Aerosol and Air Quality Research, 14: 2040–2050, 2014

ISSN: 1680-8584 print / 2071-1409 online
doi: 10.4209/aaqr.2014.01.0024

A Comparison of Multiple Combined Models for Source Apportionment, Including the PCA/MLR-CMB, Unmix-CMB and PMF-CMB Models

Guo-Liang Shi[1], Gui-Rong Liu[1], Xing Peng[1], Yi-Nan Wang[2], Ying-Ze Tian[1*], Wei Wang[2*], Yin-Chang Feng[1]

[1] *State Environmental Protection Key Laboratory of Urban Ambient Air Particulate Matter, Pollution Prevention and Control, College of Environmental Science and Engineering, Nankai University, Tianjin, 300071, China*
[2] *College of Software, Nankai University, No. 94 Weijin Road, Tianjin 300071, China*

ABSTRACT

A combined models was developed and applied to synthetic and ambient PM datasets in our prior works. In this study, multiple combined models, including the PCA/MLR-CMB, Unmix-CMB and PMF-CMB models, were developed and employed to analyzed the synthetic datasets, in order to understand 1) the accuracies of the predictions by multiple combined models; 2) the effect of Fpeak-rotation on the predictions of the PMF-CMB model; and 3) the relationship between the extracted mixed source profiles (in the first stage) and the final predictions. 50 predictions based on different combined model solutions were obtained and compared with the synthetic datasets. The average absolute errors (AAE), cluster analysis (CA), and PCA plots were applied to evaluate the precision of the predictions. These statistical methods showed that the predictions of the PCA/MLR-CMB and PMF-CMB model (with Fpeaks from 0 to 1.0) were satisfactory, those of the Unmix-CMB model were instable (some of them closely approached the synthetic values, while other them deviated from them). Additionally, it was found that the final source contributions had good correlation with their marker concentrations (obtained in the first stage), suggesting that the extracted profiles of the mixed sources can determine the final predictions of combined models.

Keywords: Receptor models; Synthetic datasets; Mixed source; Fpeak.

INTRODUCTION

Particulate matter (PM_{10} and $PM_{2.5}$) is the important pollutant in urban air ambient (Shen *et al.*, 2010; Kong *et al.*, 2011; Zheng *et al.*, 2013). Long-term exposure to particulate matter air pollution can result in increased risk of human mortality (Ozkaynak and Thurston, 1987; Russell, 2009; Yan *et al.*, 2009; Tie *et al.*, 2009; Habre *et al.*, 2011; Shen *et al.*, 2011; Cheng *et al.*, 2012; Amodio *et al.*, 201; Vernile *et al.*, 2013). In order to reduce the PM pollution, understanding the potential source categories and their contributions (source apportionment) is necessary (Zheng *et al.*, 2005). The result of source apportionment can provide the scientific supporting for air quality management decisions.

Receptor models, the useful tools for source apportionment, utilize the chemical composition of receptors for identification and apportionment of sources of PM in the atmosphere (Zheng *et al.*, 2007; Ke *et al.*, 2008; Kong *et al.*, 2010; Pant and Harrison, 2012). Among several receptor models, two main classes of models have been employed widely over the world (Hopke, 2003; Andriani *et al.*, 2011; Pant and Harrison, 2012). That is, i) Chemical Mass Balance (CMB) model and ii) multivariate factor analysis models (including Principal Component Analysis/ Multiple Linear Regression (PCA/MLR), UNMIX, and Positive Matrix Factorization (PMF)). The first class of models need both the input data of receptor and the source profiles; while the later class of models extracts source profiles and their contributions over sets of receptor samples (Hopke, 2003). The detailed introductions of the principle and applications for CMB, PCA/MLR, Unmix and PMF models have been presented in literature (Watson, 1984; Paatero and Tapper, 1994; Lee *et al.*, 1999; Watson and Chow, 2001; Song *et al.*, 2006; Chen *et al.*, 2007; Zheng *et al.*, 2007; Begum *et al.*, 2010; Harrison *et al.*, 2011; Gugamsetty *et al.*, 2012) and our prior publications (Shi *et al.*, 2011; Zhang *et al.*, 2011; Shi *et al.*, 2012; Wang *et al.*, 2012; Zhang *et al.*, 2012).

The strengths and weaknesses for the two classes of receptor models have been summarized in literature (Hopke, 2003; Pant and Harrison, 2012). Multicollinearity, arising when two different sources have similar profiles, often

* Corresponding author.
Tel.: +8602223503397
E-mail address: tianyingze@hotmail.com (Y.Z. Tian); kevinwangwei@nankai.edu.cn (W. Wang)

disturbs the predictions of the two classes of receptor models. For CMB model, near collinear sources can result in incorrect source contributions. While for multivariate factor analysis models, near collinear sources will be extracted in one factor, due to their similar signatures. In this case, a factor that contains two or more sources is usually indentified as a mixed source.

To resolve the multicollinearity, the Factor analysis-CMB combined models were developed in our prior studies (Shi *et al.*, 20091; Zeng *et al.*, 2010; Shi *et al.*, 201). The combined model was tested, and acceptable results were obtained, using synthetic datasets with collinearity (Shi *et al.*, 2009, 2011). The methodology has been quoted in a review paper on source apportionment methods by Pant and Harrison (Pant and Harrison, 2012) and involved in source apportionment works for air pollution in some cities (Shi *et al.*, 2009; Zeng *et al.*, 2010; Shi *et al.*, 2011). In our prior works (Shi *et al.*, 2009; Zeng *et al.*, 2010; Shi *et al.*, 2011), the factor analysis (PCA/MLR and PMF) and CMB model were combined together for source apportionment. On the first stage, the mixed source (containing near collinear sources) was extracted by the PMF or PCA/MLR models; and on the second stage the mixed source is treated as a new receptor and be apportioned by CMB model. For instance, Shi *et al.* (2009) applied PCA/MLR-CMB model and PMF-CMB model to the sources of PM_{10} in Zhengzhou city. On the first stage, PMF and PCA/MLR were respectively employed to identify the sources of PM_{10} in Zhengzhou city. Similar sources were obtained from two models, including the mixed source, vehicle exhaust, residual oil and secondary sulfate. Then on the second stage, the mixed source was applied as a secondary receptor and was introduced into CMB model, and soil dust, coal combustion and cement dust were identified. Although similar source categories were obtained by PCA/MLR-CMB and PMF-CMB, the slightly different source contributions were obtained (Shi *et al.*, 2009). So, the accuracy of the predictions from different combined models needs to be compared. However, the test for combined models is still very limited.

For assessing the predictions from different combined models, some issues should be focused on: (1) PCA/MLR-CMB, PMF-CMB and Unmix-CMB usually obtain different results for the same input dataset set (Shi *et al.*, 2009). So, how are the accuracies of the predictions by different combined models? (2) The extracted profiles by PMF can be rotated by setting Fpeak (Paatero *et al.*, 2004). So, how is the effect of Fpeak-rotation on the prediction of PMF-CMB model? (3) The profile of mixed source which is extracted on the first stage can determine the final results of the combined model on the second step. So, how is the relationship between the mixed source profiles and the final predictions?

