

Supplement of Atmos. Chem. Phys., 15, 12043–12063, 2015
<http://www.atmos-chem-phys.net/15/12043/2015/>
doi:10.5194/acp-15-12043-2015-supplement
© Author(s) 2015. CC Attribution 3.0 License.



Atmospheric
Chemistry
and Physics
Open Access
EGU

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 (aguha@baaqmd.gov)

The copyright of individual parts of the supplement might differ from the CC-BY 3.0 licence.

1 **S1. Determination of number of PMF factors**

2 In PMF, the choice of modeled factors in the solution is made on the basis of a qualitative
3 judgment and remains the most critical step in the interpretation of results (Ulbrich et al., 2009). A
4 number of metrics aid in this decision making process. One of these is the Q -value which represents the
5 total sum of the squares of scaled residuals. If the assumptions of bilinear model are appropriate and
6 the errors in the input data have been properly estimated such that each reproduced data point is fit to
7 within its estimated error value, then, Q/Q_{exp} should be ~ 1 . Values of $Q/Q_{exp} \gg 1$ indicate
8 underestimation of the errors or inability of the PMF solution to explain a significant portion of the
9 variability in factor profiles as the modeled sum of contributions of the chosen number of factors p .
10 Hence, the estimated Q/Q_{exp} is explored as a function of the number of factors in order to determine the
11 best modeled representation. Addition of factors (increasing p) adds more degrees of freedom to enable
12 a better fitting of the data and decreases the value of Q/Q_{exp} and if the decrease is large enough, it
13 implies that the additional factor has explained significantly more of the variation in the data and hence
14 the added factor is real (Paatero and Tapper, 1993). The % decrease in Q/Q_{exp} values or slope of the
15 curve at each step increase in p should be used as a criterion in determining the 'best' number of factors
16 in the solution. One should be careful and wise in not choosing a PMF solution solely based on Q/Q_{exp}
17 values. Choosing too many factors in a PMF solution may make a real factor further dissociate into two
18 or more non-existing sources. This phenomenon is known as *splitting* and discussed by Ulbrich et al.
19 (2009). Hence rejecting a solution involving *splitting* behavior in factors should serve as a criterion while
20 narrowing down on a PMF solution. Additional factors may also be non-unique with contributions from
21 all major classes of compounds thus rendering the apportionment of the factor useless and should be
22 used as a criterion to reject solutions. On the other hand, choosing too few factors will combine sources
23 with different emission characteristics together to produce a single factor and hence yield a solution
24 that will be difficult to interpret (Hopke, 2000). In the end, the ability to interpret a FP and issue it a
25 name of a source category, based on *a priori* knowledge of the chemical compositional profile of the
26 source, remains a qualitative but a necessary step in identification of the final PMF solution. As per P.
27 Paatero (the creator of the PMF technique), this subjectivity is a part of the PMF process and should be
28 reported in scientific publications (Ulbrich et al., 2009).

29 Figure S1.a shows the variation of Q/Q_{exp} values with increasing p for solutions including up to 10
30 factors at FPEAK = 0 (discussed in Section S.2). The Q/Q_{exp} values show a steep decrease from $p = 1$ to 5 (
31 > 10 % drop at each step) but then gradually the decrease becomes steady and is less than 10 % at each

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

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

55 **S2. Rotation of factors**

56 The bilinear PMF analysis has rotational ambiguity and is not mathematically unique. The
57 constraint of non-negativity reduces the rotational freedom in the system but does not generally
58 produce a unique solution. There may be potentially infinite linear transformations, better known as
59 “rotations”, that can reduce the rotational freedom by introducing zero values in the factor mass profile
60 (**F**) and time series (**G**) and can force the solution to produce an identical fit to the data (Ulbrich et al.,
61 2009), such that:

