



Report to
Sonoma Technology, Inc.

**INITIAL INVESTIGATION OF PAMS DATA USING
POSITIVE MATRIX FACTORIZATION (PMF)**

Prepared by

Philip K. Hopke
Department of Chemical Engineering
Clarkson University
Box 5705
Potsdam, NY 13699-5705

Introduction

Receptor modeling methods have been primarily applied to particulate species. There have been several studies that analyzed VOC data using factor analysis methods (Henry *et al.*, 1994; 1997). However, there has not been extensive application of current state-of-the-art modeling methods to the extensive VOC data obtained through the Photochemical Assessment Monitoring Station (PAMS) network sites. This monitoring effort was mandated by the Clean Air Act Amendments of 1990 for areas of moderate to severe non-attainment of the National Ambient Air Quality Standard for Ozone. In this report, positive matrix factorization will be applied to PAMS data from three sites in the northeastern US (Rider College and New Brunswick, NJ and Sherwood Island, Connecticut).

Positive Matrix Factorization

An important new approach to the factor analysis has been developed in which factor analysis has been recognized as an explicit least square problem (Paatero, 1997). This approach is called Positive Matrix Factorization (PMF). There are important advantages to this method in that it permits individual data point weighting as an optimum scaling method. The analysis gives the highest weights to the data in which there is the most confidence. Below detection limit and missing values can easily be included by giving these values low weights (high uncertainties). In environmental data where there can be extreme values in high concentration tail of the distribution, these points can be downweighted to insure that they do not have undue influence on the final results. It is also easy to impose non-negativity and other constraints into the problem as well as develop more complex models when the problem warrants it.

The data for each site will be analyzed using Positive Matrix Factorization. In this method, the factor analysis problem has been explicitly addressed as a least-squares problem. This approach allows the data to be properly scaled. The elements of the "Residual Matrix," E are defined as

$$e_{ij} = x_{ij} - \hat{x} = x_{ij} - \sum_{k=1}^p f_{ik} \cdot g_{kj} \quad (1)$$

where x_{ij} is the concentration of chemical species i in sample j , f_{ik} is the concentration of that species in particles from source k and g_{kj} is the mass concentration of particles contributed by source k to sample j with $i=1, \dots, m$ chemical species, $j=1, \dots, n$ samples and $k=1, \dots, p$ sources. The "object function," Q, that is to be minimized as a function of G and F is given by

$$Q(E) = \sum_{i=1}^m \sum_{j=1}^n \left[\frac{e_{ij}}{s_{ij}} \right]^2 \quad (2)$$

where s_{ij} is an estimate of the "uncertainty" in the i th variable measured in the j th sample. The factor analysis problem is then to minimize $Q(E)$ with respect to G and F with the constraint that each of the elements of G and F is to be non-negative. This approach permits each value to have its own weight. Thus, it can handle below detection limit and missing values. The details of the algorithm are given by Paatero (1997). The method has been applied to the apportionment of airborne particles in a number of locations (Polissar *et al.*, 1998; Patterson *et al.*, 1999; Lee *et al.*, 1999; Xie *et al.*, 1999; Ramadan *et al.*, 2000). In general, more sources and a better apportionment of the aerosol mass can be obtained using PMF than through alternative eigenvalue-based methods (Huang *et al.*, 1999). Thus, we have a flexible modeling tool that can be applied to develop source profiles and estimate source contributions. The application of this

factor analysis model has been used in at least 35 journal publications over the past 7 years and has been growing rapidly in its application to the source resolution and related problems.

Data Sets

The data sets were provided by STI following their validation of the data. Data from Rider College and New Brunswick, New Jersey and Sherwood Island, Connecticut were chosen for data completeness, quality, and representing both source and downwind areas. The basic assumption of receptor models is that the composition profiles of the various sources are preserved as the material is transported to the sampling site. There is then concern with respect to the reactivity of the olefins being measured as they are the most reactive compounds in the mixture. Thus, analyses were made with and without the olefins present.

The data were examined for below detection limit and missing values. For those samples where a majority of values were missing, the sample was deleted from the data set. In those cases where there were a limited number of missing values, zeros were substituted.

In this initial investigation, a very simple error model was used to estimate the uncertainties in the values. The uncertainties were taken to be 20% of the measured values. If we had estimated measurement uncertainties or method detection limits, we could establish a better error model in which we would have an error model of the form:

$$S_{ij} = c_1 + c_3 * x_{ij} \quad (3)$$

where c_1 is the estimated uncertainty or method detection limit value and c_3 . In this case, c_1 is 0 and c_3 is 0.20. Thus, these results should be considered preliminary in nature because we have not been able to develop as complete an error model for the data as we would typically have with particulate matter data. These sets of data are our first experience with PAMS data and we have not yet developed a good understanding of the nature of these results and their typical error structure.

The number of factors in each case was determined using multiple indicators. Solutions were obtained for various numbers of factors. In each case, the distributions of scaled residuals were examined. These distributions should be symmetric with a standard deviation of 1 if the errors are adequately estimated and the data have been fitted well. We also examine the variation of Q with the factor number looking for the value at which there is no further significant decreases in Q with an increasing numbers of factors. We also examine the profiles to ascertain that they are physically reasonable.

