Investigation of the relationship between chemical composition and size distribution of airborne particles by partial least squares and positive matrix factorization

Liming Zhou[1] and Philip K. Hopke

Center for Air Resources Engineering and Science and Department of Chemical Engineering, Clarkson University, Potsdam, New York, USA

Charles O. Stanier[2] and Spyros N. Pandis

Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

John M. Ondov and J. Patrick Pancras

Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland, USA

Received 20 May 2004; revised 17 July 2004; accepted 11 October 2004; published 9 March 2005.

[1] Two multivariate data analysis methods, partial least square (PLS) and positive matrix factorization (PMF), were used to analyze aerosol size distribution data and composition data. The relationships between the size distribution data and composition data were investigated by PLS. Three latent variables summarized chemical composition data and most variations in size distribution data especially for large particles and proved the existence of the linearity between the two data sets. The three latent variables were associated with traffic and local combustion sources, secondary aerosol, and coal-fired power plants. The size distribution, particle composition, and gas composition data were combined and analyzed by PMF. Source information was obtained for each source using size distribution and chemical composition simultaneously. Eleven sources were identified: secondary nitrate 1 and 2, remote traffic, secondary sulfate, lead, diesel traffic, coal-fired power plant, steel mill, nucleation, local traffic, and coke plant.

Citation: Zhou, L., P. K. Hopke, C. O. Stanier, S. N. Pandis, J. M. Ondov, and J. P. Pancras (2005), Investigation of the relationship between chemical composition and size distribution of airborne particles by partial least squares and positive matrix factorization, *J. Geophys. Res.*, *110*, D07S18, doi:10.1029/2004JD005050.

1. Introduction

[2] Multivariate receptor models are widely used in source apportionment of airborne particles [*Henry*, 1997, 2002; *Hopke*, 2003]. The measured chemical composition data from the samples collected at the receptor site form a matrix and this matrix can then be analyzed by UNMIX [*Henry*, 2003], PMF [*Paatero*, 1997] or other techniques to obtain two matrices representing source contribution and source profile, respectively. Recently, efforts have been made to use the methods to analyze size distribution data to identify sources [*Ruuskanen et al.*, 2001; *Wahlin et al.*, 2001; *Kim et al.*, 2004; *Zhou et al.*, 2004a, 2005].

[3] Even over a short distance (or transit time), there can be substantial changes in the size distributions of the particles emitted [*Zhu et al.*, 2002a, 2002b, 2004]. However, for the same location/transit time, the size distribution is very similar. If the size distribution coming from a source does not vary much with time, then the number concentration series measured at the receptor site have a linear relationship with the number contribution from all sources and also with their mass contributions. A previous application of multivariate receptor model with size distribution data [*Zhou et al.*, 2004a] has indicated that the number contribution of a source can be converted to its volume (mass) contribution by multiplying a constant determined by its size distribution profile.

[4] If there are linear relationships between the number concentrations and mass concentrations, it will be useful to combine the size distribution data and chemical composition data into a combined multivariate analysis. The source characteristics in both size distributions and chemical compositions may be obtained simultaneously and a better understanding of the source-receptor relationship will be provided.

[5] In this study, a small data set that includes both size distribution and composition data from the Pittsburgh Air Quality Study (PAQS) was analyzed by partial least squares (PLS) and positive matrix factorization (PMF). PLS is used to investigate the interrelationships between the number concentrations of all size intervals and the mass concen-

[1] Now at Providence Engineering and Environmental Group LLC, Baton Rouge, Louisiana, USA.

[2] Now at Department of Chemical and Biochemical Engineering, University of Iowa, Iowa City, Iowa, USA.

0148-0227/05/2004JD005050$09.00

Table 1. Missing Value Number (MN) of All Sizes and Species[a]

MN	Species	MN	Species	MN	Size, μm	MN	Size, μm	MN	Size, μm
9	Se	9	sulfate	24	0.168	29	0.0233	18	0.0032
9	Zn	0	nitrate	24	0.202	29	0.0279	18	0.0039
4	O_3	9	Al	24	0.242	29	0.0334	18	0.0046
2	NO	61	As	24	0.289	29	0.0399	18	0.0055
2	NO_x	9	Cd	24	0.3461	27	0.0478	18	0.0066
52	SO_2	9	Cr	24	0.414	26	0.0573	19	0.0079
1	CO	9	Cu	24	0.496	24	0.068	20	0.0095
3	$PM_{2.5}$	9	Fe	6	0.626	23	0.082	22	0.0113
		9	Mn	8	0.898	32	0.098	25	0.0136
		9	Ni	8	1.286	28	0.118	25	0.0163
		9	Pb	8	1.843	25	0.141	29	0.0194

[a] The total sample number is 240.

trations of all chemical species. Only if the PLS analysis can find linear relationships between the two data sets, can the two types of data can reasonably be combined and analyzed with a two-way receptor model. The results of the PMF analysis will be compared with the results in the work of *Zhou et al.* [2004b].

2. Data

[6] All the data used in this study were measured at Pittsburgh Supersite (latitude 40.4395, longitude −79.9405) on 16, 17, 18, 23 and 24 July 2001. The Pittsburgh Supersite was located in a park, around 6 miles to the east of the city center. The interstate highway, I376, extending from west to east, is around 1 to 2 km to the south of the site. There are secondary streets and minor roads rather close (<1 km) to the site. These days were chosen during the July 2001 intensive since there were complete and simultaneous measurements of both particle size distributions and chemical compositions only on these days. In particular, there was measurement of elemental species with high temporal resolution. Although these days may not represent the full month, the results will give us insights into the relationship between the size distributions and compositions of the aerosol in Pittsburgh area, and will also be useful in more completely understanding the prior size distribution analyses [*Zhou et al.*, 2004a, 2005]. The size distribution data were obtained from two scanning mobility particle spectrometers (SMPS) and an aerodynamic particle sampler (APS) with 15 min resolution. Above 583 nm, the data used in this study represent electrical mobility diameter inferred from their aerodynamic mobility and estimated density [*Khlystov et al.*, 2004]. The samples were collected at 25% relative humidity and "dry" particle distributions were obtained [*Stanier et al.*, 2004].

[7] The original size distribution data include 165 logarithmically even-spaced intervals from 0.003 μm to 2.5 μm. Every five consecutive size bins were combined into one and 33 new size intervals were produced [*Zhou et al.*, 2004a]. The 15 min number concentrations were averaged to 30 min and 240 samples were produced. The detailed description of the measurement of size distributions at Pittsburgh Supersite can be found elsewhere [*Stanier et al.*, 2004]. On 24 July there was a regional nucleation event with particle growth phenomenon and the number concentration data of that day were processed by the method introduced by *Zhou et al.* [2005].

[8] The composition data of $PM_{2.5}$, including both particle phase and gas phase, are the same as was used before in a multi time factor analysis [*Zhou et al.*, 2004b] except that all species with sampling period longer than 30 min, such as organic carbon/elemental carbon (OC/EC), were excluded in this study and all the concentrations used in this work are 30 min average. The missing values were replaced by the regressed values obtained in our previous studies [*Zhou et al.*, 2004b, 2005]. The aerosol composition data set includes sulfate and nitrate data obtained by continuous instruments of Aerosol Dynamics Inc. (ADI) [*Stolzenburg and Hering*, 2000] and metal species measured by the Semicontinuous Elements in Aerosol System (SEAS) [*Kidwell and Ondov*, 2001]. The complete description of all the measurement techniques can be found in the work of *Wittig et al.* [2003, 2004]. Table 1 summarizes the sizes and species that have been used as well as the number of missing values.

3. PLS

3.1. Method

[9] PLS is a basic tool of chemometrics for analyzing data with strongly collinear, noisy, and numerous X variables and simultaneously multiple response variables [*Wold et al.*, 2001]. For this analysis, we use X to stand for composition data and Y for size distribution data, where $X \in R^{m \times n}$ and $Y \in R^{m \times p}$ with m being the number of samples, n is the number of chemical species and p is the number of size intervals. The data in X and Y have been standardized from their original values so that each column vector in both matrices has a mean of 0 and a variance of 1. The model equations are as following:

$$X = TP' + E \tag{1}$$

$$Y = UC' + D \tag{2}$$

$$Y = TC' + H, \tag{3}$$

where T and U are score matrices, P and C are loading matrices, and E, D and H are residual matrices. In equations (1) and (2), T and U summarize the data in the X and Y matrices, respectively. Each of the column vectors in T, t_i, is called a latent variable (LV) that can be thought to be caused by a source or a source group, and so does the column vector in U, u_i. If T and U are very close, then T can also be used to explain Y and even to predict Y as indicated by equation (3). The above model is solved by the nonlinear iterative partial least squares (NIPALS)

Table 2. Correlations of Each Pair of Latent Variables *u* and *v*

Correlation Coefficient	LV
0.80	1
0.85	2
0.71	3
0.60	4
0.51	5
0.51	6
0.37	7
0.38	8
0.42	9
0.27	10
0.36	11
0.34	12

algorithm described by *Wold et al.* [2001]. Introductions to PLS are given in the work of *Manne* [1987] and *Jong* [1993]. An intuitive understanding of PLS but not a strict mathematical description can be provided as follows. The latent variables *t* and *u* in PLS analysis try to reproduce the variance in *X* and *Y*, respectively, and at the same time, each pair of *t* and *u* try to maximize their similarity (covariance) to each other. The similarities between *t* and *u* indicate similar latent structures in *X* and *Y*. For this specific study, similar *t* and *u* suggest they are both controlled by the same source or source group.