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

62 In the PMF2 algorithm, the rotated factor product is allowed to differ slightly from the product of the **G**
63 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
64 “distorted” rotations which may lead to a slightly worse but acceptable fit to the data with similar but
65 higher values of *Q* and potentially yield more physically realistic solutions (Paatero et al., 2002). After
66 the case with the *best* number of factors has been established, a subset of the “distorted” linear
67 transformations of the solution can be explored using the FPEAK parameter. Positive FPEAK values force
68 the routine to add one **G** column vector to another and subtract the corresponding **F** row vectors from
69 each other while negative FPEAK values explore the reverse scenario (Paatero, 1997; Hopke, 2000). Zero
70 values in the **F** and **G** matrices (no rotations) will limit subtractions in the matrices owing to the non-
71 negativity constraint and thus limit the scope of solutions. Only “rotations” for which the *Q*-value is not
72 significantly greater than the central case (FPEAK = 0) are considered. Prior literature suggests not
73 considering rotations for a FPEAK case in which the Q/Q_{exp} value shows an increase of 10 % or more
74 above its minimum value (usually $Q_{FPEAK=0}$) (Paatero et al., 2002). The rotation procedure produces, for
75 each FPEAK, new rotated matrices **GT** and $\mathbf{T}^{-1}\mathbf{F}$ that represents time series and factors respectively, that
76 may appear to be closer to physically real source profiles than **G** and **F**.

77 A narrow FPEAK range is more appropriate in cases where Q/Q_{exp} value for (*p*-1)-factor solution
78 (FPEAK = 0) is less than 10 % higher than the Q/Q_{exp} value for the corresponding case in the *p*-factor
79 solution. This is true in the current case of 6 versus 7-factor solution (Figure S1.a). Figure S1.b plots the
80 variation in Q/Q_{exp} values with respect to the FPEAK parameter for the 7-factor solution over a range of
81 FPEAKS from -3 to +3 in increments of 0.2 units. Solutions with narrower FPEAK range that give an
82 increase of 1 % over the minimum Q/Q_{exp} value have been investigated for acceptable PMF fits (Ulbrich,
83 2009). The FPEAK range that meets the 1 % criterion is -1.6 to 0.4 (Figure S1.b). The standard deviation
84 over this FPEAK range is the estimated error in mass fraction of each tracer in each of the seven factors.
85 We follow the guidelines in (Comero et al., 2009; Paatero et al., 2002) about behavior of Q/Q_{exp} with
86 change in the FPEAK parameter and determine the physical plausibility of the all the factor profiles at
87 each FPEAK within the shortlisted range and choose the *best* fit to the data at FPEAK = +0.6.

88 **S3. Uncertainty estimates of solution**

89 Bootstrapping in PMF is a quantitative technique that addresses the difficult topic of evaluating
90 the stability and statistical uncertainty in a candidate PMF solution (Norris et al., 2008; Ulbrich et al.,

91 2009). In the bootstrapping procedure, the PET creates a new data set by randomly selecting non-
92 overlapping blocks of consecutive samples. The new data set has the same dimensions as the original
93 data set. PMF is then applied to this new data set. In every run, each bootstrap factor is assigned to a
94 base run factor by comparing the contributions of each factor and assigning it to the one with highest
95 correlation. At the end of the user-specified number of iterations, bootstrapping statistics for all the
96 runs are generated in the PET which include average and 1σ values for each fractional component and
97 sample mass in the FP and TS, respectively. The results of bootstrapping inform the analyst of the
98 robustness of the factor profiles chosen in the base run.