Results

In each case we have the contributions for each source for each site. However, we have not been able to ascertain any assistance from these contribution plots in assigning the profiles to specific source types. We have thus not included the contribution plots in this report. We will perform some time series analyses (Fourier analysis) to determine if there are well defined frequencies in the contributions that might help with the source attributions.

Rider College

The profiles obtained from the PMF analysis of the data collected at Rider College, NJ with the olefins included in the analysis are provided in Figure 1 while the results without the olefins are shown in Figure 2. A list of the abbreviations and the full names of the species used in

the analyses are provided in Table 1. With olefins, the best solution had 6 factors while only 4 factors were required for the data set without the olefins. For the profiles with olefins (Figure 1), we can make the following tentative assignments: Factor 1 represents accumulation/aged air mass with an abundance of ethane and propane; Factor 2 is clearly biogenic with a high concentration of isoprene; Factor 3 is assigned as mostly evaporative materials from motor vehicles while Factor 4 is mostly motor vehicle combustion emissions; Factor 5 may be the accumulation/aged air mass along with a small amount of combustion products; and finally, Factor 6 appears to be diesel (C9-C11), combustion (C2-C3), and solvents (C6-C8 compounds). If we look at the combination of Factors 3 and 4, they seem to provide a good motor vehicle signature. There is rotational ambiguity in these results (Paatero *et al.*, 2002) and with limited experience in VOC profile recognition, it is difficult to be sure that these are the best possible profiles that can be extracted from the data.

For the profiles without olefins, we cannot identify the biogenic source since the most distinctive component is isoprene. The factor profiles given in Figure 2 as assigned as Factor 1 is diesel, but intermixed with lower molecular weight materials (solvents?). Factor 2 is a very typical profile for what we would expect from motor vehicles. Factors 3 and 4 appear to be mixed factors with Factor 3 including evaporative materials from gasoline with accumulation/aged air mass components. Factor 4 includes evaporative (C4-C7), accumulation/aged air mass (C2-C3) with some motor vehicle (toluene and isopentane).

New Brunswick

The profiles obtained from the PMF analysis of the data collected at New Brunswick, NJ with the olefins included in the analysis are provided in Figures 3 and 4. In this case, ten factors were found for the data set with olefins. The tentative assignments are as follows:

- Factor 1: motor vehicle (toluene and isopentane)
- Factor 2: industry + combustion
- Factor 3: aged/accumulation + diesel
- Factor 4: evaporative emissions of C3-C5
- Factor 5: C4 evaporative emissions
- Factor 6: unknown
- Factor 7: biogenic (isoprene)
- Factor 8: pentane/gasoline evaporative emissions
- Factor 9: natural gas + accumulation/aged air mass
- Factor 10: motor vehicle (diesel and gasoline)

The results without the olefins are shown in Figure 5. In this case, we have made the following factor assignments:

- Factor 1: unknown
- Factor 2: toluene/solvent emissions
- Factor 3: accumulation/aged air mass
- Factor 4: natural gas + accumulation/aged air mass
- Factor 5: solvent + diesel
- Factor 6: evaporative

Factor 7: motor vehicle

Sherwood Island

The profiles obtained from the PMF analysis of the data collected at Sherwood Island, CT with the olefins included in the analysis are provided in Figure 6 while the results without the olefins are shown in Figure 7. For the results of the analysis of the data with olefins (Figure 6), Factor 1 is similar to the other accumulation/aged air mass with high concentrations of ethane and propane. Factor 2 has high concentrations of cis-2-Butene, n-butane, n-pentane and 2,3,4-trimethyl pentane and is not immediately identifiable. Factor 3 is biogenic with high isoprene. Factor 4 has high propane and 2,3,4-trimethyl pentane. It also includes the largest values of ethylbenzene, benzene and toluene and thus, is likely to involve gasoline vehicle emissions.

For the data without olefins (Figure 7), Factor 1 is dominated by ethane and propane and represents the accumulation of material in aged air masses. Factors 2 and 3 are similar to the other motor vehicle profiles at the other sites. Factor 2 may represent more gasoline emissions relative to Factor 3 being more related to diesel, but clearly not fully separating the contributions of the two major types of motor vehicles.

Summary

These initial analyses of PAMS data are encouraging. We have been able to separate the data into a number of components some of which can be reasonably identified. However, at this point, we cannot be certain we have the best solutions since we are uncertain of the simplistic modeling of the errors used to weight the data in the analysis and the potential for rotations that might provide somewhat more realistic profiles. We need to accumulate more experience with VOC profiles and these types of data before we can have the same level of confidence in the solutions that we have with particulate compositional data sets. We also need to do some additional work with the source contribution estimates to see if their time variation provides additional insights into the likely source types contributing to the observed VOC concentrations. We would suggest that this approach shows promise of identifying sources and will be worth further exploration to fully determine the feasibility of PMF analysis to provide useful information for making air quality management decisions.

Table 1. Abbreviations used to designate the species included in the PMF analyses.