3.2. Results and Discussion

[10] Table 2 shows the correlation coefficients for each pair of latent variables (LV), t_i and u_i. After the first three latent variables, the correlations become poor, suggesting that there is no further relationships between the residual matrices. We define the two variables *Rx* and *Ry*:

$$Rx = 1 - \mathrm{var}(E)/\mathrm{var}(X) \tag{4}$$

$$Ry = 1 - \mathrm{var}(H)/\mathrm{var}(Y), \tag{5}$$

where var means variance. These two variables describe how much of the variance has been explained by the latent variables. As indicated in Table 3, when using three LVs, most of the variance in the *X* matrix has been explained, but much less of the variance in the *Y* matrix has been explained. This situation may suggest that the linear relationships amongst the chemical species are better than those amongst the sizes so that the size distribution data have larger residuals. Another reason is that number concentrations of all sizes were used in *Y* matrix, but not all chemical species were included in *X* matrix. Since some important species were not included in *X* matrix, such as OC/EC, some number concentrations in *Y* matrix will not be explained without those chemical species that characterize them. This is also one reason that PLS can identify less sources than PMF. Most of the variance of the small particles in the *Y* matrix is not explained since the small particles have little mass contribution. Thus the chemical species reflect the mass concentrations rather than the number concentrations. If they have no unique marker species, they only produce small variations in *X* matrix and hence are not summarized by the latent variables. This phenomenon also supports our prior results showing very weak correlations between number concentration and mass concentration for sources dominated by smallest sizes and largest number concentration [*Zhou et al.*, 2004a, 2005].

[11] Figure 1 shows the loadings of the first LV (the first column vectors of *P* and *C*). When the correlation coefficient between the LV score and the concentration series of a size interval or a species is over 0.7, a black bar is used to denote that value. The first LV (LV1) explains most of the variations of nitrate, NO, NO_x, and CO as well as several metal elements indicating emissions from traffic and other point combustion sources like the coke plant in the south. In our previous study [*Zhou et al.*, 2004b], As, Cd and Mn are associated with point industrial sources such as metal working, and Zn was thought to be from traffic in our previous multi time analysis [*Zhou et al.*, 2004b]. The size range of the first LV is wide, from 10 nm up to 200 nm, and this size range is also found to be related to traffic and other point sources by the analyses with only size distribution data [*Zhou et al.*, 2004a, 2005]. The high loadings between 1 to 2 μm are consistent with the volume size distribution profiles of the traffic and combustion sources [*Zhou et al.*, 2004a].

[12] The second LV (LV2) is mostly associated with sulfate and the size range 0.3 to 0.8 μm as indicated in Figure 2. These are particles from distant sources, converted from the precursor SO_2 via photochemical reactions during the transport [*Zhou et al.*, 2004a, 2005].

[13] The third LV (LV3) also explains some sulfate but more SO_2, as shown in Figure 3 and Table 4. The coal-fired power plants within 100 km from the receptor site are the probable sources. Because of the short distance, most of the SO_2 cannot be converted during the transport. The trimodal distribution implied by the size loadings are most likely caused by the conversions when the plume traveled from the source to the receptor. The newly formed particles are small while the aged ones are large. Since the growth of the particles is susceptible to meteorological and other condi-

Table 3. *Rx* and *Ry* of All Sizes and Species

Rx	Species	*Rx*	Species	*Ry*	Size, μm	*Ry*	Size, μm	*Ry*	Size, μm
0.45	Se	0.84	sulfate	0.62	0.168	0.26	0.0233	0.14	0.0032
0.70	Zn	0.81	nitrate	0.63	0.202	0.28	0.0279	0.23	0.0039
0.71	O_3	0.13	Al	0.67	0.242	0.24	0.0334	0.20	0.0046
0.76	NO	0.59	As	0.75	0.289	0.18	0.0399	0.14	0.0055
0.89	NO_x	0.71	Cd	0.69	0.3461	0.21	0.0478	0.04	0.0066
0.55	SO_2	0.44	Cr	0.63	0.414	0.22	0.0573	0.07	0.0079
0.71	CO	0.63	Cu	0.61	0.496	0.40	0.068	0.14	0.0095
		0.34	Fe	0.58	0.626	0.30	0.082	0.21	0.0113
		0.74	Mn	0.49	0.898	0.46	0.098	0.31	0.0136
		0.15	Ni	0.64	1.286	0.39	0.118	0.29	0.0163
		0.46	Pb	0.77	1.843	0.58	0.141	0.31	0.0194

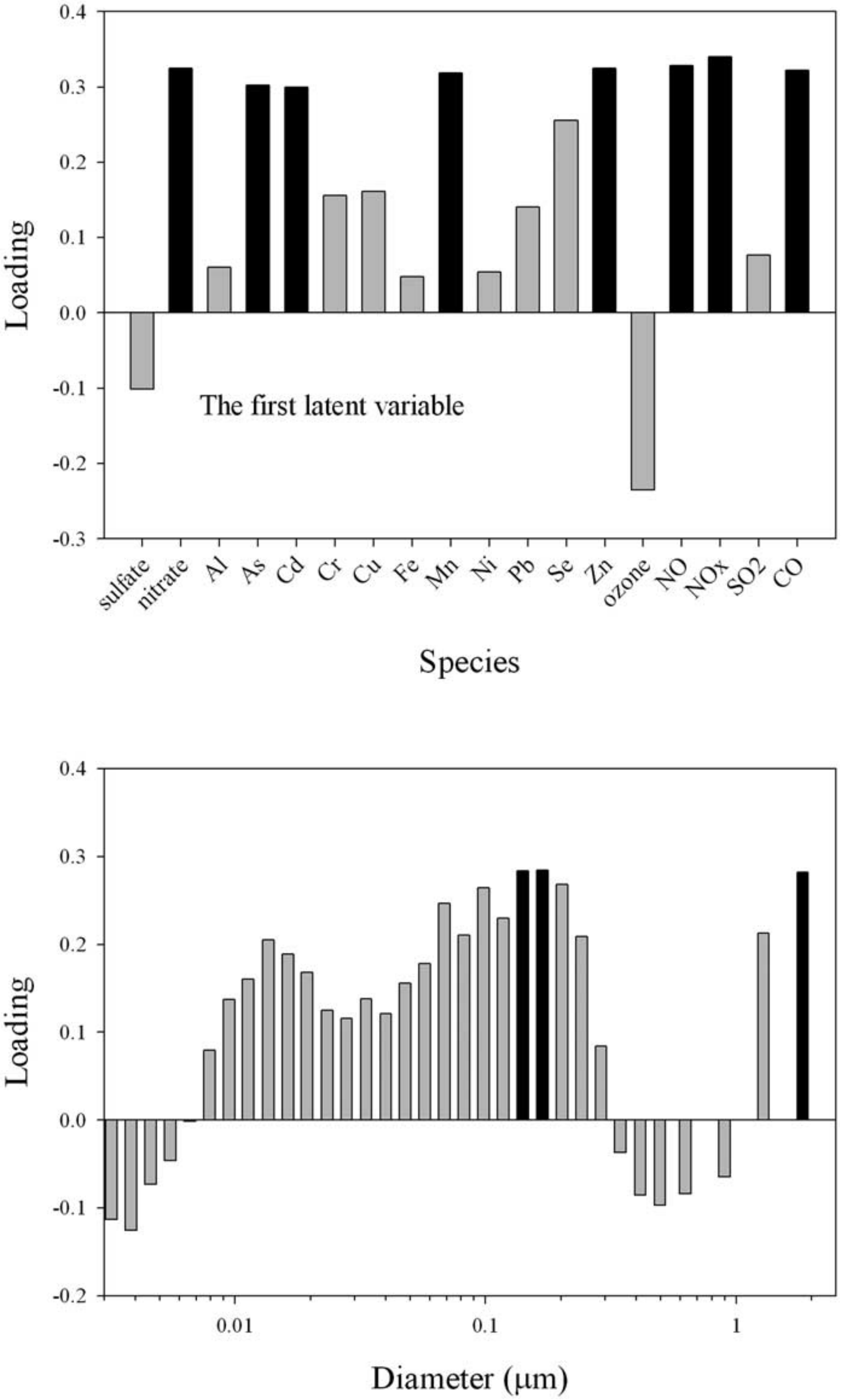

Figure 1. Loadings of the first LV for (top) chemical species and (bottom) sizes.

tions, the linearity between the number concentration and mass concentration is worse than for the first two LVs.