In this work, three kinds of combined models (PCA/MLR-CMB, Unmix-CMB and PMF-CMB) were employed to analyze synthetic datasets in order to quantify the contribution of the source to the synthetic datasets and to understand accuracies of the predictions by multiple combined models. For this purpose, five synthetic datasets were developed firstly. Then, PMF-CMB model solutions were performed by setting a set of different Fpeak values, to study the impacts of Fpeak-rotation on the prediction of PMF-CMB model. The predictions would be compared with the synthetic contributions to estimate the accuracy of the results by multiple combined models. Next, the variety of extracted profiles (on the first stage) and final predictions of PMF-CMB with different Fpeaks were discussed, to analyze the theoretical causes for different predictions. Finally, the correlations between final predicted source contributions and the estimated concentrations of marker species in the mixed source (extracted on the first stage) were analyzed. The findings in this work can provide useful information for the application of multiple combined models.

METHODS

Principle of Combined Model

As described in our prior studies (Shi *et al.*, 2009; Zeng *et al.*, 2010; Shi *et al.*, 2011), the combined model mainly contains two stages.

For the first stage, the PM receptor dataset was introduced into the factor analysis model (PCA/MLR, Unmix or PMF). The factor profiles and contributions can be calculated, as follows:

$$X_{(n\times m)} = G_{(n\times p)}F_{(p\times m)} + E_{(n\times m)} \quad (1)$$

where, n is the number of samples; m is the number of the chemical components; p is the number of the extracted factors; $X_{(n\times m)}$ is the matrix of ambient concentrations with m species and n samples; F is the profile matrix for extracted factors (sources) with m species and p factors, G is the contributions matrix for extracted factors (sources) with n samples and p factors and E is the residual matrix with m species and n samples.

The extracted factors by the factor analysis model can be identified as different source categories, basing on the marker species. Pant and Harrison have summarized some commonly source profiles and their key marker species (Pant and Harrison, 2012). For the extracted factors, if one factor can be identified as one source category, it is referred to as an extracted simplex source; whereas if it contains two or more source categories, it is referred to as an extracted mixed source. The sources included in the mixed source can be called sub-sources.

For the second stage, the mixed source was treated as a new receptor. The profile and its uncertainty were calculated on the first stage, according to our prior studies (Shi *et al.*, 2009; Zeng *et al.*, 2010; Shi *et al.*, 2011). Then, according to the mixed source profile as well as the emission inventory of the monitory area, the categories of the sub-sources can be identified, and the sub-sources profiles should be measured in the studied area. Next, the profiles of new receptor (mixed source extracted on the first stage) and the sub-source were introduced into the CMB model. And the contributions of the sub-sources to the new receptor would be calculated by CMB model, as follows (US EPA, 2004):

$$S_{(k\times l)} = (F^{T}_{(K\times m)}(V_{e(\mathrm{m}\times\mathrm{m})})^{-1}F_{(\mathrm{m}\times\mathrm{k})})^{-1}F^{T}_{(K\times m)}(V_{e(m\times m)})^{-1}C_{(m\times l)} \quad (2)$$

where, k is the number of the sub-sources; $S_{(k\times 1)}$ is the vector of the estimated source contributions of k sub-sources; $C_{(m\times 1)}$ is the profile of new receptor (mixed source in stage 1); $F_{(m\times k)}$ profile matrix of k sub-sources; and V_e is the effective variance matrix.

Finally, the contributions of the simplex sources obtained on the first stage, as well as the contributions of the sub-sources predicted on the second stage were the final results of combined model.

In this work, the PCA/MLR-CMB, PMF-CMB and Unmix-CMB models were developed and tested. The SPSS 16.0, PMF2, EPA Unmix 6.0 and EPA CMB 8.2 softwares were employed.

Development of the Synthetic Receptor Datasets

In order to compare the results of different combined models, five synthetic receptor datasets (Datasets A to E) were developed. The synthetic receptor datasets were generated by the actual primary profiles and secondary sources. For each synthetic receptor dataset, seven actual PM_{10} source categories were included: resuspended dust (RD), soil dust, coal combustion, cement dust, vehicle exhaust, secondary sulfate and secondary nitrate. The actual primary source profiles were obtained in five different cities in China (see Tables S1–S5 in Supplementary Material), and reported in our prior studies (Bi *et al.*, 2007). For each source profile, 24 species were included (Table S6).

The construction of the synthetic datasets referred to the relative literature and our prior works (Brinkman *et al.*, 2006; Shi *et al.*, 2011). An $n \times m$ matrix (n is the number of samples and m is the number of chemical species) of concentrations ($\mu g/m^3$) X was developed as follows:

$$X_{(n\times m)} = G_{(n\times p)}F_{(p\times m)} + E_{(n\times m)} \quad (3)$$

where p is the number of the sources; the $X_{(n\times m)}$,$G_{(n\times p)}$ and $F_{(p\times m)}$ appear in similar role as ambient concentrations matrix, profile matrix and contributions matrix, respectively. While $E_{(n\times m)}$ is the noise matrix.

In this work, a dataset with 80 samples was constructed, simulating an 80-days sample campaign. Similar to the references (Brinkman *et al.*, 2006; Shi *et al.*, 2011), the $G_{n\times p}$ values were subjectively varied to reflect differences in the source emission patterns and the influence of metrological conditions. In this way, an 80 × 24 synthetic receptor dataset was obtained (80 samples and 24 chemical species). The synthetic source contributions to the synthetic receptor dataset and their standard deviations are shown in Table 1. In addition to the seven source categories above, 80 daily contributions of noise (unknown sources) were added into the synthetic dataset. The method of noise generation was referred to our prior study (Shi *et al.*, 2011). In order to generate the noise, firstly, 80 profiles were simulated by MATLAB; next, the contributions ($\mu g/m^3$) of noise for 80 days were simulated by MATLAB, with a normal distribution that the average contributions and standard deviations were 30.51 ± 11.53. Thus, the species concentrations for each noise were obtained as:

$$c_{ij} = g_i \times f_{ij}$$

where, c_{ij} is the concentration ($\mu g/m^3$) of j^{th} species in i^{th} noise; g_i is the simulated contributions ($\mu g/m^3$) of i^{th} noise in i^{th} sample; f_{ij} is the simulated fractions (g/g) of j^{th} species in profile of i^{th} noise. The mean value and its standard deviation for 80 contributions of noise were presented in Table 1.

So, for constructing the five synthetic datasets, the five different F (source profiles) matrices were employed. To conveniently compare the results from different models, the same G (source contributions) matrices were applied in the five synthetic datasets. And the synthetic datasets were calculated according to Eq. (3).

RESULT AND DISCUSSION

Source Apportionment by Combined Models, A Case of Synthetic Dataset A

In this section, the Dataset A was analyzed by the multiple combined models firstly, and then the results of combined models (estimated source contributions) will be compared to the synthetic contributions. Secondly, all the five synthetic datasets were studied, and the results will be discussed.