AIRS-code	Abbreviation	Hydrocarbon-Species	Group
43206	acety	Acetylene	O
43203	ethyl	Ethylene	O
43202	ethan	Ethane	P
43205	prpyl	Propylene	O
43204	propa	Propane	P
43214	isbta	Isobutane	P
43280	v1bute	1-Butene	O
43212	nbuta	n-Butane	P
43216	t2bte	trans-2-Butene	O
43217	c2bte	cis-2-Butene	O
43221	ispna	Isopentane	P
43220	npnta	n-Pentane	P
43243	ispre	Isoprene	O
43226	t2pne	trans-2-Pentene	O
43227	c2pne	cis-2-Pentene	O
43244	v22dmb	2,2-Dimethylbutane	P
43242	cypna	Cyclopentane	P
43284	v23dmb	2,3-Dimethylbutane	P
43285	v2mpna	2-Methylpentane	P
43230	v3mpna	3-Methylpentane	P
43246	v2m1pe	2-Methyl-1-Pentene	O
43231	nhexa	n-Hexane	P
43262	mcpna	Methylcyclopentane	P
43247	v24dmp	2,4-Dimethylpentane	P
45201	benz	Benzene	A
43248	cyhxa	Cyclohexane	P
43263	v2mhxa	2-Methylhexane	P
43291	v23dmp	2,3-Dimethylpentane	P
43249	v3mhxa	3-Methylhexane	P
43250	v224tmp	2,2,4-Trimethylpentane	P
43232	nhept	n-Heptane	P
43261	mcyhx	Methylcyclohexane	P
43252	v234tmp	2,3,4-Trimethylpentane	P
45202	tolu	Toluene	A
43960	v2mhhep	2-Methylheptane	P
43253	v3mhhep	3-Methylheptane	P
43233	noct	n-Octane	P
45203	ebenz	Ethylbenzene	A
45109	m_pty	m/p-Xylene	A
45204	oxyl	o-Xylene	A
43235	nnon	n-Nonane	P
45207	v135tmb	1,3,5-Trimethylbenzene	A
45208	v124tmb	1,2,4-Trimethylbenzene	A
45212	metol	m-Ethyltoluene	A
45213	petol	p-Ethyltoluene	A
45225	v123tmb	1,2,3-trimethylbenzene	A
43238	ndec	n-Decane	P
43954	nundc	n-Undecane	P

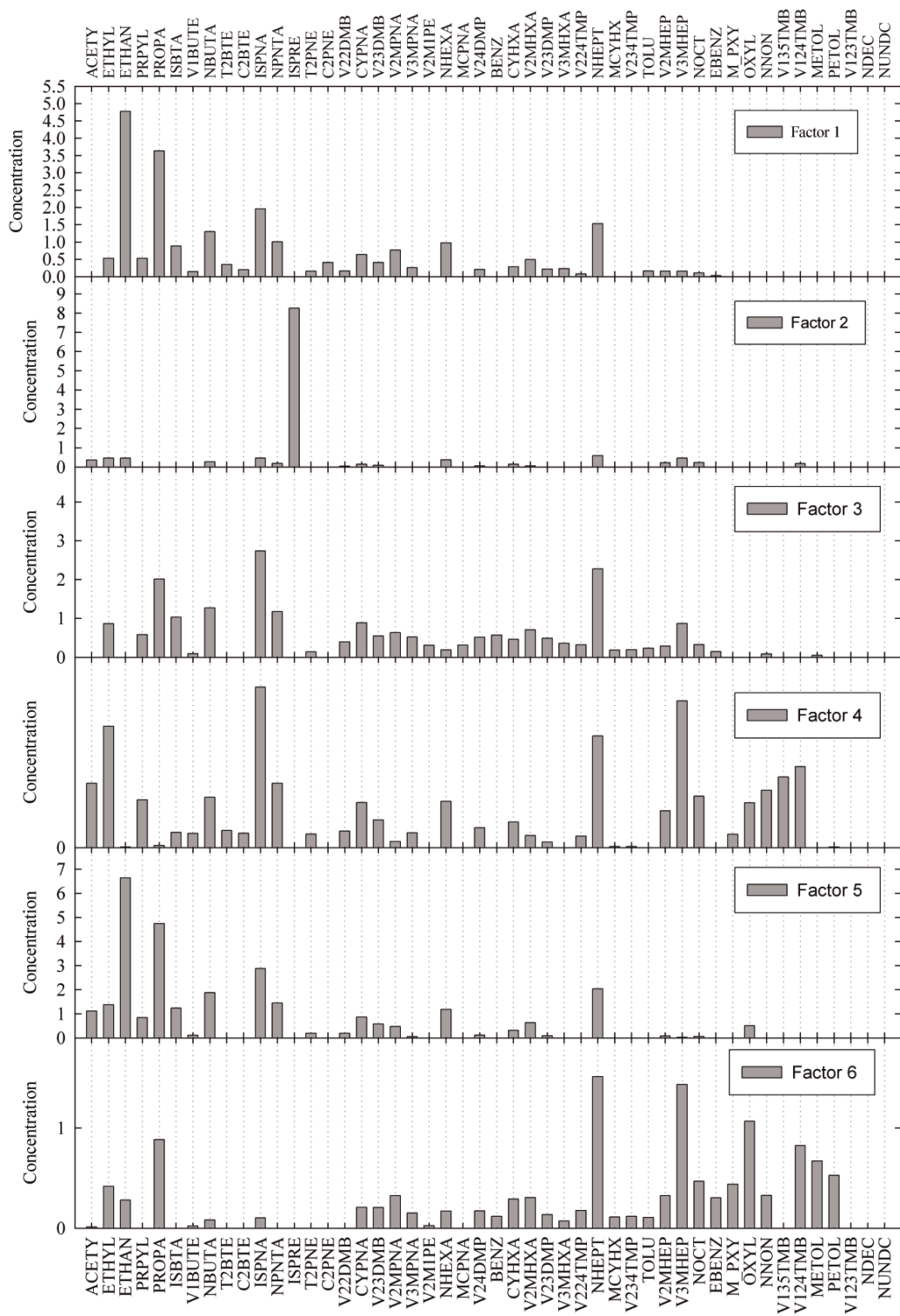


Figure 1. VOC profiles derived from the PAMS data from the Rider College, NJ site including the olefins.

