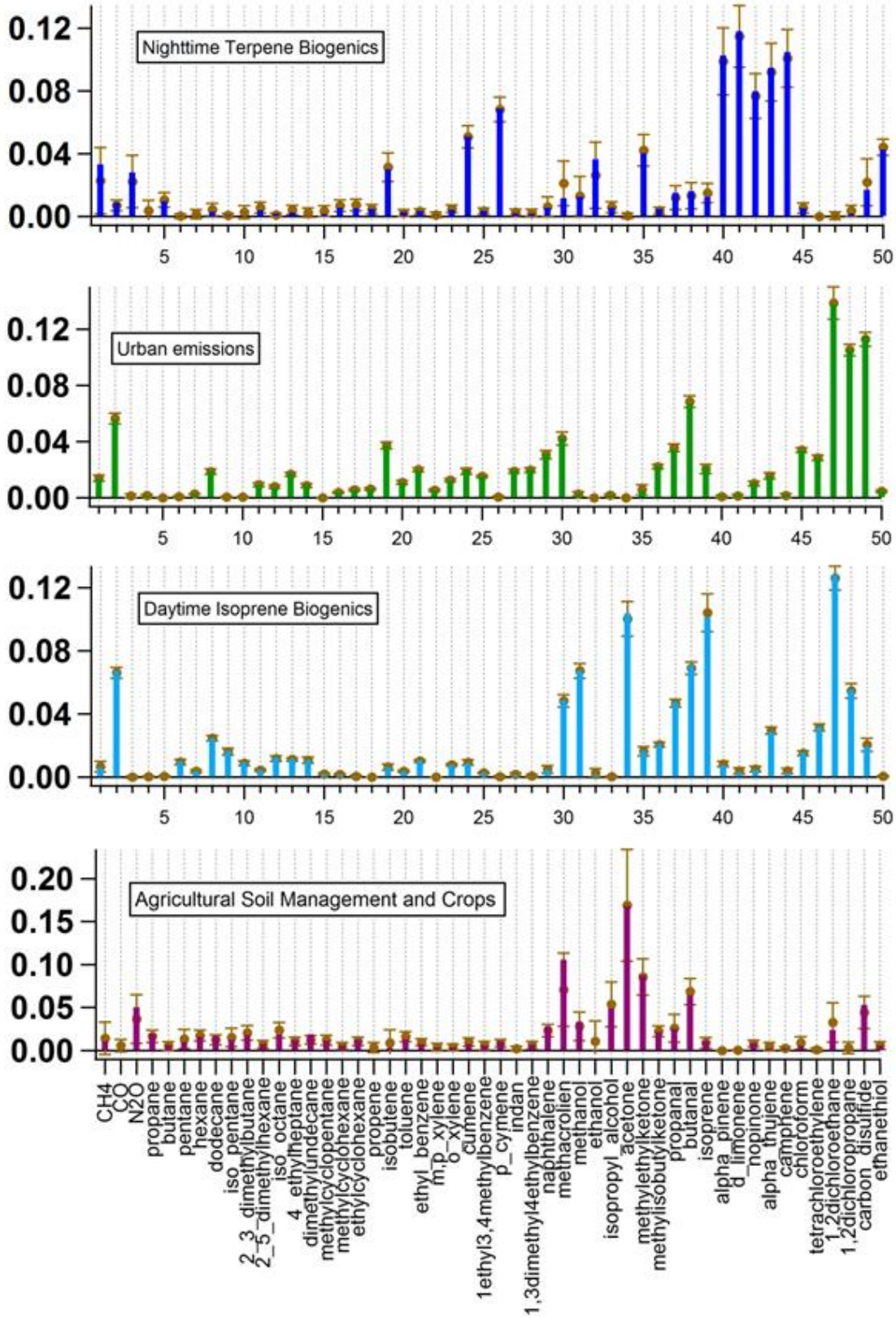
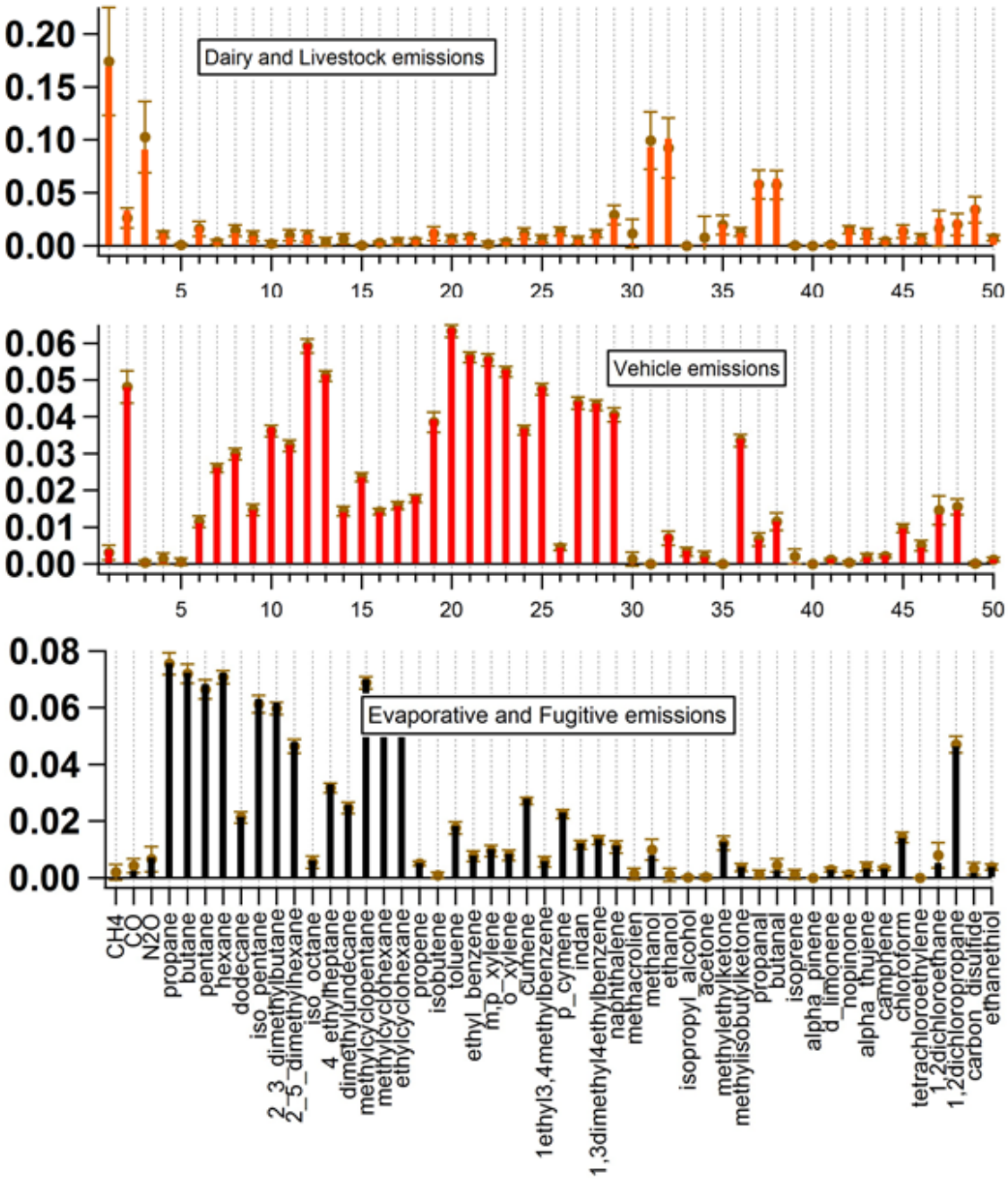


Figure S3. PMF 8-factor profile (FP). The source factors are: evaporative/fugitive (in black), vehicles (in red), dairy and livestock (in orange), daytime biogenics + secondary organics (in light blue), urban (in green), nighttime anthropogenic + terpene biogenics (in navy blue) and two *split* sources (in grey and brown, respectively) which resemble a disintegration of the agricultural + soil management source (in purple) from the 7-factor solution (*Figure S4*).

166 **Figure S4**



167



168

169 **Figure S4.** Source profile of the seven factors (at FPEAK = +0.6) with uncertainty estimates generated
170 from 100 bootstrapping runs. The source factors are **(a)** nighttime anthropogenics + terpene biogenics
171 **(b)** urban **(c)** daytime biogenics + secondary organics **(d)** agricultural + soil management **(e)** dairy and
172 livestock **(f)** vehicles and **(g)** evaporative and/or fugitive. The x-axis represents the normalized fraction of
173 mass in each source factor, while the y-axis lists all the chemical species included in the PMF analysis.
174 The numbers on the y-axis pertains to the tracer nomenclature adopted in Table 2. The solid brown

markers denote the average of the 100 bootstrapping runs and the error bars represent the standard deviation about the average.