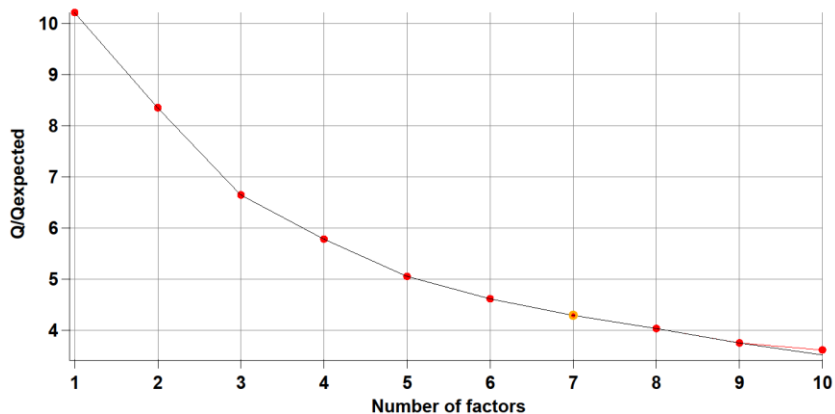
99 Bootstrapping was applied to the base run (7-factor solution, FPEAK = +0.6, SEED = 0) with 100
100 runs. The FP of the seven factor profiles with their bootstrapping averages and standard deviation range
101 is plotted in Figure S4. The fractional contributions to a source factor from tracers that occur in relatively
102 high proportions in the base run (indicated by colored bars) is quite similar to the averages over the 100
103 bootstrapping runs (dots) in all the seven factors. The plot also shows the uncertainty in each mass
104 fraction represented by the standard deviation (1σ) of these averages (indicated by whiskers about the
105 dots). For e.g. the uncertainty in the normalized fractional proportion of CH_4 in the dairy and livestock
106 source factor is 29 % (1σ confidence interval) of the mean CH_4 mass fraction while the uncertainty in
107 PMF-derived N_2O fraction of agricultural and soil management factor is 70 % of the mean fraction of N_2O
108 apportioned to that factor. The overall averaged mass fraction of compounds in all factors from the
109 bootstrapping runs is similar to the factors from the base run (Figure S4) that suggests that the chemical
110 profile of each factor is reproduced consistently in the bootstrapping runs. Within a factor, the
111 uncertainties in individual mass fractions are lower for major constituents while minor constituents have
112 larger uncertainties. The uncertainties of the tracers that occur in relatively minor proportions in each
113 source factor can be high which is a known limitation as PMF is weak in its partitioning of the mixing
114 ratio signals arising from collocated sources and artifacts arising due to meteorology (like strong daytime
115 mixing), and hence suffers from the 'mixing' and 'splitting' phenomena (discussed above). But in spite of
116 these high uncertainties, the 1σ confidence interval of the mass fraction of these minor CH_4 (or N_2O)
117 fractions is significantly smaller than the confidence intervals of the major CH_4 (or N_2O) mass fractions,
118 and as such are insignificant contributors to the overall apportionment. We conclude that the
119 bootstrapping results show a robust 7-factor PMF solution with reasonable uncertainties for tracers that
120 are major contributors to a source factor. The uncertainties also confirm that PMF analysis does not
121 yield a unique solution but rather presents a range of possible combinations of mass fractions of
122 compounds, all with low Q/Q_{exp} ratios. The uncertainties generated in the factor profile and the time

123 series from the bootstrapping runs are propagated to determine the uncertainties in the relative
124 apportionment of the trace gas distribution by source type (in Figure 7).