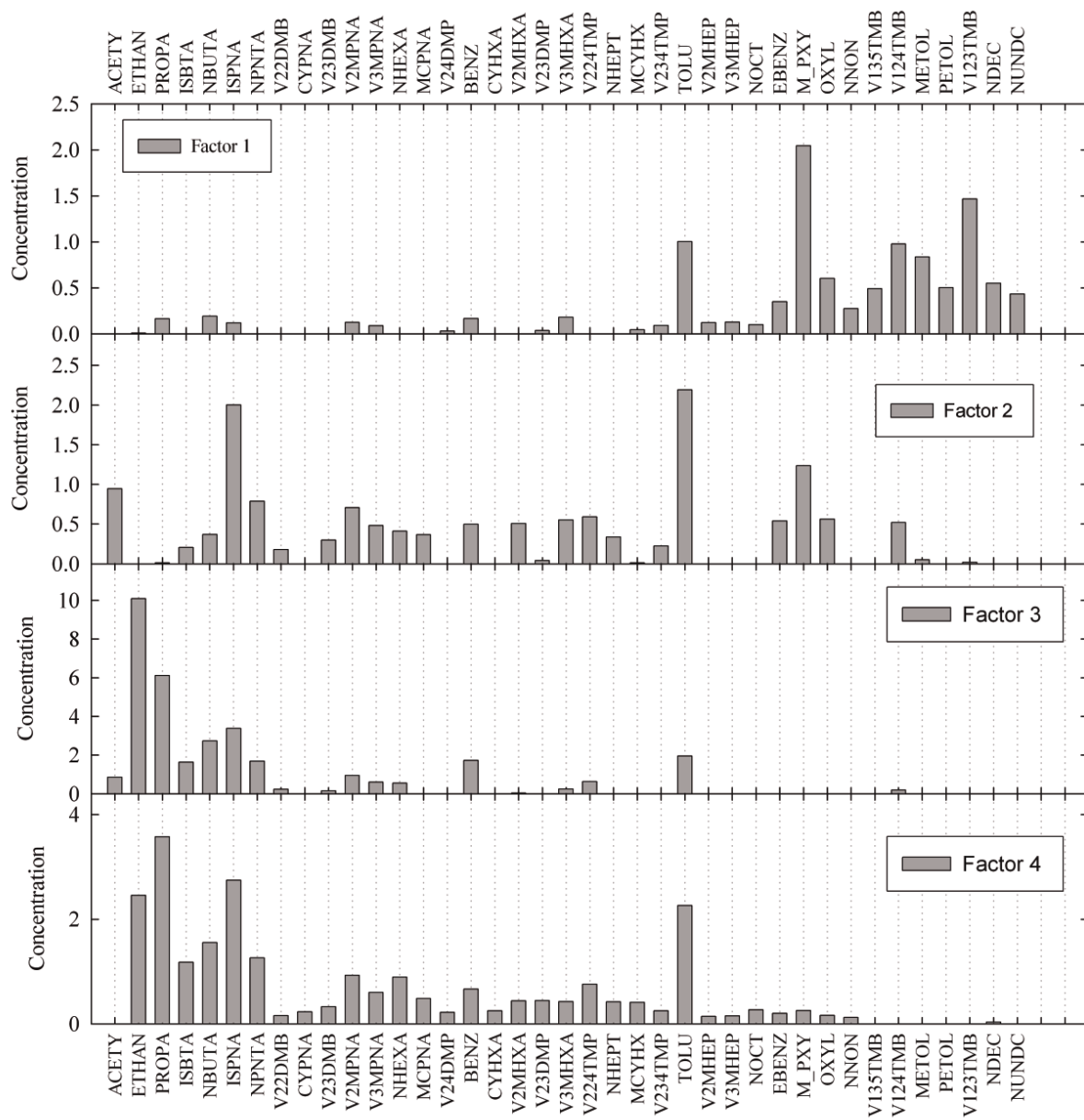


Figure 2. VOC profiles derived from the PAMS data from the Rider College, NJ site without the olefins.