4. PMF

4.1. Method

[14] A two way receptor model was solved by PMF, an explicit least squares regression tool developed by *Paatero* [1997]. The model equation is:

$$X = GF + E, \qquad (6)$$

or in the elemental form,

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij}, \qquad (7)$$

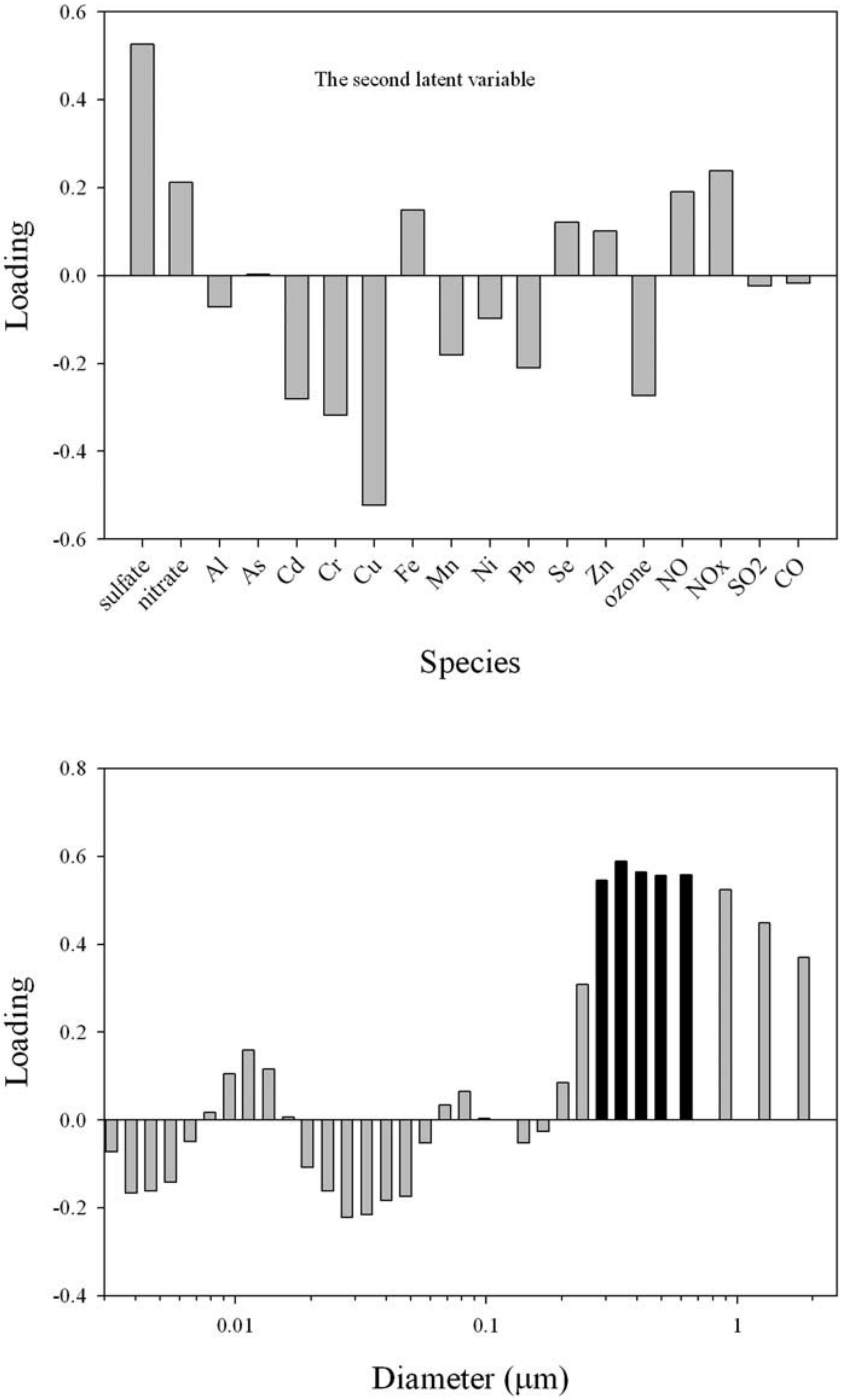

Figure 2. Loadings of the second LV for (top) chemical species and (bottom) sizes.

where X is the matrix of observed data, the element x_{ij} is the concentration value of the ith sample at the jth size interval or species. G and F are respectively the source contributions and source profiles to be estimated. E is a matrix of residuals.

[15] The residual sum of squares (Q) is defined by equation (8) and minimized by finding the optimal F and G.

$$Q = \left\| \frac{(X - GF)}{S} \right\|_{F,G}^{2} = \sum_{i} \sum_{j} \left(\frac{e_{ij}}{s_{ij}} \right)^{2}. \tag{8}$$

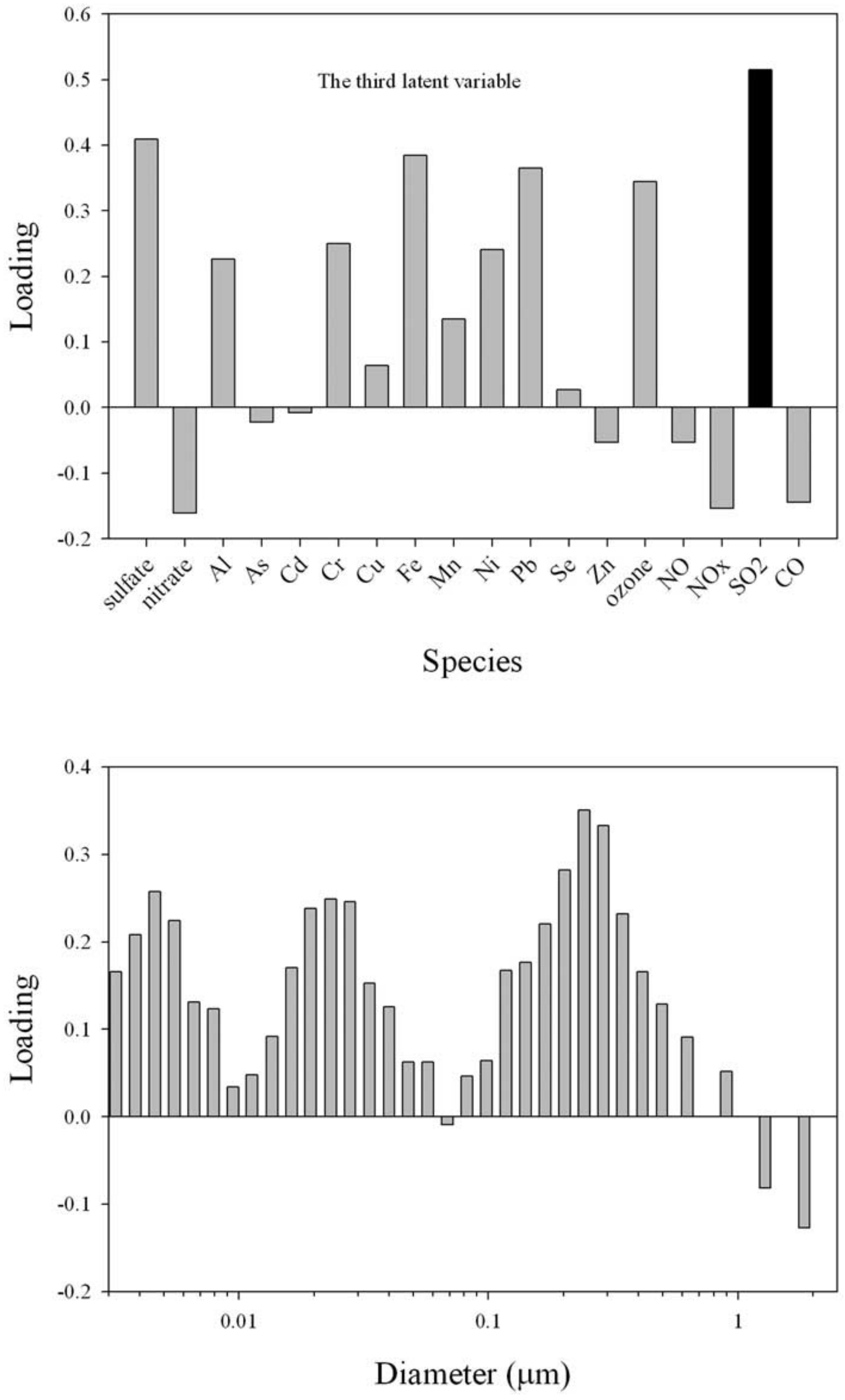

Figure 3. Loadings of the third LV for (top) chemical species and (bottom) sizes.