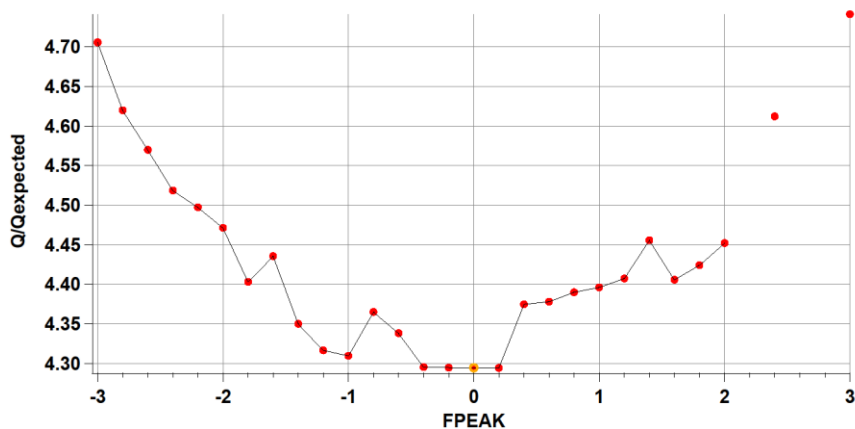
125 **References**

- 126 Bon, D. M., Ulbrich, I. M., de Gouw, J. a., Warneke, C., Kuster, W. C., Alexander, M. L., Baker, a.,
127 Beyersdorf, a. J., Blake, D., Fall, R., Jimenez, J. L., Herndon, S. C., Huey, L. G., Knighton, W. B., Ortega, J.,
128 Springston, S. and Vargas, O.: Measurements of volatile organic compounds at a suburban ground site
129 (T1) in Mexico City during the MILAGRO 2006 campaign: measurement comparison, emission ratios, and
130 source attribution, *Atmos. Chem. Phys.*, 11(6), 2399–2421, doi:10.5194/acp-11-2399-2011, 2011.
- 131 Comero, S., Capitani, L., and Gawlik, B. M.: Positive Matrix Factorization - An introduction to the
132 chemometric evaluation of environmental monitoring data using PMF, JRC Scientific and Technical
133 Reports, EUR 23946 EN-2009.
- 134 Hopke, P.: A guide to positive matrix factorization, Work. UNMIX PMF as Appl. to PM2, 1–16 [online]
135 Available from: <ftp://128.153.5.141/users/h/o/hopkepk/IAEA/PMF-Guidance.pdf> (Accessed 13 March
136 2013), 2000.
- 137 Norris, G., Vedantham, R., Wade, K. and Brown, S.: EPA positive matrix factorization (PMF) 3.0
138 fundamentals & user guide, Prep. US ... [online] Available from:
139 [http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:EPA+Positive+Matrix+Factorization+\(P](http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:EPA+Positive+Matrix+Factorization+(PMF)+3.0+Fundamentals+&+User+Guide#0)
140 [MF\)+3.0+Fundamentals+&+User+Guide#0](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.
- 141 Paatero, P.: Least squares formulation of robust non-negative factor analysis, *Chemom. Intell. Lab. Syst.*,
142 37(1), 23–35, doi:10.1016/S0169-7439(96)00044-5, 1997.
- 143 Paatero, P., Hopke, P. K., Song, X.-H. and Ramadan, Z.: Understanding and controlling rotations in factor
144 analytic models, *Chemom. Intell. Lab. Syst.*, 60(1-2), 253–264, doi:10.1016/S0169-7439(01)00200-3,
145 2002.
- 146 Paatero, P. and Tapper, U.: Analysis of different modes of factor analysis as least squares fit problems,
147 *Chemom. Intell. Lab. Syst.*, 18(2), 183–194, doi:10.1016/0169-7439(93)80055-M, 1993.
- 148 Ulbrich, I. M. et al.: Interpretation of organic components from Positive Matrix Factorization of aerosol
149 mass spectrometric data, *Atmos. Chem. Phys.*, 9(9), 2891–2918, doi:10.5194/acp-9-2891-2009, 2009.