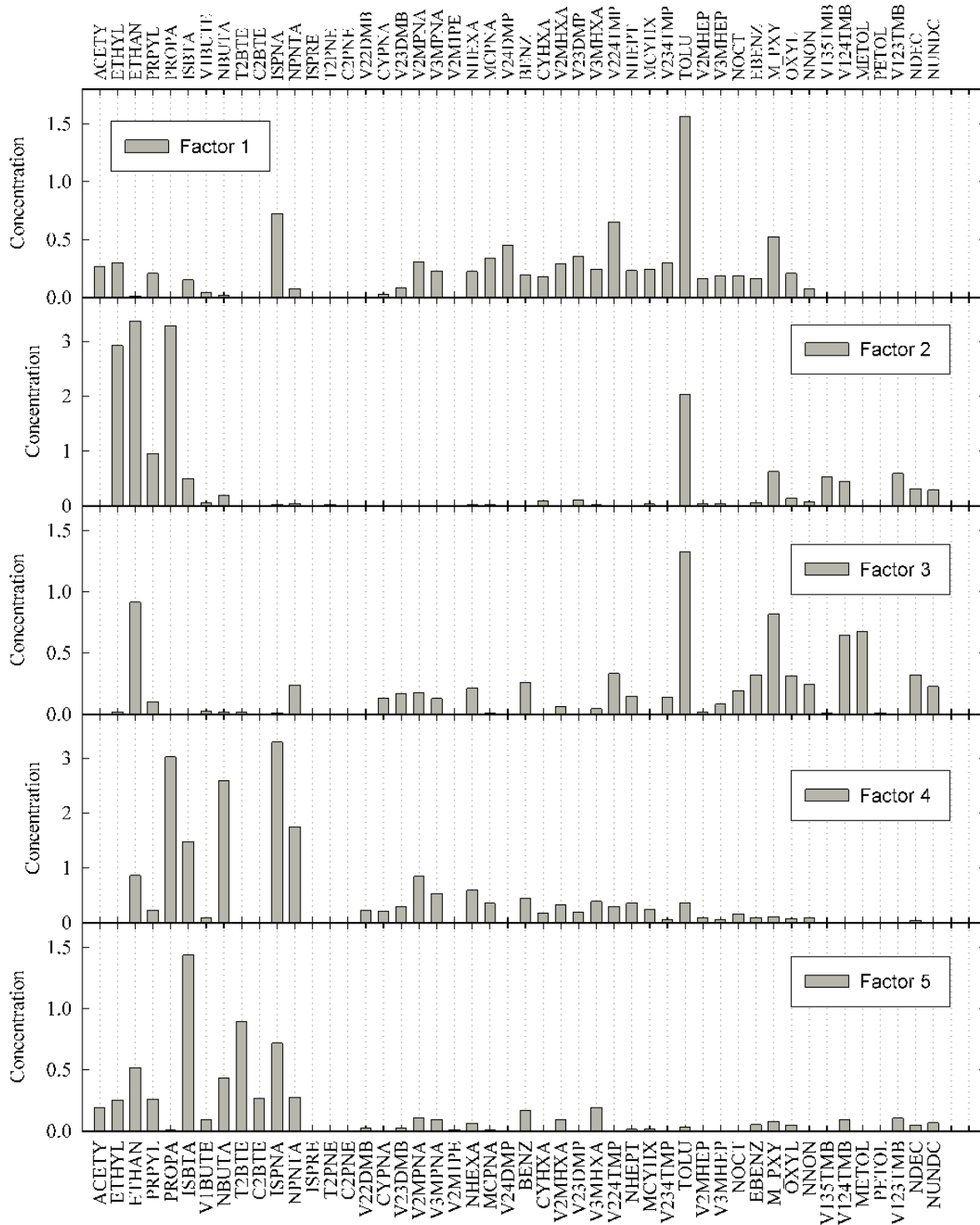


Figure 3. VOC profiles 1 to 5 derived from the PAMS data from the New Brunswick, NJ site including the olefins.