PCA/MLR-CMB Model

(1) PCA/MLR Stage (The First Stage)

The Dataset A was introduced into PCA/MLR-CMB model, firstly. Six factors (with eigenvalue greater than 1) were extracted after varimax rotation, accounting for 78% of the total variance. The relatively lower value of explained variability by interpretable factors might related to high noise introduced in the synthetic database. The loadings of six factors were listed in Table S6.

The first factor (31.84% of the variance) highly related to the species including Al, Si, Ca, Ti. These species are used as markers of crustal sources (Han *et al.*, 2009; Shi *et al.*, 2011; Pant and Harrision, 2012). TC also presented relatively higher weighting in this factor. So, this factor

Table 1. The synthetic contributions of seven sources ($\mu g/m^3$) (mean ± sd).

Sources	Synthetic contributions	Sources	Synthetic contributions
RD	50.15 ± 19.03	vehicle	40.00 ± 15.88
soil	29.94 ± 12.65	secondary sulfate	20.00 ± 7.74
coal combustion	39.87 ± 16.22	secondary nitrate	10.00 ± 4.83
cement	20.01 ± 8.25	noise	30.00 ± 11.34
Total PM mass	239.96 ± 57.47		

might be a mixed source (called mixed source 1), including RD, soil dust, coal combustion and cement sources.

The secondary factor (11.86% of the variance) is explained by TC, which is the marker of vehicle exhaust (Bhave *et al.*, 2007; Ke *et al.*, 2008; Robles *et al.*, 2008). Hence factor 2 can be identified as vehicle exhaust emission source.

The fourth factor (8.32% of the variance) got high loadings of NH_4^+, SO_4^{2-} and NO_3^-. So this factor might be a mixed source (called mixed source 2) which contained secondary sulfate and nitrate sources (Ho *et al.*, 2006; Ke *et al.*, 2008; Pant and Harrision, 2012).

The other three factors got high loadings for the species such as Na, Co, Cu, etc. In this study, these factors might be noises.

Next, the predicted profiles ($\mu g/m^3$) and contributions of extracted sources were estimated and shown in Table S7. And the estimated contributions of extracted factors were compared with the simulated values, as shown in Fig. S1.

(2) CMB Stage (The Secondary Stage)

In this stage, the mixed sources 1 and 2 (in Table S7) were treated as receptors and introduced into CMB model.

For mixed source 1, the values of the receptor concentrations were listed in Table S7; the standard deviations for concentrations were calculated according to our priori works (Shi *et al.*, 2009, 2011). The concentrations and their standard deviations for mixed source 1, as well as the source profiles (including RD, soil dust, coal combustion and cement) were introduced into EPACMB 8.2 model, to estimate the source contributions. After converging, the predicted contributions of four sources were obtained and then shown in Fig. 1. In this stage, the performance indices met the requirements of the CMB model: χ^2 was 0.00, R^2 was 1.00 and the percentage of mass (PM) accounted for was 100.75%.

Similarly, the concentrations and standard deviations for mix source 2, as well as the source profiles (including secondary sulfate and nitrate) were introduced into EPACMB 8.2 model. The calculated contributions of secondary sulfate and nitrate were also described in Fig. 1. The performance indices of results also met the requirements (χ^2: 0.00, R^2: 1.00 and PM: 100.75%).

Unmix-CMB Model

The sources identified by Unmix model were similar to those by PCA/MLR model, as shown in Table S8. The first factor got high level of Al, Si, Ca, etc., so this factor might be the mix source (called mixed source 1), including RD, soil dust, coal combustion and cement. The fifth factor which was characterized by TC, NH_4^+, SO_4^{2-} and NO_3^-, can be identified as a mixed source (called mixed source 2), containing the vehicle exhaust, secondary sulfate and nitrate. The contributions of the two mixed sources were listed in Table S8, as well. Also, the comparison of contributions of extracted factors and true values is described in Fig. S2.

On the CMB stage, the mixed source 1 and 2 were treated as the receptor and introduced into EPACMB8.2 model. The estimated source contributions of Unmix-CMB model were described in Fig. 1. For the two mixed sources, the performance indices of results were also in the range of the requirement.

PMF-CMB Model

Three factors were extracted by PMF model, as presented

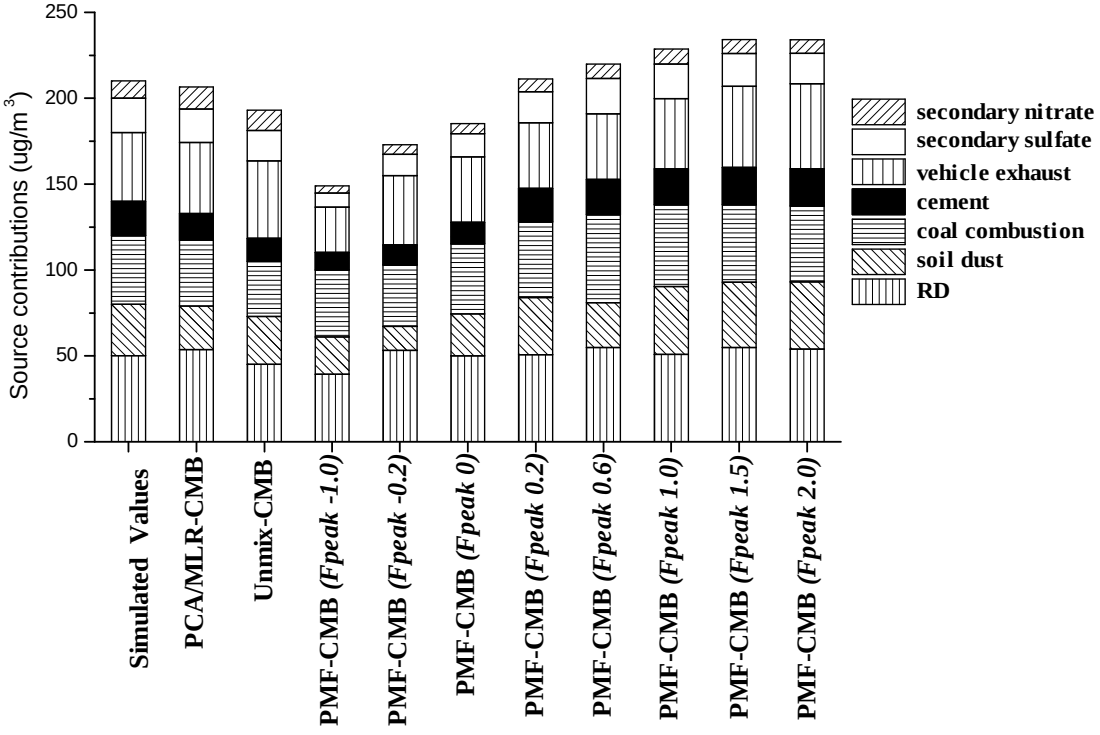

Fig. 1. Source contributions of synthetic and predicted values, for multiply combined models, the case of synthetic Dataset A.

in Table S9. The equation recommended in User's Guide for PMF2 by Paatero (2007) was used to calculate the "uncertainties", which is calculated as follows: std-dev(x_{ij}) is = 5% of x_{ij} plus two units of the least significant digit reported for x_{ij}, where x_{ij} the j^{th} specie concentration in the i^{th} sample. In this case, the Fpeak was set as 0. Factor 1 can be identified as the mix sources including RD, soil dust, coal combustion and cement; and factor 2 can be associated to the mix sources including vehicle, secondary sulfate and nitrate sources. Fig. S3 illuminates the regression plots between the estimated contributions of extracted factors and true values and the final results of PMF-CMB model were presented in Fig. 1. Additionally, the calculated uncertainties for some tracers in the associated mixed sources for final result of Dataset A were listed in Table S10.