The uncertainties *s* were computed based on the measurement errors by equation (9):

$$s_{ij} = \sigma_{ij} + C_3 \max\left(|x_{ij}|, |y_{ij}|\right), \qquad (9)$$

where y_{ij} is the calculated value for x_{ij}, σ_{ij} is the measurement error, and C_3 is a dimensionless constant value, 0.08 in this study. The estimation of the measurement errors of size distribution data were based on the combination of size bins and the detailed procedure was provided in the work of *Zhou et al.* [2004a]. C_3 is used as the estimation of the relative uncertainties of large values (see P. Paatero, User's Guide for positive matrix factorization programs PMF2 and PMF3, Part 2: Reference,

Table 4. The Correlations of the Latent Variables by PLS With All Chemical Species

	LV1	LV2	LV3
Sulfate	−0.26	0.68	0.57
Nitrate	0.83	0.27	−0.22
Al	0.15	−0.09	0.31
As	0.77	0.00	−0.03
Cd	0.76	−0.36	−0.01
Cr	0.40	−0.41	0.35
Cu	0.41	−0.67	0.09
Fe	0.12	0.19	0.53
Mn	0.81	−0.23	0.19
Ni	0.14	−0.13	0.33
Pb	0.36	−0.27	0.51
Se	0.65	0.16	0.04
Zn	0.83	0.13	−0.07
Ozone	−0.60	−0.35	0.48
NO	0.83	0.25	−0.07
NO_x	0.86	0.31	−0.21
SO_2	0.19	−0.03	0.71
CO	0.82	−0.02	−0.20

available by FTP at ftp://ftp.clarkson.edu/pub/hopkepk/pmf/). FPEAK is a parameter in PMF for controlling rotations [*Paatero et al.*, 2002]. When the FPEAK value is positive, the following additional term is included in the object function *Q*:

$$Q^P = \beta^2 \left(\sum_{k=1}^{p} \sum_{j=1}^{n} f_{kj} \right)^2, \qquad (10)$$

where β^2 corresponds to the FPEAK value. The term defined above attempts to pull the sum of all the elements of *F* toward zero and makes the program do elementary transformations for *F* and *G* by subtracting the *F* vectors from each other and adding corresponding *G* vectors to obtain a more physically realistic solution. The FPEAK value was chosen as 0.1 since there was no clearly defined edges in *G* space [*Paatero et al.*, 2005].

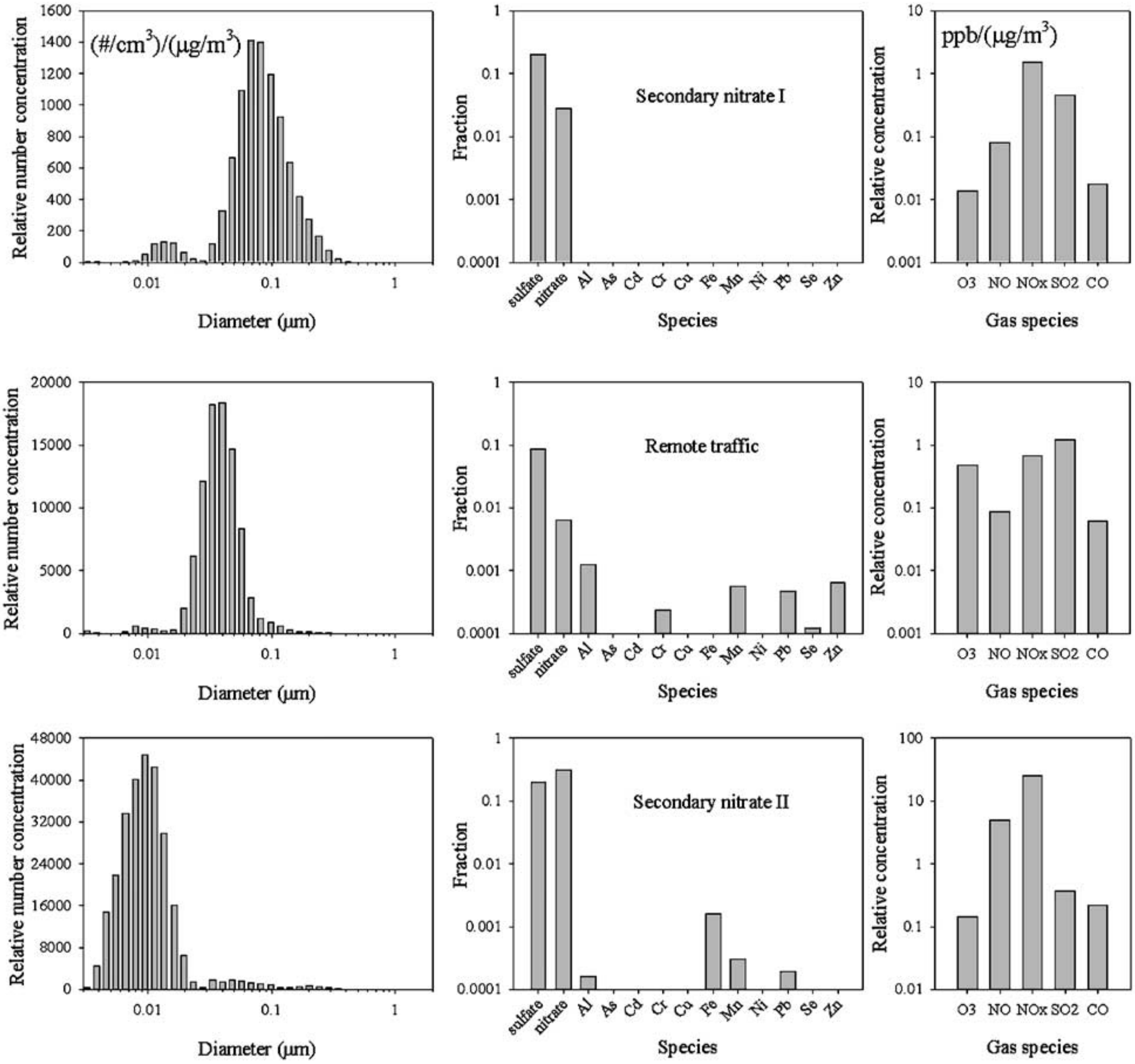

Figure 4. Source profiles from PMF analysis (secondary nitrate 1, remote traffic, and secondary nitrate 2).