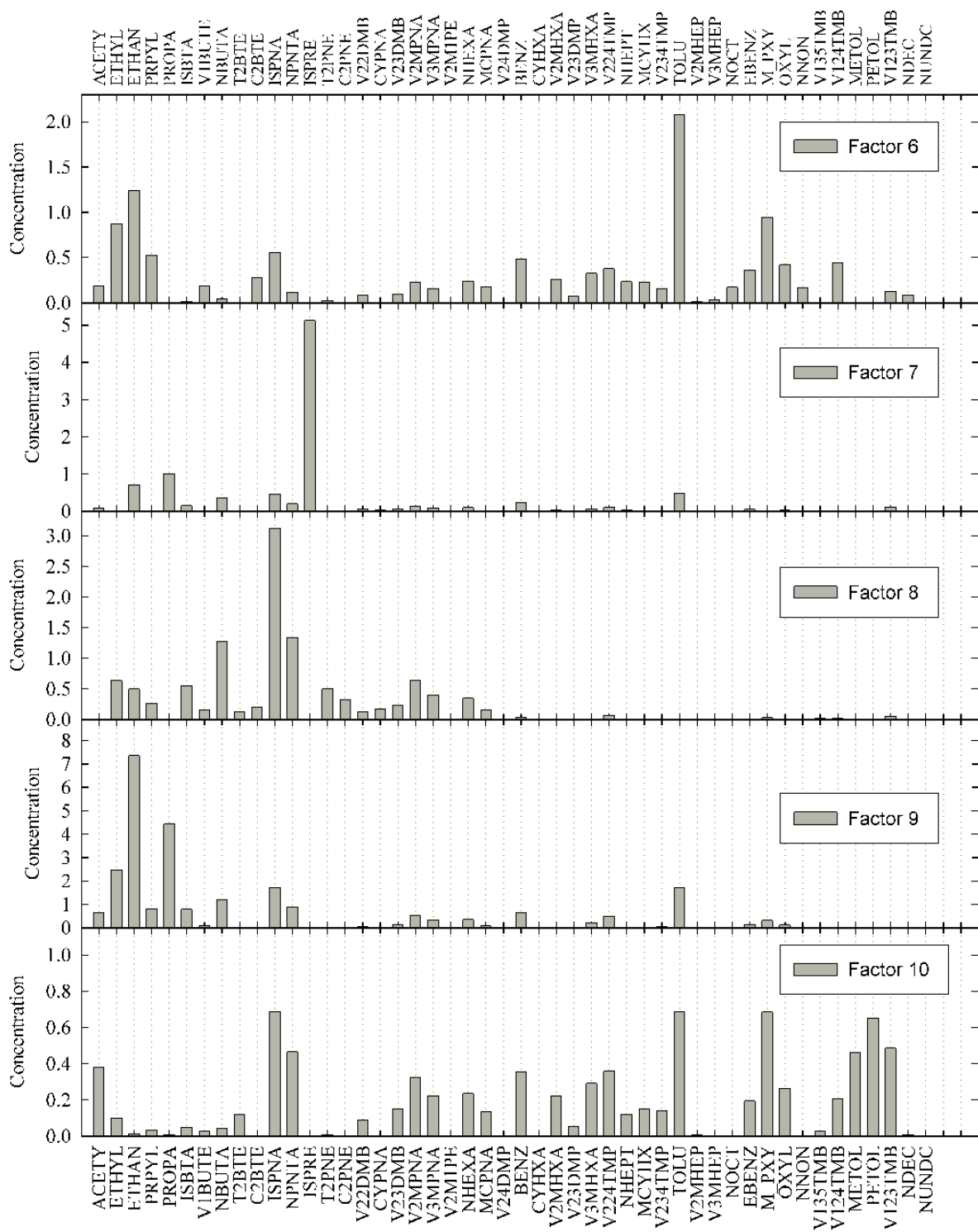


Figure 4. VOC profiles 6 to 10 derived from the PAMS data from the New Brunswick, NJ site including the olefins.