Additionally, in order to understand the impact of the Fpeak-rotation on the final results of PMF-CMB model, different Fpeak values were tested. In this work, the values of –1.0, –0.2, 0, 0.2, 0.6, 1, 1.5 and 2 were set, the estimated contributions of extracted factors and true values were compared in Figs. S4–S10, the Q values for corresponding Fpeak were showed in Fig. S11 and the results of PMF-CMB model for these Fpeak were shown in Fig. 1. Additionally, the synthetic dataset had been analyzed by the single CMB model. However, ill estimated source contributions had been obtained by the model, due to the high collinearity.

Comparison of Multiply Combined Models

In Fig. 1, source apportionment results of are presented. In order to compare these different results, relative error (RE) and average absolute error (AAE) (Javitz *et al.*, 1988) were employed. The calculations of RE and AAE are described as follows:

$$RE_j = (E_j - T_j)/T_j \times 100 \quad (4)$$

where, E_j is the predicted contribution (μg/m^3) of j^{th} source; T_j is the synthetic contribution (μg/m^3) of j^{th} source. So, if RE_j was positive value, it indicates that the predicted contribution of j^{th} source was overestimated; oppositely, negative value means the underestimated contribution by the combined model. The RE_j can reflect the discriminate of predicted and synthetic contributions for j^{th} source category.

$$AAE_r = \frac{1}{n} \times \sum_{j=1}^{n} (\left| E_j - T_j \right| / T_j) \times 100 \quad (5)$$

where, n is the number of the source categories (in this study, n = 7). As the estimated source contributions are more close to the true source contributions, the values of *AAE* become smaller. The AAE_T used here to quantify the total difference between predicted and synthetic contributions for all seven source categories.

The *RE* and AAE_T values for multiply combined models are presented in Table 2. It can be found that most of the *REs* were in the range of –50% to 50% and AAE_T were from 9.22% to 35.89%. Javitz *et al.* (1988) suggested that AAE_T less than 50% would represent acceptable precision. In this work, PCA/MLR-CMB, Unmix-CMB, and PMF-CMB (Fpeak = 0, 0.2, 0.6, 1.0, 1.5, 2) got relative low AAE_T values, ranging from 9.22% to 19.38%. It suggests that the multiple combined models could get reasonable precisions. And basing on the AAE_T values, the predicted concentrations obtained by PMF-CMB model (Fpeak = 0.2) got the most accurate predictions, with AAE_T = 9.22%; followed by PCA/MLR-CMB model (AAE_T = 11.86%) and Unmix model (AAE_T = 16.32%).

The Impaction of Fpeak-rotation to Final Results

According to Fig. 1 and Table 2, the PMF-CMB model got different results for multiply Fpeak values. In our prior studies (Shi *et al.*, 2009; Zeng *et al.*, 2010; Shi *et al.*, 2011), it suggests that the final results of combined model are mostly influenced by the profile of the mixed source which extracted by the model on the first stage.

The predicted concentrations of marker species for extracted factors (obtained by PMF-CMB model for different Fpeaks, as well as PCA/MLR and Unmix models) and synthetic values were compared in Fig. 2. The synthetic values of mixed sources total values by adding each source that was included in mixed source. It can be found that the marker species concentrations were relative lower than the

Table 2. RE and AAE_T values for multiply combined models, the case of synthetic Dataset A.

	RE							AAE_T
	RD	soil	coal	cement	vehicle	SS [a]	SN [b]	
PCA/MLR-CMB	6.87	–15.36	–3.78	–22.09	3.33	–2.70	28.91	11.86
Unmix-CMB	–9.97	–7.53	–19.35	–33.01	13.10	–12.04	19.22	16.32
PMF-CMB (Fpeak -1.0)	–21.59	–27.45	–2.21	–48.91	–33.96	–59.18	–57.91	35.89
PMF-CMB (Fpeak -0.2)	6.02	–53.17	–10.45	–41.73	1.10	–38.44	–44.20	27.87
PMF-CMB (Fpeak -0)	–0.43	–18.22	2.46	–37.49	–5.17	–32.26	–39.62	19.38
PMF-CMB (Fpeak 0.2)	0.98	11.19	10.32	–2.61	–4.33	–9.73	–25.41	9.22
PMF-CMB (Fpeak 0.6)	9.05	–11.94	28.11	2.68	–4.68	3.91	–17.85	11.17
PMF-CMB (Fpeak 1.0)	1.33	31.80	19.57	4.18	2.31	0.80	–12.94	10.42
PMF-CMB (Fpeak 1.5)	9.17	27.42	13.03	8.58	18.05	–5.12	–16.07	13.92
PMF-CMB (Fpeak 2.0)	7.44	30.64	11.26	8.48	23.47	–11.89	–19.56	16.11

[a] SS: secondary sulfate.
[b] SN: secondary nitrate.