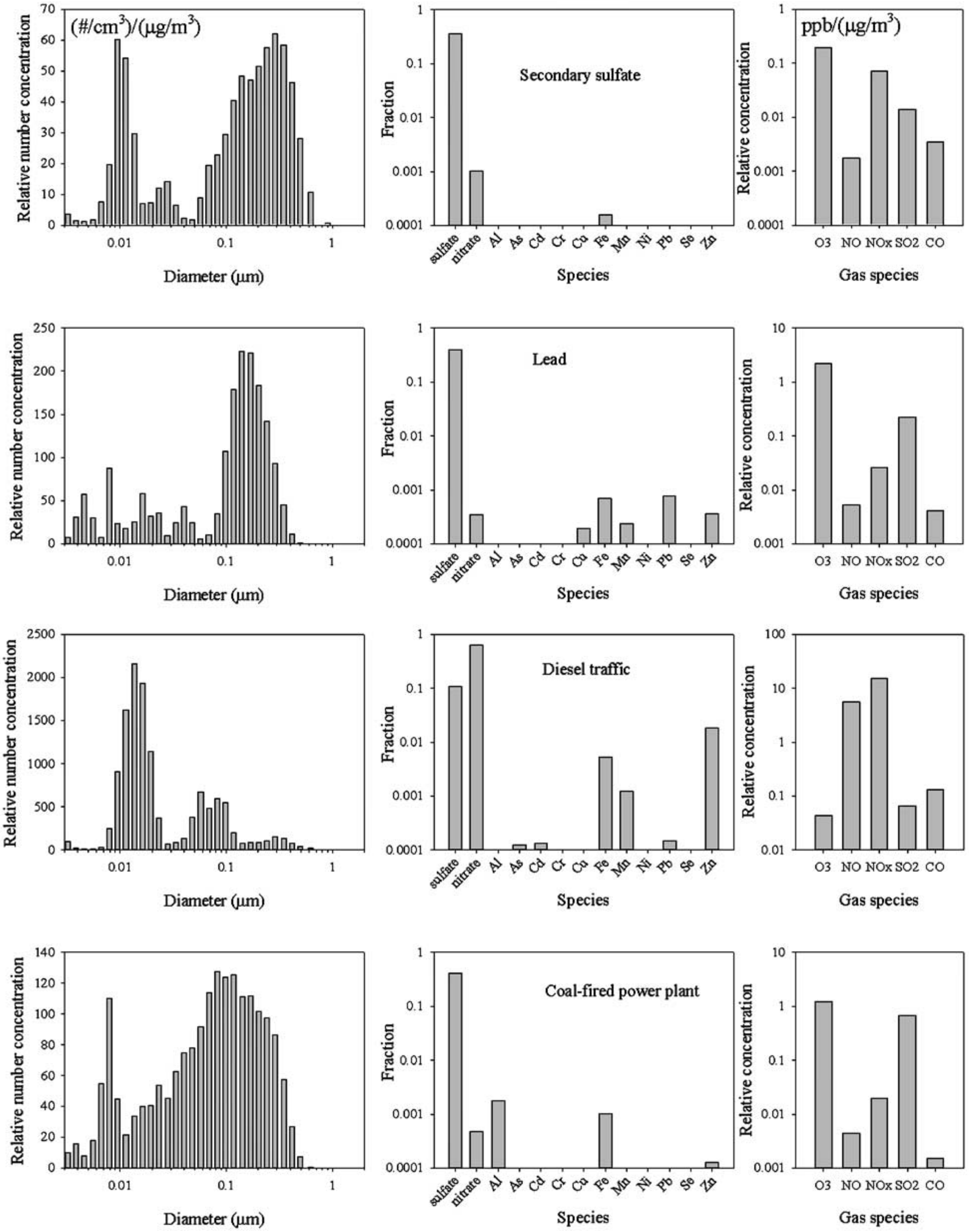

Figure 5. Source profiles from PMF analysis (secondary sulfate, lead, diesel traffic, and local coal-fired plant).

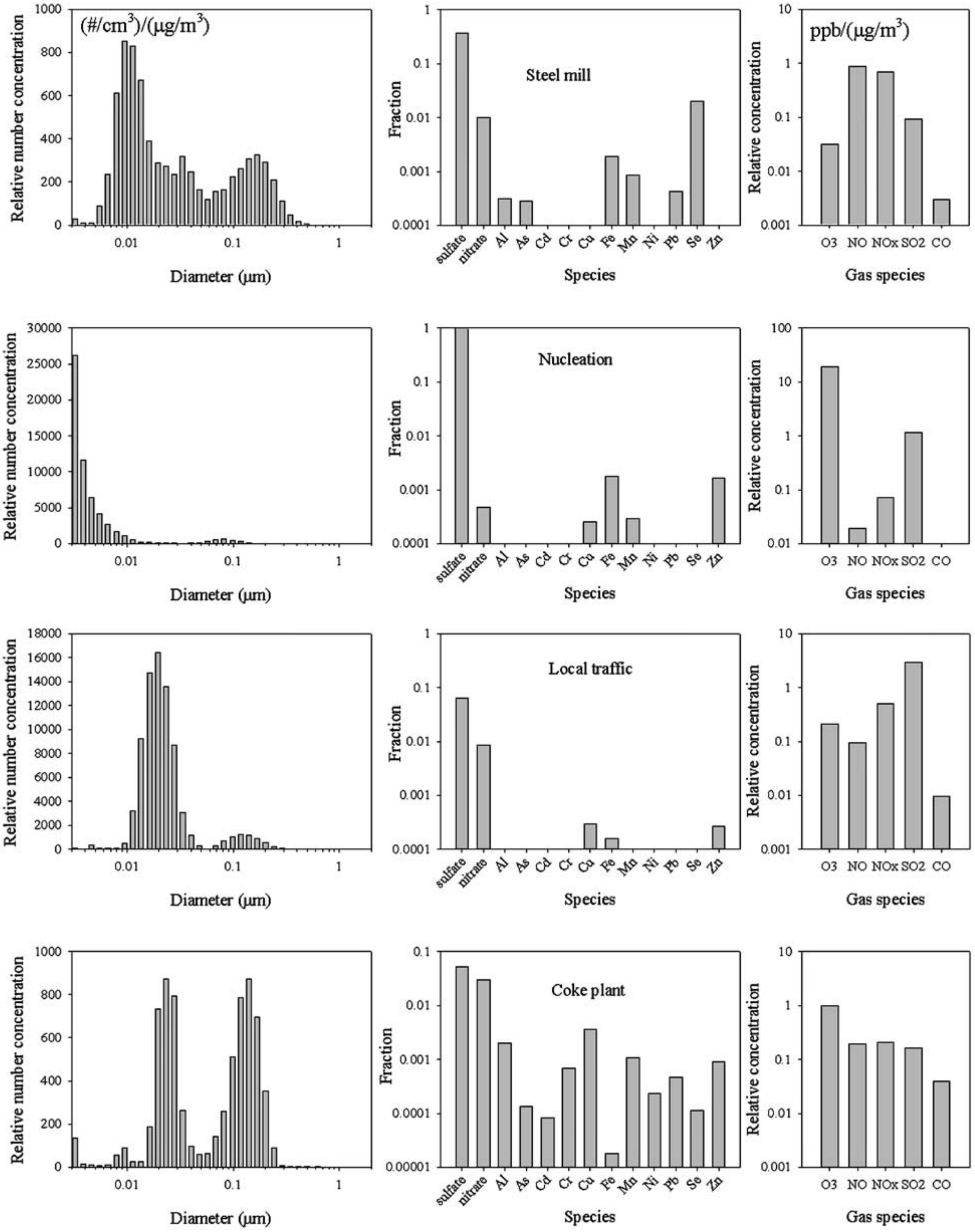

Figure 6. Source profiles from PMF analysis (steel mill, nucleation, local traffic, and coke plant).

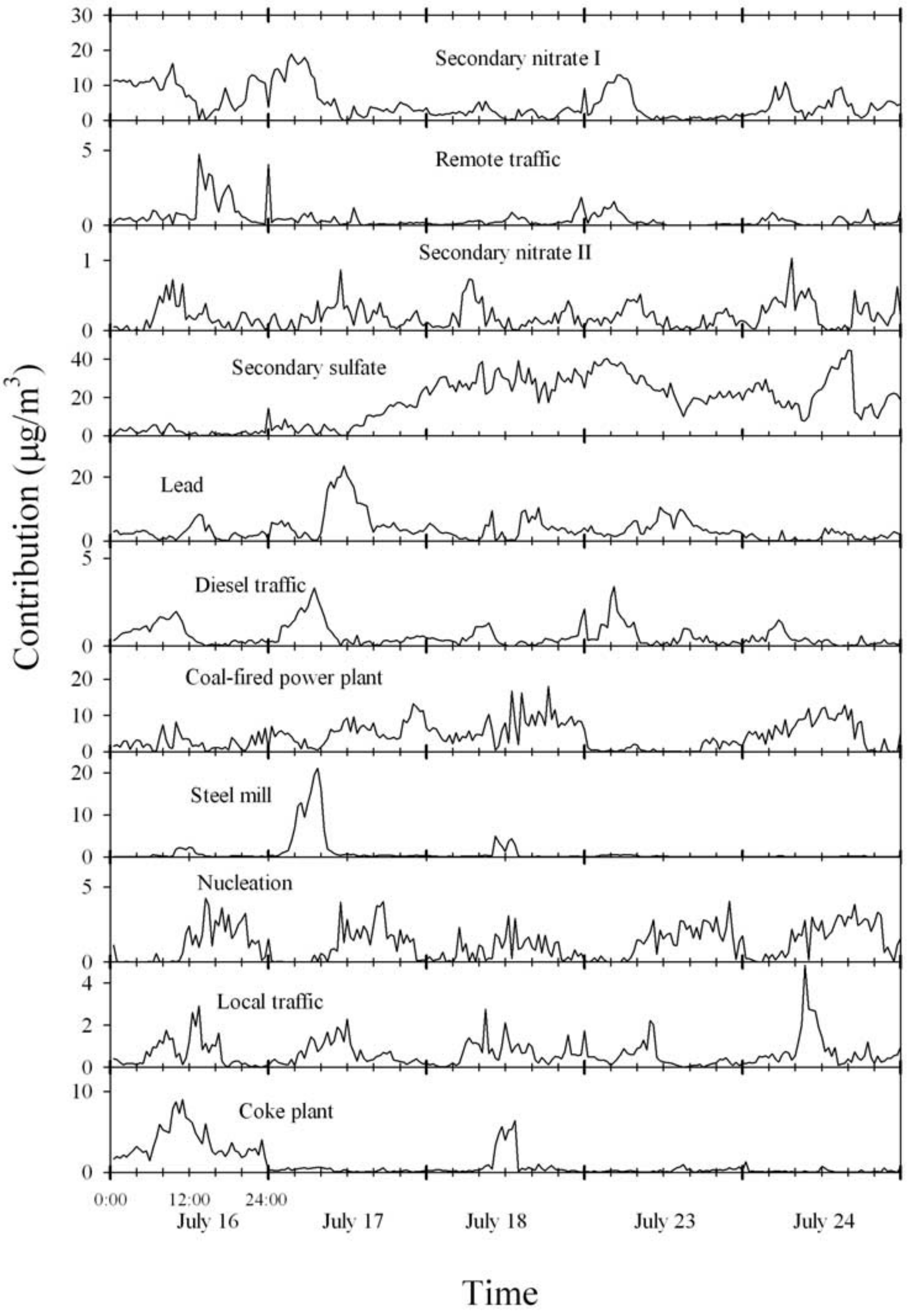

Figure 7. Source contributions from PMF analysis.