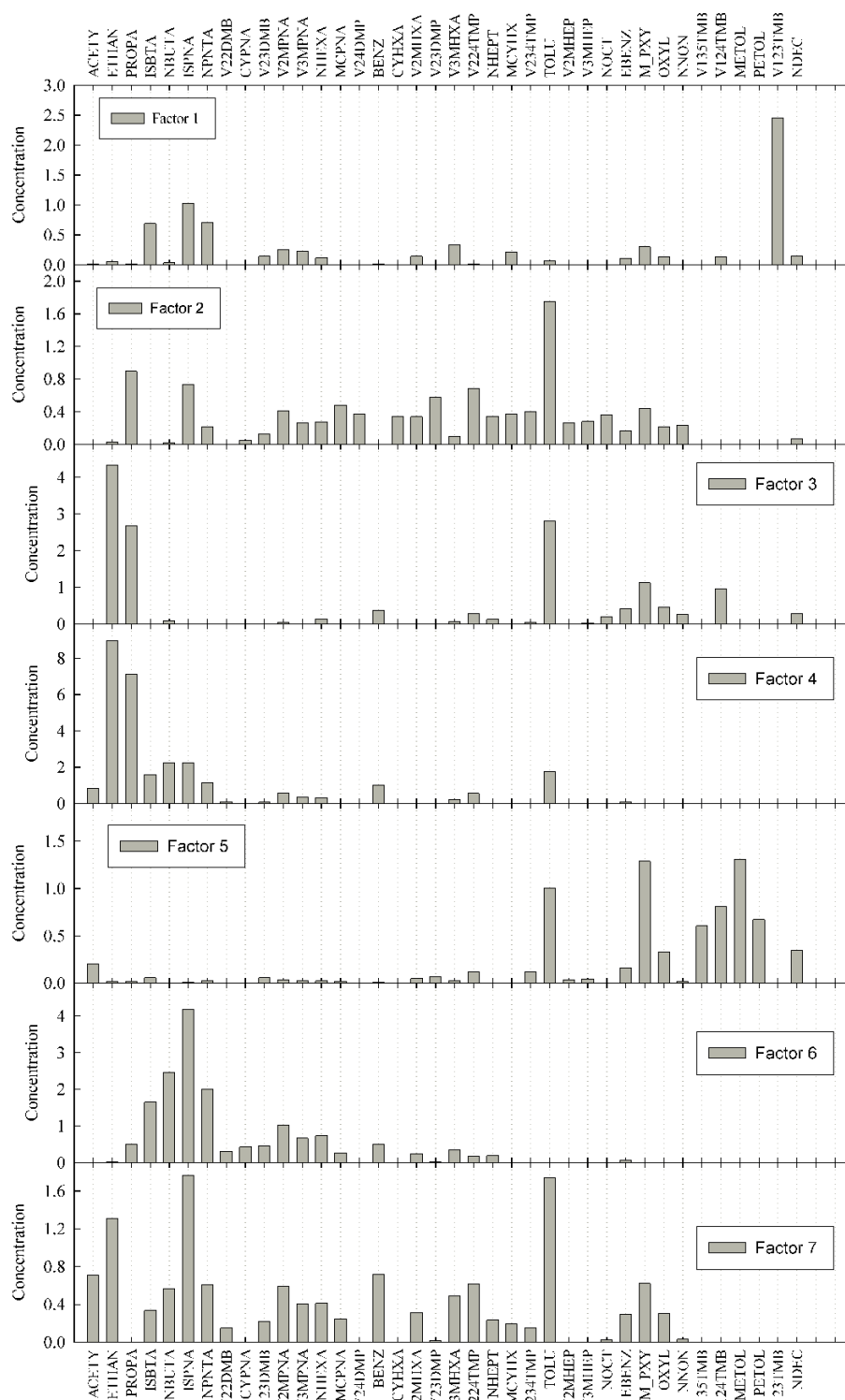


Figure 5. VOC profiles derived from the PAMS data from the New Brunswick, NJ site without the olefins.

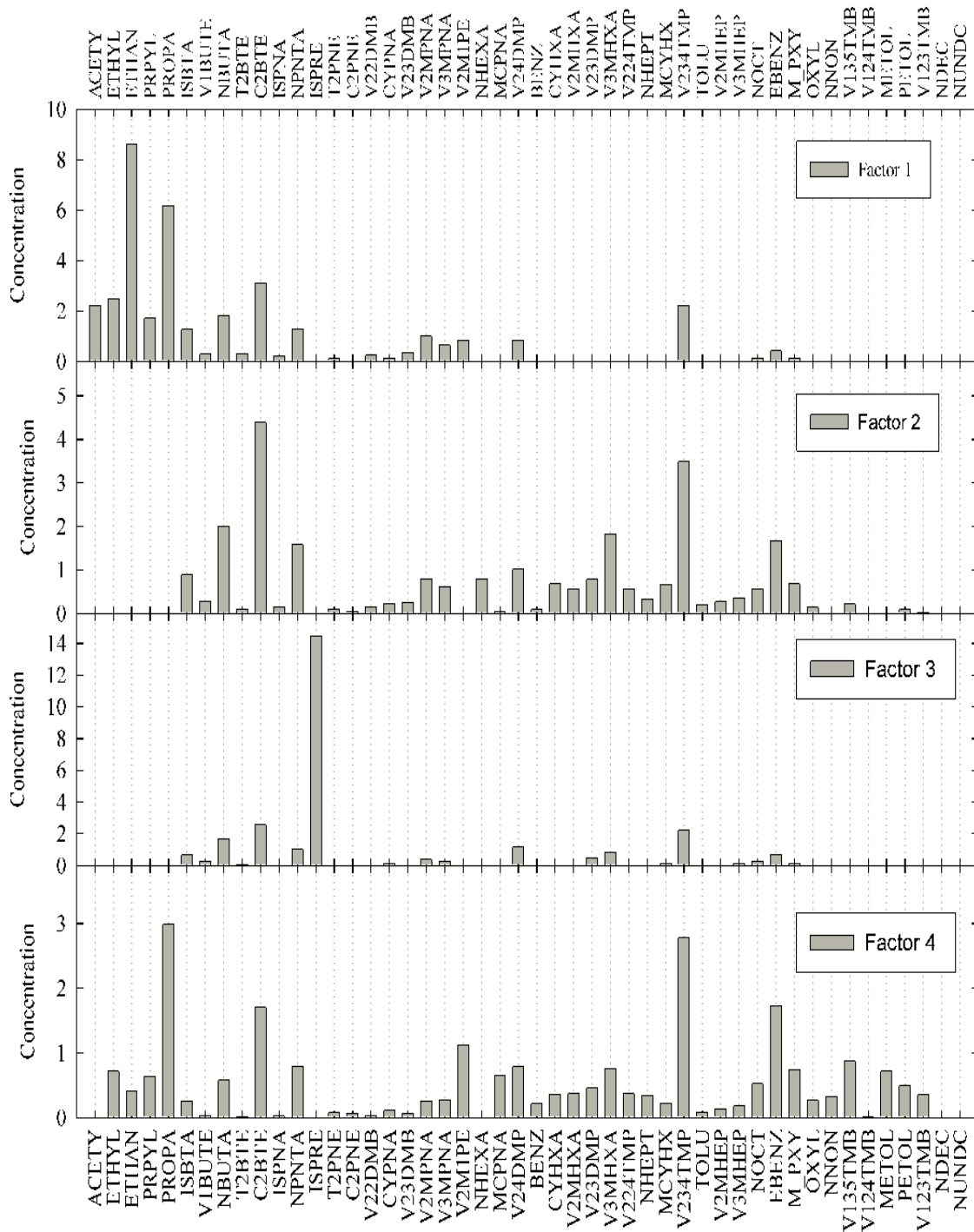


Figure 6. VOC profiles derived from the PAMS data from the Sherwood Island, CT site including the olefins