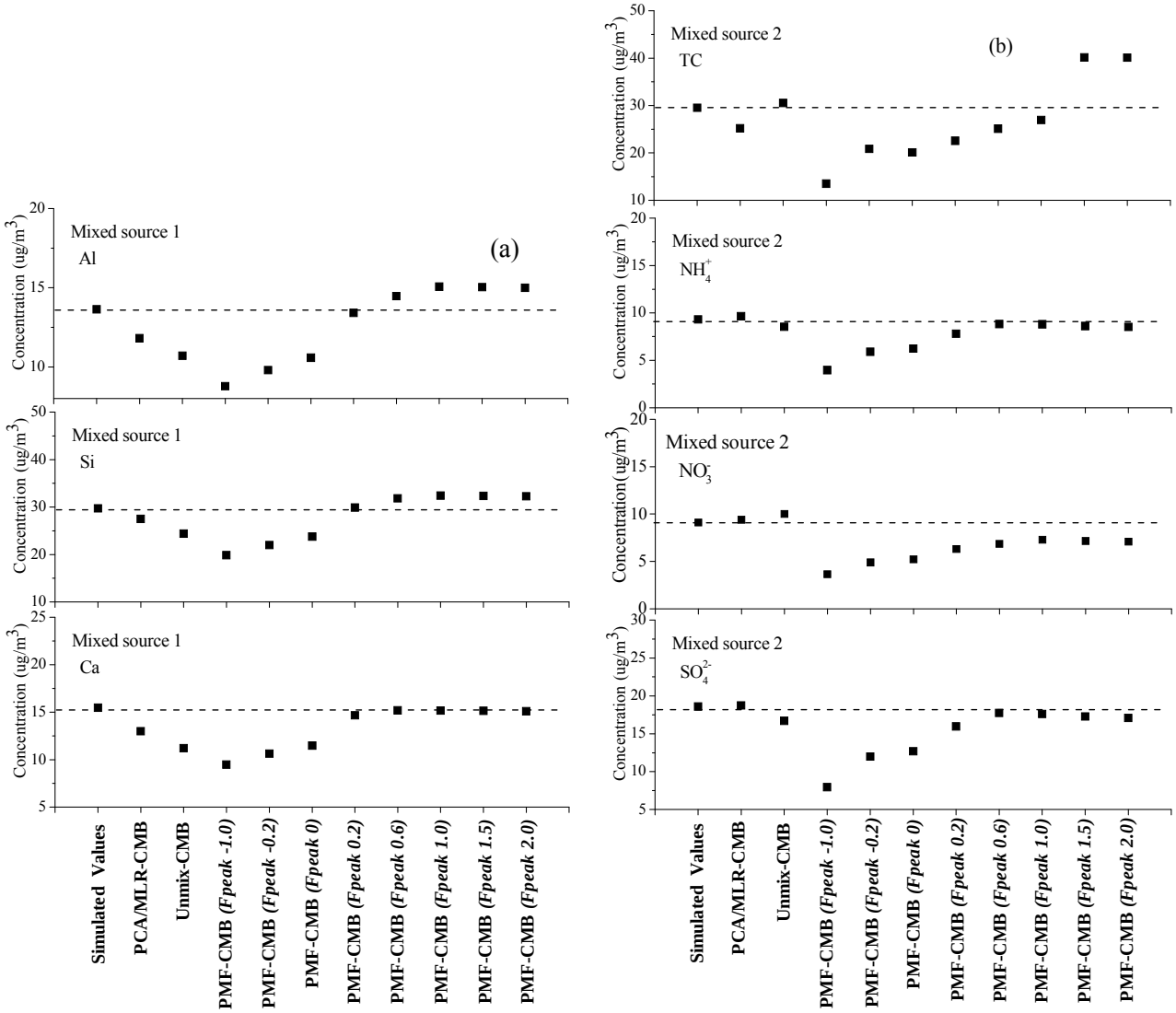

Fig. 2. Species contributions of synthetic and predicted values, for multiply combined models, the case of synthetic Dataset A.

(a) Mixed source 1 (including RD, Soil, Coal and Cement).

The synthetic concentration of species is calculated as follow:

$$C_{ij} = \sum_{k=1}^{4} S_{ik} f_{kj}$$

where s_{ik} is the synthetic contributions of k^{th} source (RD, Soil, Coal and Cement) for i^{th} day; f_{kj} is the fraction of j^{th} species in k^{th} source profile.

(b) Mixed source 2 (including vehicle, secondary sulfate and nitrate).

The synthetic concentration of species is calculated as follow:

$$C_{ij} = \sum_{k=1}^{3} S_{ik} f_{kj}$$

where s_{ik} is the synthetic contributions of k^{th} source (vehicle, secondary sulfate and nitrate) for i^{th} day; f_{kj} is the fraction of j^{th} species in k^{th} source profile.

synthetic ones when the Fpeaks were setting as negative values. Along with the Fpeaks increasing, the predicated concentrations became higher (close or beyond the synthetic values). The changes of the predicted concentrations of

species markers were influenced by the selected different Fpeaks. That is, positive Fpeak values can "sharpen" the profile matrix and negative Fpeak values "smear" the profile matrix (Paatero, 2004; USEPA, 2008).

Combining the results of Fig. 2 and Table 2, we can conclude that: when the predicted concentration of the marker species was below the synthetic one, the source contribution usually was underestimated (RE got negative value in Table 2) on the CMB-stage; on the contrary, the overestimated source contribution (RE got positive value in Table 2) was usually because of the high concentration of marker species. The further discussion of the relationship between the final results of combined models and the concentrations of marker species (obtained on the first stage) will be presented later.

Additionally, the estimated contributions of extracted factors by different models in stage 1 can also influence the final results of the combined models. According to Figs. S1–S10, the slope and Pearson's r values changed largely for the different models. Summary, the results with good slope and Pearson's r values in stage 1 often obtained well final results for the combined models.

Source Apportionment for Five Synthetic Datasets

In this section, the other four synthetic datasets were studied by the multiply combined models, as well. The results for the four datasets were illustrated in Figs. S12–S15. The trends of the results for Dataset B–E were somehow similar to that for Dataset A. As shown in Fig. 1 and Figs. S12–S15, a total of 50 combined model solutions were carried out, for the five synthetic datasets.

Fig. 3 describes the overall results of 50 solutions as well as the synthetic contribution, for each source category. According to the scatter in the retrieved concentrations, most of the predicted contributions were approached the synthetic values, suggesting that these results might be satisfactory. However, there are some points deviated from the synthetic values, indicating the unreasonable results. In this work, most of these unreasonable results were obtained by the PMF-CMB solutions, with negative or large positive (1.5 or 2.0) Fpeak values. The PMF-CMB solutions with negative Fpeak usually got low predicted concentrations of marker species for factors identified as the mixed source on the PMF-stage, resulting in the underestimated contributions on the CMB-stage. On the other hand, very large positive values of FPEAK may not lead to the desired solution as well (Paatero, 2004), due to the excessive overestimation of the marker species' concentrations on the PMF-stage. What's more, the concentrations of all factors are typically slightly overestimated according to Fig. 3, because contributions of noise were added into the synthetic datasets; and the use of positive and negative noise might help to reduce this positive bias in the future.

Addition to Fig. 3, AAE_j for j^{th} source category was employed to help analyze the total accuracy of the predictions. The AAE_j was obtained according to the equation as follows:

$$AAE_j = \frac{1}{50} \times \sum_{i=1}^{50} (\left| E_{ij} - T_{ij} \right| / T_{ij} \times 100 \quad (6)$$

where E_j is the estimated contribution of j^{th} source for i^{th} solution; 50 is the number of the combined model solutions.

The values of AAE_j for 50 solutions are listed in Table 3. All the AAE_j values were less than 50%, indicating that the predictions can be acceptable (Javitz *et al.*, 1988).

What's more, as discussed above, PMF-CMB solutions with negative and large positive (1.5 and 2.0) Fpeaks got relative bad predictions. So, AAE_j values were calculated again, only containing 30 solutions (excluding the PMF-CMB solutions with negative and large positive (1.5 and 2.0) Fpeaks):

$$AAE_j = \frac{1}{30} \times \sum_{i=1}^{30} (\left| E_{ij} - T_{ij} \right| / T_{ij} \times 100 \quad (7)$$

AAE_js for 30 solutions are shown in Table 3 as well. The AAE_js for 30 solutions were relative lower than those for 50 solutions, suggesting that the results of these solutions might be more accuracy.