[16] The mass apportionment conditions [*Hopke et al.*, 1980] are satisfied by re-scaling the source contribution series and source profiles as shown in equation (11).

$$x_{ij} = \sum_{p=1}^{P} f_{ip} \cdot \frac{w_p}{w_p} \cdot g_{pj} \tag{11}$$

The scaling constants in the above equation, w_p, were determined by regressing PM$_{2.5}$ mass concentrations (v_j) against the estimated source contributions as indicated in equation (12).

$$v_j = \sum_{p=1}^{P} w_p \cdot g_{pj} \tag{12}$$

[17] The source profiles include three parts, number concentrations for all size intervals, mass fractions for all species and volume concentration for all gases. In Figures 4–6, for each source profile, the unit of the

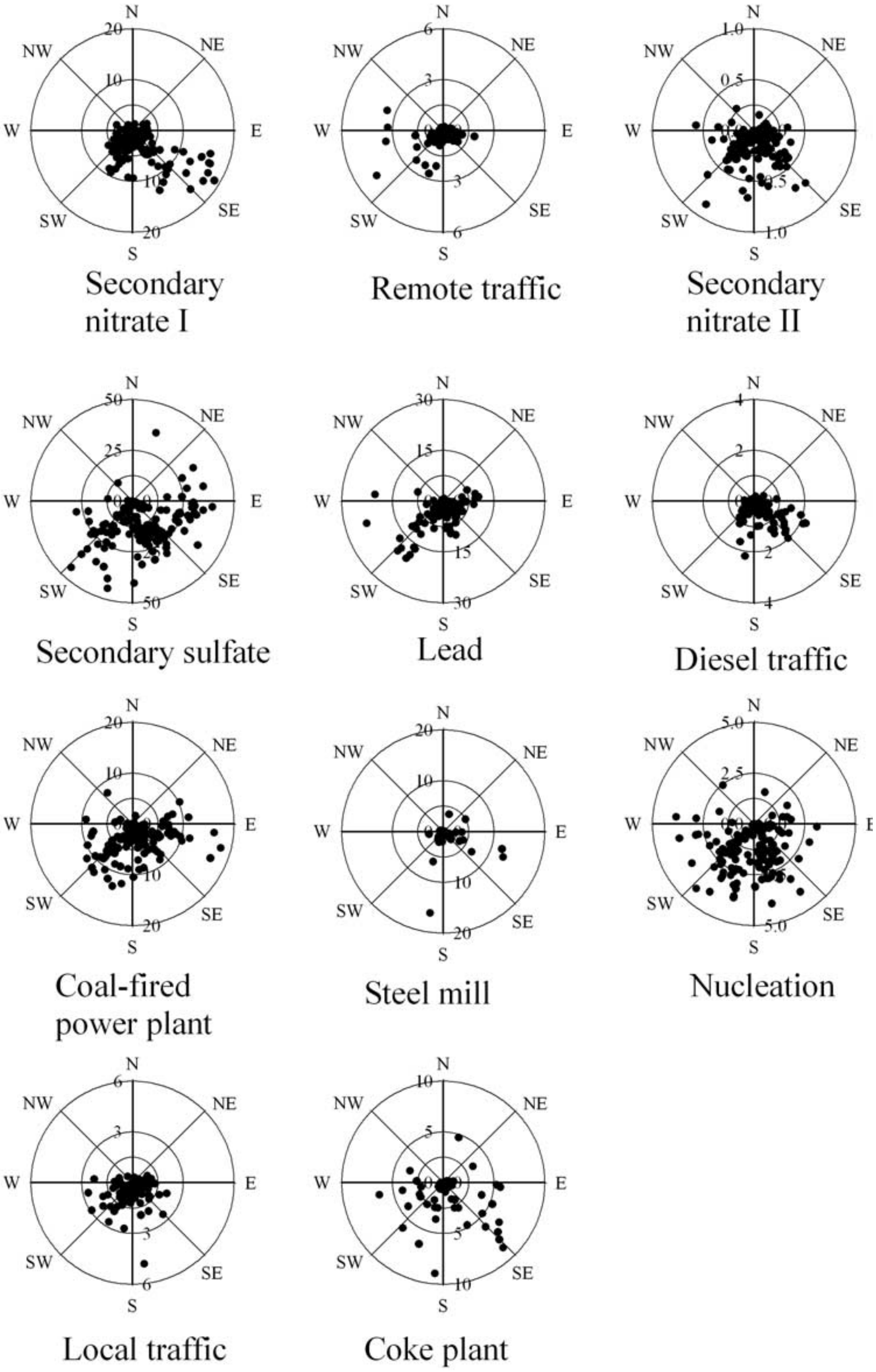

Figure 8. The relationships of source contributions with wind directions.

vertical axis in each of the three rows, from left to right, is (number/cm^3)/(μg/m^3), 1 and ppb/(μg/m^3) (for CO, the unit is ppm/(μg/m^3)), respectively.

4.2. Results and Discussion

[18] Eleven factors were found to provide the best solution. The sources are identified as secondary nitrate 1 and 2, remote traffic, secondary sulfate, lead, diesel traffic, coal-fired power plant, steel mill, nucleation, local traffic, and coke plant. When using an additional factor, the nucleation factor is separated into two factors and thus, additional factors were not warranted. With fewer factors, there were apparently mixed sources or poorly fit variables. The results will be discussed and compared with our previous PMF