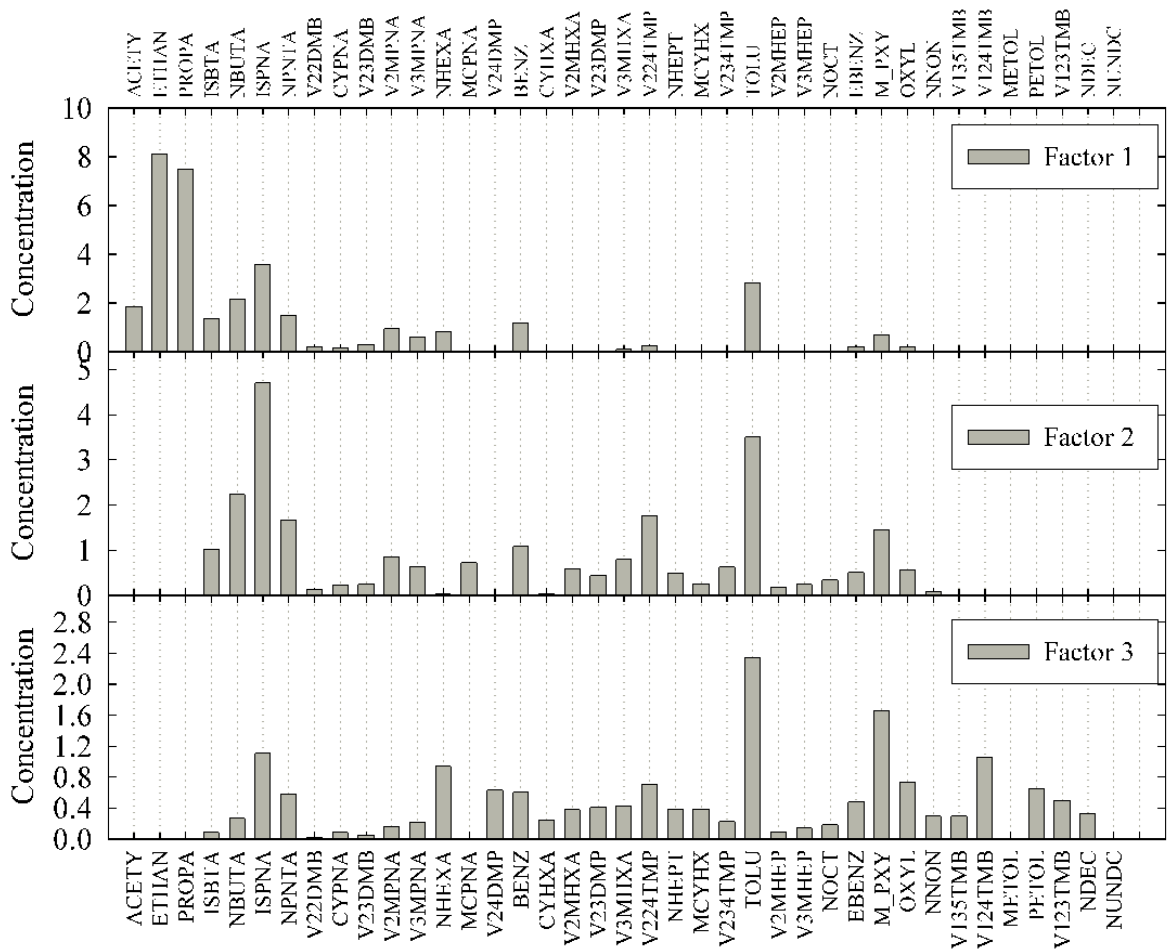


Figure 7. VOC profiles derived from the PAMS data from the Sherwood Island, CT site without the olefins.

REFERENCES

- Henry, R.C., C.W. Lewis, and J.F. Collins (1994) Vehicle-Related Hydrocarbon Source Compositions from Ambient Data: The GRACE/SAFER Methods, *Environ. Sci. Technol.* 28:823-832.
- Henry, R.C., C.H. Spiegelman, J.F. Collins, and E. Park (1997) Reported emissions of organic gases are not consistent with observations, *Proc. Natl. Acad. Sci. USA* 94:6596-6599.
- Lee, E., C.K. Chan, and P. Paatero (1999) Application of Positive Matrix Factorization in Source Apportionment of Particulate Pollutants in Hong Kong, *Atmospheric Environ.* 33:3201-3212.
- Paatero, P. (1997) Least Squares Formulation of Robust, Non-Negative Factor Analysis, *Chemom. Intell. Lab. Syst.* 37:23-35.
- Paterson, K.G. , J.L. Sagady, D.L. Hooper , S.B. Bertman , M.A. Carroll , and P.B. Shepson (1999) Analysis of Air Quality Data Using Positive Matrix Factorization, *Environ. Sci. Technol.* 33: 635-641
- Polissar, A.V., P.K. Hopke, and J.M. Harris (2001a) Source Regions for Atmospheric Aerosol Measured at Barrow, Alaska, *Environ. Sci. Technol.* 35: 4214-4226.
- Polissar, A.V. , P.K. Hopke, and R.L. Poirot (2001b) Atmospheric Aerosol over Vermont: Chemical Composition and Sources, *Environ. Sci. Technol.* 35: 4604-4621.
- Xie, Y. L., Hopke, P., Paatero, P., Barrie, L. A., and Li, S. M. (1999). Identification of source nature and seasonal variations of Arctic aerosol by positive matrix factorization. *J. Atmos. Sci.* 56:249-260.