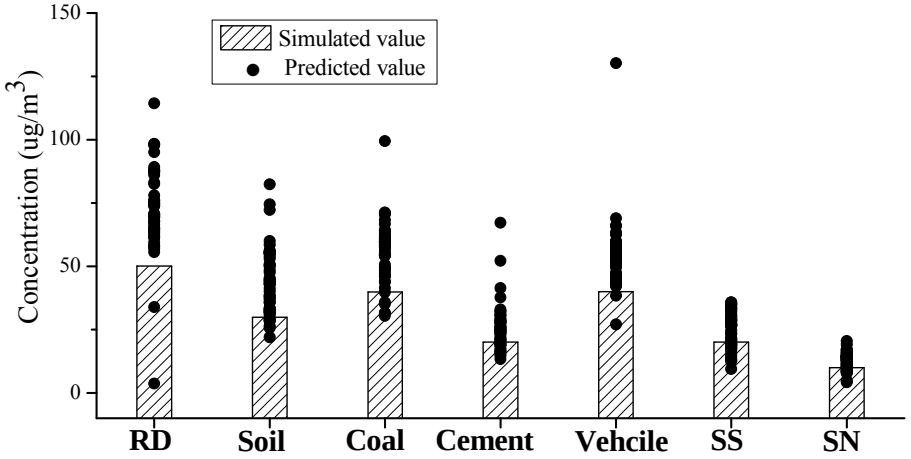

Fig. 3. The results of 50 combined model solutions as well as the synthetic contribution, for each source category. *SS: secondary sulfate; SN: secondary nitrate.

Table 3. AAE_j values for each source category.

Sources	AAE(%)				
	[a]	[b]	Group 3	Group 2	Group1
RD	17.62	15.50	7.76	34.29	11.39
soil	26.57	19.69	27.86	26.24	27.75
coal combustion	17.18	12.49	8.70	19.96	22.73
cement	29.20	24.90	21.36	31.16	35.84
vehicle	13.60	8.83	7.37	13.73	14.69
secondary sulfate	27.10	20.62	21.92	31.20	32.88
secondary nitrate	31.77	27.24	20.25	12.75	40.32

[a] for all 50 solutions.
[b] for 30 solutions, excluding the PMF-CMB solutions with negative and large positive (1.5 and 2.0) Fpeaks.

Besides, an application of combing the source contributions with cluster analysis (CA) was employed to analyze the similarity between synthetic contributions and predictions for 50 solutions. In this work, the application was similar to the process of factor cluster analysis (FCA) (Masiol *et al.*, 2012). In the work of Masiol *et al.* (2012), the source contributions were estimated by the PCA/MLR model, and then the predicted contributions were used as input data for the clustering process using the Ward's hierarchical method and the squared Euclidean distances (Masiol *et al.*, 2012). In our study, the predictions obtained by the 50 solutions (in Fig. 1, Figs. S12–S15) as well as the synthetic values were used as the input data. The dendrogram of CA on the source contributions is described in Fig. S16 and three groups were discriminated. The synthetic values were clustered in group 3, suggesting that the predictions in group 3 were the most close to the synthetic values. While the predictions in group 1 were the most different from the synthetic ones. The AAEs for the solutions in different groups were presented in Table 3 and show the agreement conclusion of dendrogram. Additionally, the input data of CA was also analyzed by PCA to identify the similarity between synthetic contributions and 50 predictions. Two factors were extracted and the factor scores are plotted in Fig. S17. The points which close to the synthetic values indicated that these predictions were relative accurate. The result of Fig. S17 was also consistent to that of CA dendrogram.

According to Figs. S16 and S17, it can be found that the PCA/MLR-CMB solutions can obtain the accurate predictions (in group 3). For PMF-CMB model, the solutions with Fpeaks from 0 to 1.0 usually got relative desired results (mostly in group 3 and 2) while solutions with negative or large positive Fpeaks (1.5 or 2.0) often present bad predictions (mostly in group 1). Some solutions with Fpeaks for 1.5 or 2.0 were clustered in group 2 or 3, however, their performance indices on CMB-stage did not meet the requirements of the CMB mode. And the accuracy of predictions for of Unmix-CMB solutions seems relative instable (Either clustered in group 1 or group 3).

As discussed above, the final predictions of combined models might be influenced by the estimated concentrations of marker species in the mixed source which obtained on the first stage. So in this section, the relationship between markers' concentrations and the predictions were studied. The correlation between the estimated markers' concentrations and the predictions for the solutions were shown in Fig. 4. Good positive correlations were obtained for the predicted source contributions and the concentrations of their markers, suggesting that the estimated markers' concentrations can determine the final results of combined models.

Some duplicate experiments for synthetic dataset tests were carried out by the combined models, and similar conclusions were obtained.

CNOCLUSION

In this work, PCA/MLR-CMB, Unmix-CMB and PMF-CMB models were employed to study the synthetic datasets, and RE and AAE were employed to assess the precision of the results for these multiply combined models. The results suggest that the multiple combined models could get reasonable precisions.

Different Fpeaks were set, for the PMF-CMB solutions, to investigate the impaction of Fpeak on the rotation. The results of PMF-CMB solutions with 0 or positive Fpeaks (0–1.0) were relative accurate, while the PMF-CMB solutions with negative or large positive Fpeaks often got unsatisfactory results, due to the unreasonable extracted profile of mixed source on the first stage.

The predictions of Unmix-CMB solutions were relative instable: some of them were close to the synthetic contributions, while others were deviated from the synthetic values.

Finally, it can be found that the source contributions were good correlated to their markers' concentrations (obtained on the first stage), suggesting that the extracted profiles of mixed sources can determine the final results of combined models. The findings of this study can provide information for application of combined models. Generally, the users can use more than one combined models to enhance the accuracy of the final results. For the PMF-CMB and Unmix-CMB combined models, the selection of the factor number can determine the final results. Thus, a PCA can be employed before the combined models, to help gain the information of the factors. Finally, for the ambient dataset source apportionment, more information beyond the PM concentrations might be needed to help to obtain the acceptable result, such as the field surveys, emission inventory and source investigation.