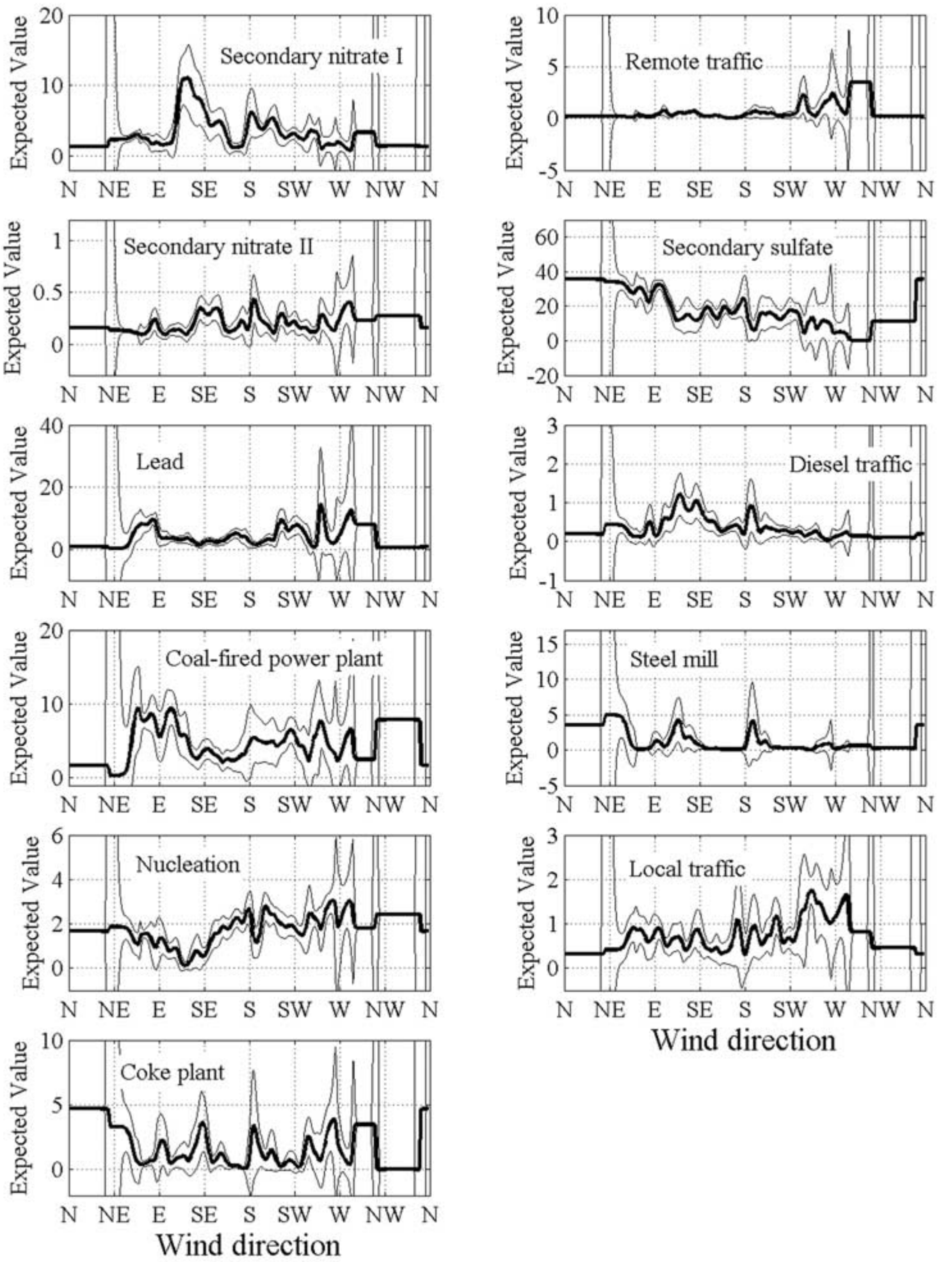

Figure 9. Nonparametric regression (NPR) analysis results for each source. (The unit of the expected value is μg/m^3.)

analyses of the size distribution data [*Zhou et al.*, 2004a, 2005] and multi time analyses of composition data [*Zhou et al.*, 2004b].

[19] The major mode of secondary nitrate 1 is at 0.08 μm as indicated in Figure 4. Figure 7 shows that the source contribution is high in the early morning when the low temperature favors the formation of nitrate. These particles are associated with aged NO_x emissions and grow into large sizes during the transport. In Figure 8, where the scatter plot of source contribution and corresponding wind directions are presented, the highest contribution is from southeast. Nonparametric regression (NPR) was also used to investigate the wind direction effects. In this method, averaged and smoothed source contributions are plotted against wind direction, confidential intervals are also given. High values indicate more transport from that direction. The detailed description of this method can be found elsewhere [*Henry et al.*, 2002; *Zhou et al.*, 2004b; *Kim and Hopke*, 2004] and will not be repeated here. The NPR results in Figure 9 also indicates that southeast is the major source direction of secondary nitrate 1, where two thin lines give the 95% confidential intervals.

[20] The size range of remote traffic is similar to our previous results of the size distribution data analysis [*Zhou et al.*, 2005], where the particles in this size range were found to be from traffic emissions several miles away or some other unknown point source emissions. The source contribution only has high peaks on the afternoon of 16 July

Table 5. Correlations of the Source Contributions by PMF With All Chemical Species

	Sulfate	Nitrate	Al	As	Cd	Cr	Cu	Fe	Mn	Ni	Pb	Se	Zn	O_3	NO	NO_x	SO_2	CO
Secondary nitrate 1	−0.26	0.61	−0.02	0.42	0.41	0.12	0.24	−0.12	0.42	−0.11	−0.03	0.34	0.51	−0.51	0.57	0.74	0.17	0.61
Remote traffic	−0.24	0.09	0.04	0.08	0.19	0.33	0.24	−0.04	0.27	0.03	0.10	0.01	0.14	−0.02	0.09	0.11	0.14	0.32
Secondary nitrate 2	−0.10	0.26	0.06	−0.09	0.17	−0.01	0.07	−0.08	0.05	−0.24	−0.01	−0.01	0.11	−0.26	0.30	0.33	0.17	0.18
Secondary sulfate	0.77	−0.11	−0.11	−0.34	−0.42	−0.23	−0.47	0.20	−0.50	0.02	−0.34	−0.26	−0.25	0.01	−0.17	−0.17	−0.10	−0.26
Lead	0.13	−0.21	0.04	0.15	0.16	−0.17	0.02	−0.06	0.11	0.11	0.81	−0.07	0.01	0.30	−0.16	−0.20	0.20	−0.21
Diesel traffic	−0.10	0.79	0.02	0.62	0.54	0.09	0.13	0.03	0.65	0.00	0.13	0.59	0.94	−0.65	0.75	0.84	−0.01	0.68
Local coal-fired plant	0.45	−0.17	0.60	−0.21	−0.20	−0.05	−0.03	0.17	−0.15	−0.08	−0.02	−0.14	−0.18	0.17	−0.21	−0.23	0.40	−0.19
Steel mill	−0.04	0.43	0.00	0.75	0.24	0.18	−0.02	0.32	0.72	0.15	0.30	0.98	0.57	−0.27	0.58	0.50	0.07	0.37
Nucleation	0.09	−0.55	−0.08	−0.34	−0.31	0.01	−0.01	0.06	−0.14	0.03	0.07	−0.19	−0.35	0.70	−0.34	−0.51	0.18	−0.43
Local traffic	−0.01	0.13	0.07	−0.03	0.12	0.16	0.09	0.13	0.13	−0.01	0.20	0.09	0.08	0.05	0.18	0.13	0.40	0.14
Coke plant	−0.40	0.24	0.24	0.21	0.64	0.88	0.70	0.28	0.48	0.43	0.20	0.02	0.21	−0.08	0.24	0.19	0.12	0.44

when there is much transport. The present species only explain a small fraction of the total particle mass contribution and some other species like OC/EC seem to be the major components. Figure 8 shows the source emission is from west, the direction of Pittsburgh city center, but the confidence interval at that direction in Figure 9 is large and this can be attributed to the short duration and small number of samples for the high source contribution episode. However, the major reason to think it as remote traffic is from the time series in the work of *Zhou et al.* [2005]. On the basis of the information from this study, we cannot exclude the possibility that it is from a point source.

[21] As shown in Figure 4, most particles of secondary nitrate 2 are at around 10 nm. This source has very high relative concentrations of NO and NO_x and also a very high fraction of nitrate. This source is associated with fresh NO_x emissions and the particles are formed in the vicinity of the receptor site and thus they have not much time to grow into large ones. The source contribution is high around the morning rush hour, implying the relationship with NO_x emissions from local traffic. Figures 8 and 9 indicate no clear dominating directions for this source.

[22] The secondary sulfate factor has a strong correlation with sulfate as shown in Table 5. It is composed of the largest particles and also has the largest mass contribution. This source corresponds to LV2 in the PLS analysis. The particles are formed during transport from distant sources. Compared with the other sources, this source has the lowest relative concentration of SO_2 around 0.01 ppb/(μg/m^3) as shown in Figure 5, while the relative concentration is 0.1 to 1 ppb/(μg/m^3) for the other sources. This situation suggests that most of the SO_2 is from local sources. It is shown in Figures 8 and 9 that this source seems to be from many directions.

[23] In our previous work [*Zhou et al.*, 2004b], lead was found to be from an unidentified source. Table 5 indicates a strong correlation between the source contribution and lead concentration series. The lead source here may be a local point source. Both Figures 8 and 9 indicate that the source is from southwest, where a metal working plant is located.

[24] Diesel traffic is similar to the traffic source identified before [*Zhou et al.*, 2004b] and both of them are strongly correlated with zinc. This source has the highest relative concentration for NO and NO_x as shown in Figure 5, and also has the highest correlation with NO and NO_x as shown in Table 5. In Figure 7, it can be found that the source contribution is only high in the early mornings. This phenomenon can be explained by the fact that the heavy-duty truck drivers avoid driving in the morning rush hours. The southeast and south directions indicated in Figures 8 and 9 are the directions of highway I376.

[25] The mode of coal-fired power plant is around 0.1 μm. This source is corresponding to LV3 in the PLS analysis. Its particle composition is close to secondary sulfate and it also explains some sulfate. However, its gas composition and size distribution profile is different from the secondary sulfate factor and that enables the separation into two factors. The dominating direction is between southwest and south may also include the direction between east and southeast. Since there are more than one coal-fired power plants near Pittsburgh area especially in the south and southwest directions [*Zhou et al.*, 2004a], this source may not be a single point source.

[26] Nucleation features the smallest particles. The small mode at 0.1 μm explains most of the related mass concentration and this mode is probably caused by simultaneous condensation with nucleation. The major chemical component is sulfate and this is consistent with other theoretical and experimental results on nucleation studies. The concentration of ozone for nucleation is the highest of all sources. Obviously, ozone is not a primary emission. Thus this profile suggests a strong relationship between ozone as a measure of photochemical activity and nucleation events. The elevated source contribution and ozone concentration can both be attributed to the increased photochemical reaction activities around noon as shown

in Figure 7. The nucleation source does not have clear dominating directions.