150 **Figure S1**

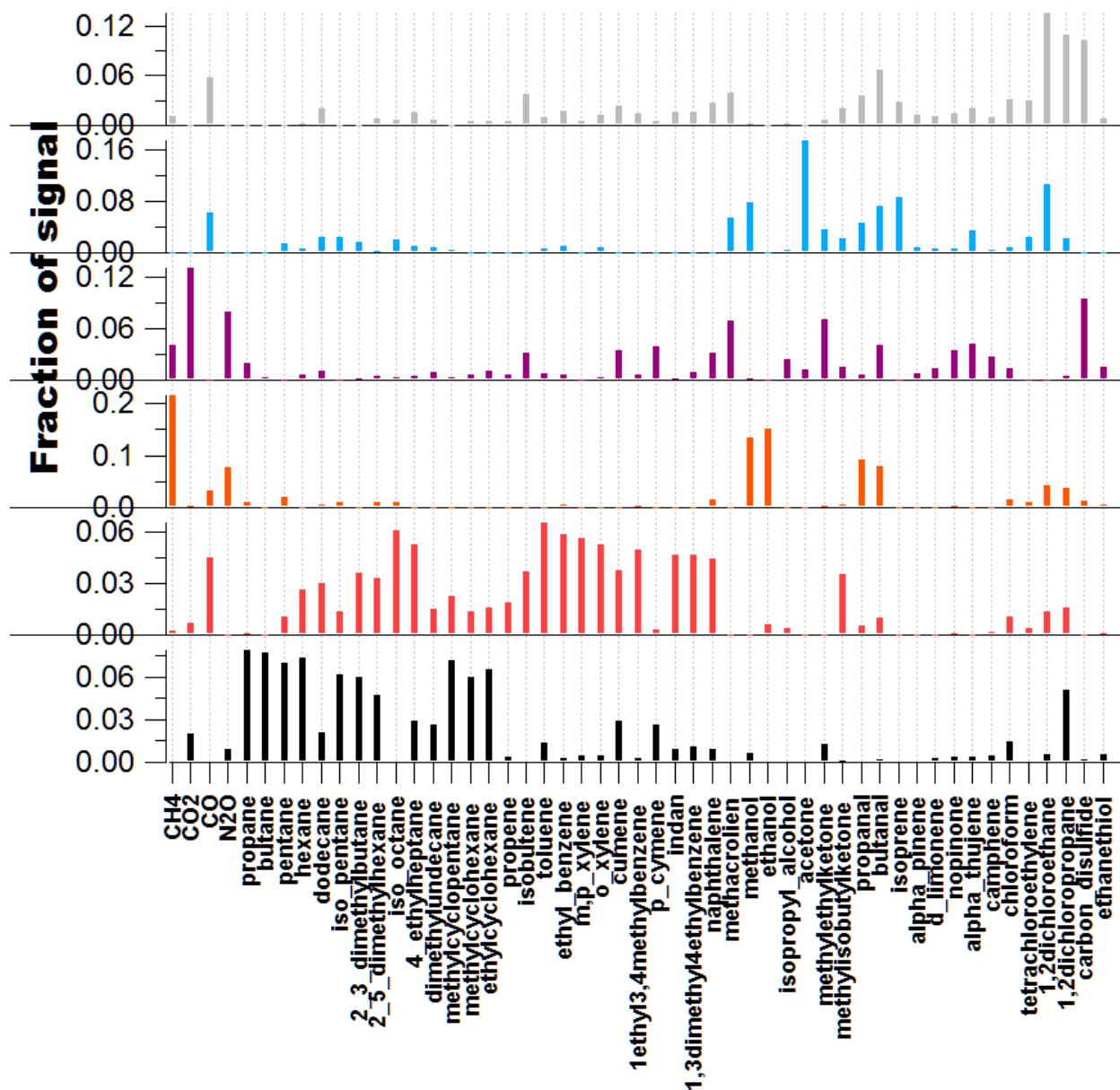


151

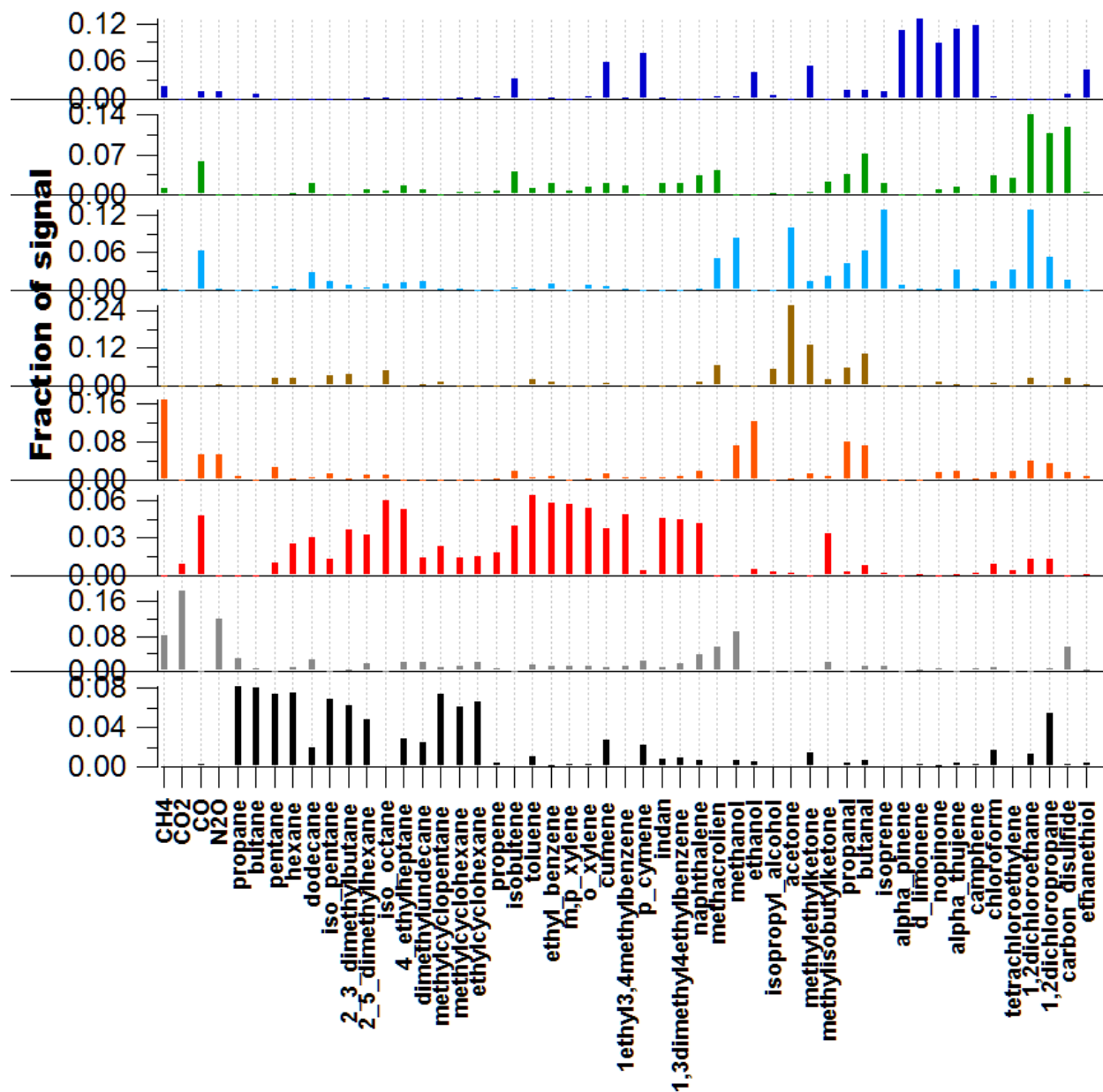


152

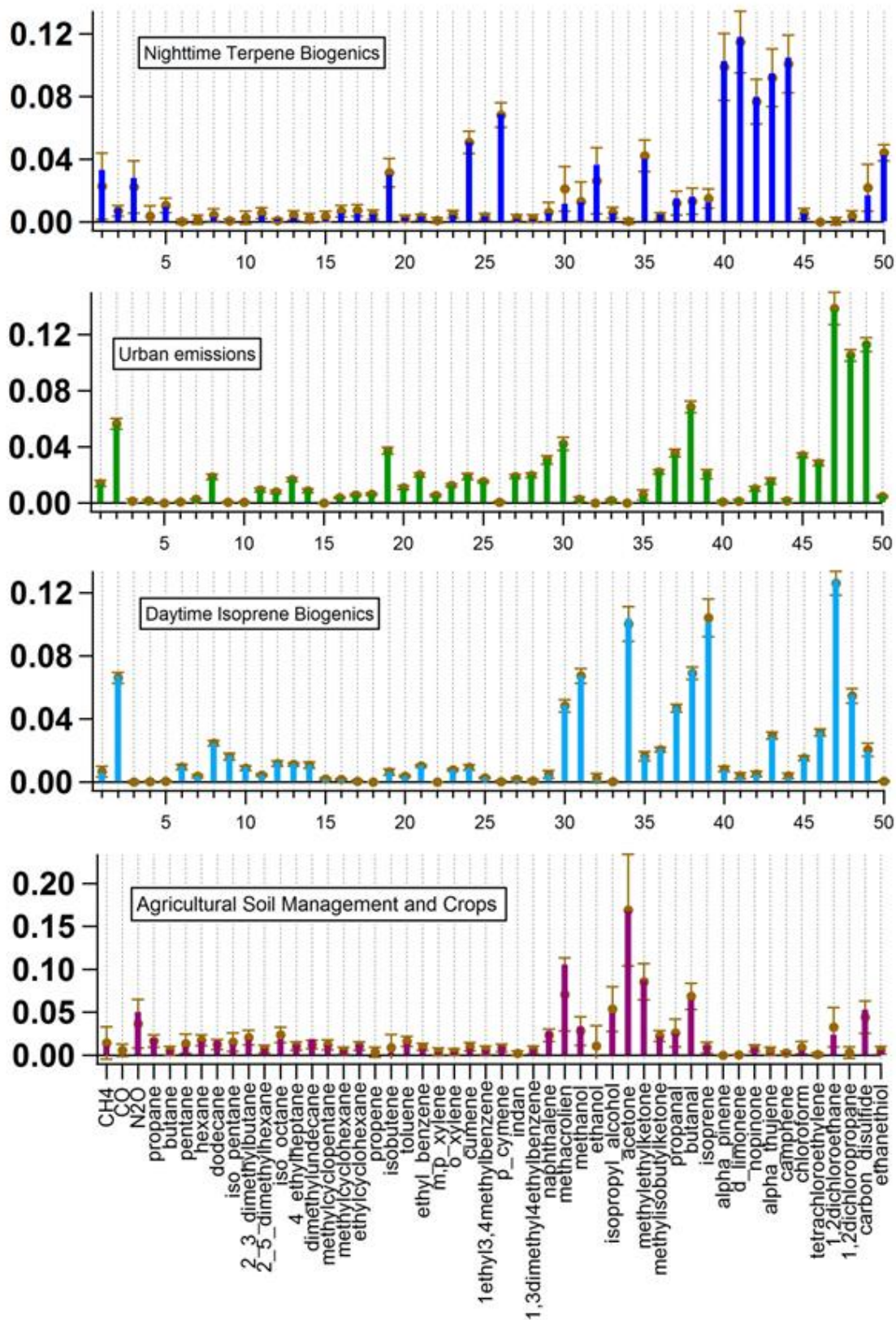
153 **Figure S1.** (top) Change in the quality of fit parameter (Q/Q_{exp}) with increasing number of factors at
154 FPEAK = 0. The % change in the Q/Q_{exp} value is larger than 10 % at each successive step until $p = 5$. For p
155 > 5 , % change in Q/Q_{exp} value < 10 % for each successive step increase in p . (bottom) Change in the
156 values of Q/Q_{exp} for the FPEAK range from -3 to +3. The Q/Q_{exp} values change by ~ 10 % from the
157 minimum of 4.3 at FPEAK = 0 over this FPEAK range.