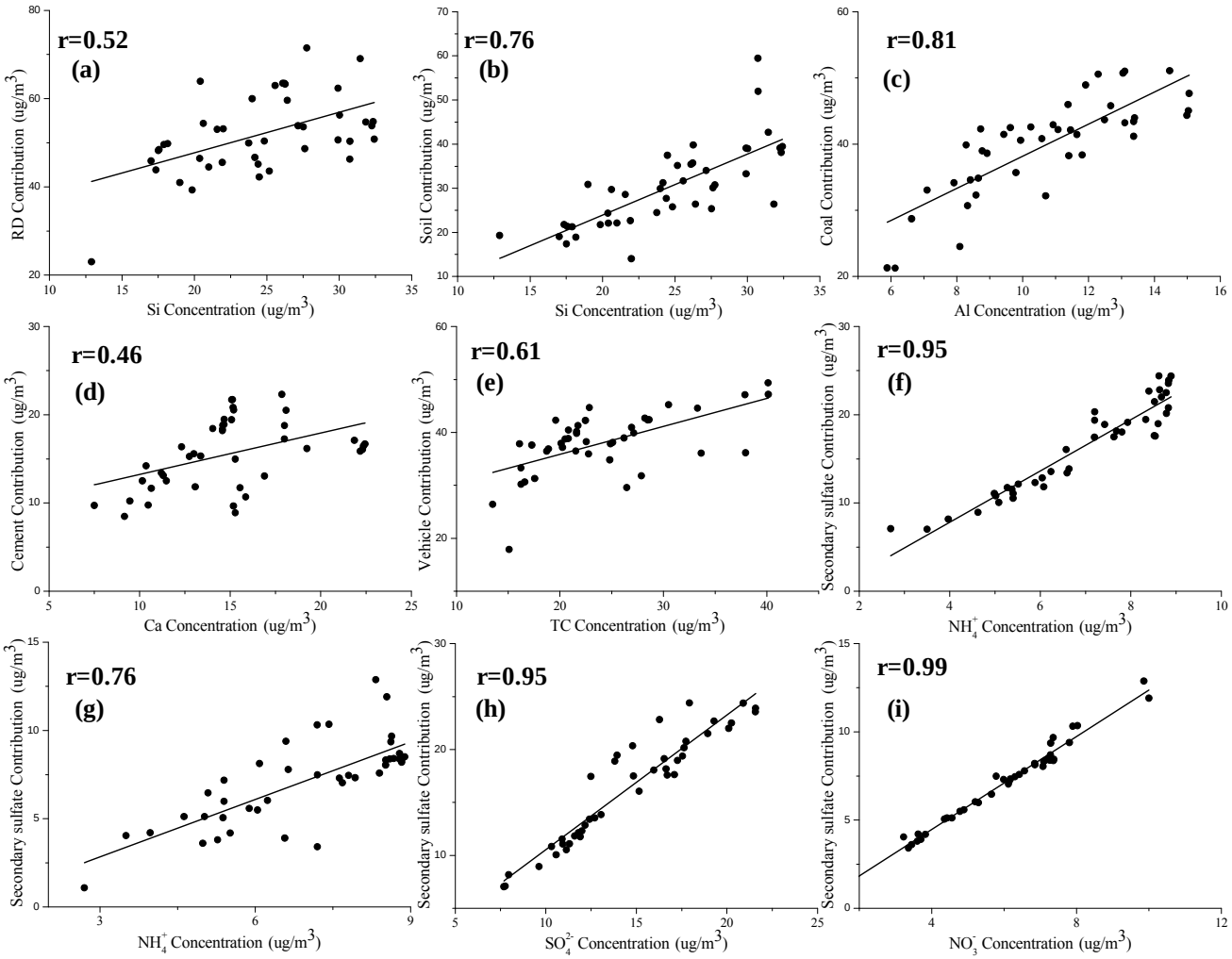

Fig. 4. The correlation between the estimated markers' concentrations and the predictions for the solutions.

ACKNOWLEDGEMENTS

This study is supported by the National Natural Science Foundation of China (21207070, 41205089, 21207069) and the Ph.D. Candidate Research Innovation Fund of Nankai University.

SUPPLEMENTARY MATERIALS

Supplementary data associated with this article can be found in the online version at http://www.aaqr.org.

REFERENCE

Amodio, M., Andriani, E., de Gennaro, G., Di Gilio, A., Ielpo, P., Placentino, C.M. and Tutino, M. (2013). How a Steel Plant Affects Air Quality of a Nearby Urban Area: A Study on Metals and PAH Concentrations. *Aerosol Air Qual. Res.* 13: 497–508.

Begum, B.A., Biswas, S.K., Markwitz, A. and Hopke, P.K. (2010). Identification of Sources of Fine and Coarse Particulate Matter in Dhaka, Bangladesh. *Aerosol Air Qual. Res.* 10: 345–353.

Bhave, P.V., Pouliot, G.A. and Zheng, M. (2007). Diagnostic Model Evaluation for Carbonaceous $PM_{2.5}$ Using Organic Markers Measured in the Southeastern US. *Environ. Sci. Technol*. 5: 1577–1583.

Bi, X.H., Feng, Y.C., Wu, J.H., Wang, Y.Q. and Zhu, T. (2007). Source Apportionment of PM_{10} in Six Cities of Northern China. *Atmos. Environ.* 41: 903–912.

Brinkman, G., Vance, G., Hannigan, M.P. and Milford, J.B. (2006). Use of Synthetic Data to Evaluate Positive Matrix Factorization as a Source Apportionment tool for $PM_{2.5}$ Exposure Data. *Environ. Sci. Technol*. 40: 1892–1901.

Chen, L.W.A., Watson, J.G. and Chow, J.C. (2007). Quantifying $PM_{2.5}$ Source Contributions for the San Joaquin Valley with Multivariate Receptor Models. *Environ. Sci. Technol*. 8: 2818–2826.

Cheng, Y.H., Liu, Z.S. and Ya, J.W. (2012). Comparisons of PM_{10}, $PM_{2.5}$, Particle Number, and CO_2 Levels inside Metro Trains Traveling in Underground Tunnels and on Elevated Tracks. *Aerosol Air Qual. Res.* 12: 879–891.

Gugamsetty, B., Wei, H., Liu, C.N., Awasthi, A., Hsu, S.C., Tsai, C. J., Roam, A.D., Wu, Y.C. and Chen, C.F. (2012). Source Characterization and Apportionment of

PM_{10}, $PM_{2.5}$ and $PM_{0.1}$ by Using Positive Matrix Factorization. *Aerosol Air Qual. Res.* 12: 476–491.

Habre, R., Coull, B. and Koutrakis, P. (2011). Impact of Source Collineaity in Simulated $PM_{2.5}$ Data on the PMF Receptor Model Solution. *Atmos. Environ.* 45: 6938–6946.

Han, Y.M., Cao, J.J., Jin, Z.D. and An, Z.S. (2009). Elemental Composition of Aerosols in Daihai, a Rural Area in the Front Boundary of the Summer Asian Monsoon. *Atmos. Res.* 2: 229–235.

Harrison, R.M., Beddows, D.C. and Dall'Osto, M. (2011). PMF Analysis of Wide-Range Particle Size Spectra Collected on a Major Highway. *Environ. Sci. Technol*. 13: 5522–5528.

Ho, K.F., Cao, J.J., Lee, S.C. and Chan, C.K. (2006). Source Apportionment of $PM_{2.5}$ in Urban Area of Hong Kong. *J. Hazard. Mater.* 1: 73–85.

Hopke, P.K. (2003). Recent Developmens in Receptor Modeling. *J. Chemom*. 17: 255–265.