[27] The size range of local traffic, indicated in Figure 6, is similar to our previous results [*Zhou et al.*, 2004b], where we found the particles in this range showed strong diurnal patterns, including the contribution peak around morning rush hours and significant weekday/weekend difference. Figure 7 indicates that the temporal variations of the source contribution show peaks at morning rush hours. The correlations of the source contribution with the gases (NO, NO_x and CO) are weak as indicated in Table 5 since most of these gases are emitted from other sources such as diesel traffic.

[28] A coke plant and a steel mill are two sources that were also found by analyses of composition data. They have two number modes at 10–20 nm to and 0.1–0.2 μm. In the analyses of size distribution data [*Zhou et al.*, 2004a, 2005], the particles with the size range around 0.1 μm were thought to be from the local combustion sources. This conclusion is consistent with the presence of the number modes at the large size while the number modes at 10–20 nm were not found to be related to point sources by the analysis with only size distribution data.

[29] The two point sources, coke plant and steel mill, as well as secondary nitrate 1 and the diesel traffic are included in LV1. Their size ranges are similar and are thus summarized in one latent variable in the PLS analysis. These sources explain most variations of NO, NO_x and CO.

5. Conclusion

[30] Partial least squares and positive matrix factorization have been used to analyze aerosol sized distribution data and composition data together. PLS analyses found there are linear relationships between the number concentrations of large sized particles and the mass concentrations of most of the chemical species. Since the linear relationship between the two data sets was proved by PLS, PMF can be used for source apportionment and it can even identify the sources with small chemical mass concentrations but high number concentrations caused by small particle sizes, such as nucleation and local traffic.

[31] The two methods have revealed source information including both size distribution and chemical composition at the same time. These results are helpful for understanding the results by the analysis of size distribution data.

[32] **Acknowledgments.** This research was conducted as part of the Pittsburgh Air Quality Study which was supported by U.S. Environmental Protection Agency under contract R82806101 and the U.S. Department of Energy National Energy Technology Laboratory under contract DE-FC26-01NT41017. This paper has not been subject to EPA's required peer and policy review, and therefore does not necessarily reflect the views of the agency. No official endorsement should be inferred.

References

Henry, R. C. (1997), History and fundamentals of multivariate air quality receptor models, *Chemom. Intell. Lab. Syst.*, *37*, 525–530.

Henry, R. C. (2002), Multivariate receptor models—Current practice and future trends, *Chemom. Intell. Lab. Syst.*, *60*, 43–48.

Henry, R. C. (2003), Multivariate receptor modeling by N-dimensional edge detection, *Chemom. Intell. Lab. Syst.*, *65*, 179–189.

Henry, R. C., Y.-S. Changa, and C. H. Spiegelman (2002), Locating nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction, *Atmos. Environ.*, *36*, 2237–2244.

Hopke, P. K. (2003), Recent developments in receptor modeling, *J. Chemom.*, *17*, 255–265.

Hopke, P. K., R. E. Lamb, and D. F. C. Natusch (1980), Multielemental characterization of urban roadway dust, *Environ. Sci. Technol.*, *14*, 164–172.

Jong, S. (1993), SIMPLS: An alternative approach to partial least squares regression, *Chemom. Intell. Lab. Syst.*, *18*, 251–263.

Khlystov, A., C. Stanier, and S. N. Pandis (2004), An algorithm for combining electrical mobility and aerodynamic size distributions when measuring ambient aerosol, *Aerosol. Sci. Technol.*, *38*, 229–238.

Kidwell, C. B., and J. M. Ondov (2001), Elemental analysis of sub-hourly ambient aerosol collections, *Aerosol Sci. Technol.*, *35*, 596–601.

Kim, E., and P. K. Hopke (2004), Comparison between conditional probability function and nonparametric regression for fine particle source directions, *Atmos. Environ.*, *38*, 4667–4673.

Kim, E., P. K. Hopke, T. V. Larson, and D. S. Covert (2004), Analysis of ambient particle size distributions using UNMIX and positive matrix factorization, *Environ. Sci. Technol.*, *38*, 202–209.

Manne, R. (1987), Analysis of two partial-least-squares algorithms for multivariate calibration, *Chemom. Intell. Lab. Syst.*, *2*, 187–197.

Paatero, P. (1997), Least squares formulation of robust non-negative factor analysis, *Chemom. Intell. Lab. Syst.*, *37*, 23–35.

Paatero, P., P. K. Hopke, X. H. Song, and Z. Ramadan (2002), Understanding and controlling rotations in factor analytic models, *Chemom. Intell. Lab. Syst.*, *60*, 253–264.

Paatero, P., P. K. Hopke, B. A. Begum, and S. K. Biswas (2005), A graphical diagnostic method for assessing the rotation in factor analytical models of atmospheric pollution, *Atmos. Environ.*, *39*, 193–201.

Ruuskanen, J., et al. (2001), Concentrations of ultrafine, fine and PM2. 5 particles in three European cities, *Atmos. Environ.*, *35*, 3729–3738.

Stanier, C., A. Khlystov, W. R. Chan, M. Mandiro, and S. N. Pandis (2004), A method for the in-situ measurement of fine aerosol water content of ambient aerosol: The Dry-Ambient Aerosol Spectrometer (DAASS), *Aerosol Sci. Technol.*, *38*(S1), 215–228.

Stolzenburg, M. R., and S. V. Hering (2000), Method for the automated measurement of fine particle nitrate in the atmosphere, *Environ. Sci. Technol.*, *34*, 907–914.

Wahlin, P., F. Palmgren, R. V. Dingenen, and F. Raes (2001), Experimental studies of ultrafine particles in streets and the relationship to traffic, *Atmos. Environ.*, *35*(S1), S63–S69.

Wittig, B., N. Anderson, A. Y. Khlystov, S. N. Pandis, C. Davidson, and A. L. Robinson (2003), Pittsburgh Air Quality Study overview and preliminary scientific findings, *Atmos. Environ.*, *38*, 3107–3125.

Wittig, A. E., S. Takahama, A. Y. Khlystov, S. N. Pandis, S. Heringe, B. Kirbye, and C. Davidson (2004), Semi-continuous PM2.5 inorganic composition measurements during the Pittsburgh Air Quality Study, *Atmos. Environ.*, *38*, 3201–3213.

Wold, S., M. Sjöström, and L. Eriksson (2001), PLS-regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, *58*, 109–130.

Zhou, L., E. Kim, P. K. Hopke, C. Stanier, and S. N. Pandis (2004a), Advanced factor analysis on Pittsburgh particle size distribution data, *Aerosol Sci. Technol.*, *38*, 118–132.

Zhou, L., P. K. Hopke, P. Paatero, J. M. Ondov, P. J. Pancras, N. J. Pekney, and C. I. Davidson (2004b), Advanced factor analysis for multiple time resolution aerosol composition data, *Atmos. Environ.*, *38*, 4909–4920.

Zhou, L., E. Kim, P. K. Hopke, C. Stanier, and S. N. Pandis (2005), Mining airborne particulate size distribution data by positive matrix factorization, *J. Geophys. Res.*, doi:10.1029/2004JD004707, in press.

Zhu, Y., W. C. Hinds, S. Kim, S. Shen, and C. Sioutas (2002a), Study of ultrafine particles near a major highway with heavy-duty diesel traffic, *Atmos. Environ.*, *36*, 4323–4335.

Zhu, Y., W. C. Hinds, S. Kim, and C. Sioutas (2002b), Concentration and size distribution of ultrafine particles near a major highway, *J. Air Waste Manage. Assoc.*, *52*, 1032–1042.

Zhu, Y., W. C. Hinds, S. Kim, and C. Sioutas (2004), Seasonal trends of concentration and size distribution of ultrafine particles near major highways in Los Angeles, *Aerosol Sci. Technol.*, *38*(A1), 5–13.

P. K. Hopke, Center for Air Resources Engineering and Science, Department of Chemical Engineering, Clarkson University, PO Box 5708, Potsdam, NY 13699-5708, USA. (hopkepk@clarkson.edu)

J. M. Ondov and J. P. Pancras, Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, USA.

S. N. Pandis, Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

C. O. Stanier, Department of Chemical and Biochemical Engineering, University of Iowa, 4122 Seamans Center, Iowa City, IA 52242, USA. (cstanier@engineering.uiowa.edu)

L. Zhou, Providence Engineering and Environmental Group LLC, 6160 Perkins Road, Suite 100, Baton Rouge, LA 70808, USA. (limingzhou@providencebr.com)