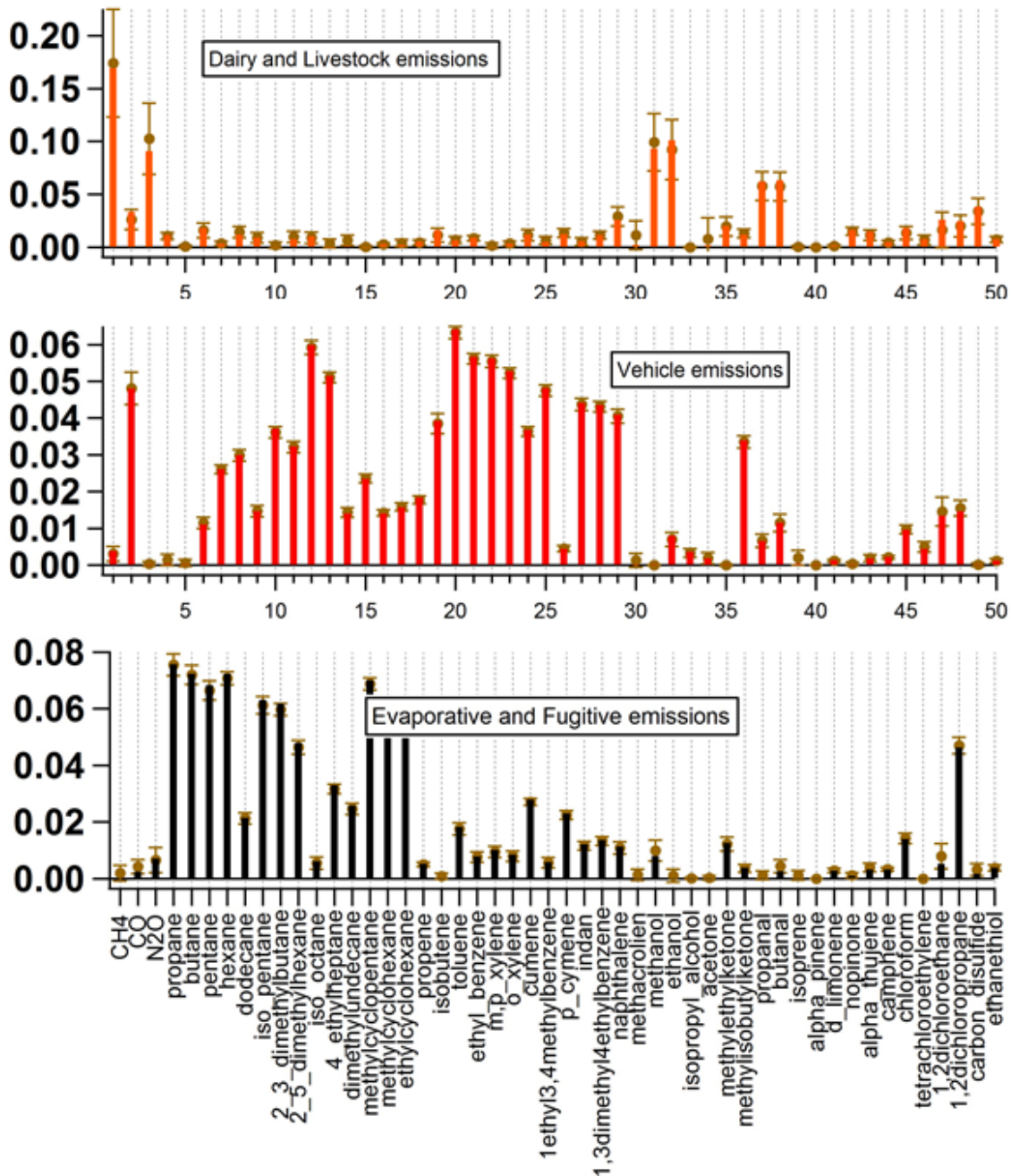


159
 160 **Figure S2.** PMF 6-factor profile (FP). The source factors are: evaporative/fugitive (in black), vehicles (in
 161 dairy and livestock (in orange), agricultural + soil management (in purple), daytime biogenics +
 162 secondary organics (in light blue) and a *mixed* source factor (in grey) which is not unique and has
 163 contributions from more than one source.



165
 166 **Figure S3.** PMF 8-factor profile (FP). The source factors are: evaporative/fugitive (in black), vehicles (in
 167 red), dairy and livestock (in orange), daytime biogenics + secondary organics (in light blue), urban (in
 168 green), nighttime anthropogenic + terpene biogenics (in navy blue) and two *split* sources (in grey and
 169 brown, respectively) which resemble a disintegration of the agricultural + soil management source (in
 170 purple) from the 7-factor solution (*Figure S4*).





173
 174 **Figure S4.** Source profile of the seven factors (at FPEAK = +0.6) with uncertainty estimates generated
 175 from 100 bootstrapping runs. The source factors are (a) nighttime anthropogenics + terpene biogenics
 176 (b) urban (c) daytime biogenics + secondary organics (d) agricultural + soil management (e) dairy and
 177 livestock (f) vehicles and (g) evaporative and/or fugitive. The x-axis represents the normalized fraction of
 178 mass in each source factor, while the y-axis lists all the chemical species included in the PMF analysis.
 179 The numbers on the y-axis pertains to the tracer nomenclature adopted in Table 1. The solid brown

180 markers denote the average of the 100 bootstrapping runs and the error bars represent the 1σ standard
181 deviation about the average.