Javitz, H.S., Watson, J.G. and Robinson, N. (1988). Performance of the Chemical Mass Balance Model with Simulated Local-Scale Aerosols. *Atmos. Environ.* 22: 2309–2322.

Ke, L., Liu, W., Wang, Y.H., Russell, A.G., Edgerton, E.S. and Zheng, M. (2008). Comparison of $PM_{2.5}$ Source Apportionment Using Positive Matrix Factorization and Molecular Marker-Based Chemical Mass Balance. *Sci. Total Environ.* 2–3: 290–302.

Kong, S.F., Ding, X., Bai, Z.P., Han, B., Chen, L., Shi, J.W. and Li, Z.Y. (2010). A Seasonal Study of Polycyclic Aromatic Hydrocarbons in $PM_{2.5}$ and $PM_{2.5-10}$ in Five Typical Cities of Liaoning Province, China. *J. Hazard. Mater.* 1: 70–80.

Kong, S.F., Ji, Y.Q., Lu, B., Chen, L., Han, B., Li, Z.Y. and Baim Z.P. (2011). Characterization of PM_{10} Source Profiles for Fugitive dust in Fushun-a City Famous for Coal. *Atmos. Environ.* 30: 5351–5365.

Lee, E., Chan, C.K. and Paatero, P. (1999). Application of Positive Matrix Factorization in Source Apportionment of Particulate Pollutants in Hong Kong. *Atmos. Environ.* 19: 3201–3212.

Masiol, M., Squizzato, S., Ceccato, D., Rampazzo, G. and Pavoni, B. (2012). A Chemometirc Approach to Determine Local and Regional Sources of PM_{10} and its Geochemical Composition in a Coastal Area. *Atmos. Environ.* 54: 127–133.

Ozkaynak, H. and Thurston, G.D. (1987). Associations between 1980 U.S. Mortality Rates and Alternative Measures of Airborne Particle Concentration. *Risk Anal.* 7: 449–460.

Paatero, P. and Tapper, U. (1994). Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics* 5: 111–126.

Paatero, P. (2004). User's Guide for Positive Matrix Factorization Programs PMF2 and PMF3.

Pant, P.P. and Harrison, R.M. (2012). Critical Review of Receptor Modeling for Particulate Matter: A Case Study of India. *Atmos. Environ.* 49: 1–12.

Robles, L.A., Fu, J.S. and Reed, G.D. (2008). Modeling and Source Apportionment of Diesel Particulate Matter. *Environ. Int.* 34: 1–11.

Russell, A.G. (2009). A focus on Particulate Matter and Health. *Environ. Sci. Technol*. 43: 4620–4625.

Shen, G.F., Wang, W., Yang, Y.F., Zhu, C., Min, Y.J., Xue, M., Ding, J.N., Li, W., Wang, B., Shen, H.Z., Wang, R., Wang, X.L. and Tao, S. (2010). Emission Factors and Particulate Matter Size Distribution of Polycyclic Aromatic Hydrocarbons from Residential Coal Combustions in Rural Northern China. *Atmos. Environ.* 39: 5237–5243.

Shen, G.F., Wang, W., Yang, Y.F., Ding, J.N., Xue, M., Min, Y.J., Chen, Z., Shen, H.Z., Li, W., Wang, B., Wang, R., Wang, X.L., Tao, S. and Russell, A.G. (2011). Emissions of PAHs from Indoor Crop Residue Burning in a Typical Rural Stove: Emission Factors, Size Distributions, and Gas- Particle Partitioning. *Environ. Sci. Technol.* 4: 1206–1212.

Shi, G.L., Li, X., Feng, Y.C., Wang, Y.Q., Wu, J.H., Li, J. and Zhu, T. (2009). Combined Source Apportionment, Using Positive Matrix Factorization-Chemical Mass Balance and Principal Component Analysis/Multiple Linear Regression-Chemical Mass Balance Models. *Atmos. Environ.* 43: 2929–2937.

Shi, G.L., Zeng, F., Li, X., Feng, Y.C., Wang, Y.Q., Liu, G.X. and Zhu, T. (2011). Estimated Contributions and Uncertainties of PCA/MLR-CMB Results: Source Apportionment for Synthetic and Ambient Datasets. *Atmos. Environ.* 45: 2811–2819.

Shi, G.L., Tian, Y.Z., Guo, C.S., Feng, Y.C., Xu, J. and Zhang, Y. (2012). Sediment-Pore Water Partition of PAH Source Contributions to the Yellow River Using Two Receptor models. *J. Soils Sediments* 12: 1154–1163.

Song, Y., Xie, S.D., Zhang, Y.H., Zeng, L., Salmon, L.G. and Zheng, M. (2006). Source Apportionment of $PM_{2.5}$ in Beijing Using Principal Component Analysis/Absolute Principal Component Scores and UNMIX. *Sci. Total Environ.*1: 278–286.

Tie, X.X., Wu, D. and Brasseur, G. (2009). Lung Cancer Mortality and Exposure to Atmospheric Aerosol Particles in Guangzhou, China. *Atmos. Environ.* 14: 2375–2377.

US Environmental Protection Agency. (2004). EPA CMB 8.2 User's manual Office of Air Quality Planning and Standards, Research Triangle Park NC 27711.

US Environmental Protection Agency. (2008). EPA PMF 3.0 User's Manual Office of Air Quality Planning and Standards, Research Triangle Park NC 27711.

Vernile, P., Tutino, M., Bari, G., Amodio, M., Spagnuolo, M., de Gennaro, G., and de Lillo, E. (2013). Particulate Matter Toxicity Evaluation Using Bioindicators and Comet Assay. *Aerosol Air Qual. Res.* 13: 172–178.

Wang, Z.S., Wu, T., Shi, G.L., Fu, X., Tian, Y.Z., Feng, Y.C., Wu, X.F., Wu, G., Bai, Z.P. and Zhang, W.J. (2012). Potential Source Analysis for PM_{10} and $PM_{2.5}$ in Autumn in a Northern City in China. *Aerosol Air Qual. Res.* 12: 39–48.

Watson, J.G. (1984). Overview of Receptor Model Principles. *J. Air Waste Manage. Assoc.* 34: 619–623.

Watson, J.G. and Chow, J.C. (2001). Source

Characterization of Major Emission Sources in the Imperial and Mexicali Valleys along the US/Mexico Border. *Sci. Total Environ.*1–3: 33–47.

Yan, B., Zheng, M., Hu, Y.T., Ding, X., Sullivan, A.P., Weber, R.J., Baek, J., Edgerton, E.S. and Russell, A.G. (2009). Roadside, Urban, and Rural Comparison of Primary and Secondary Organic Molecular Markers in Ambient $PM_{2.5}$. *Environ. Sci. Technol*. 12: 4287–4293.

Zeng, F., Shi, G.L., Li, X., Feng, Y.C., Bi, X.H., Wu, J.H. and Xue, Y.H. (2010). Application of a Combined Model to Study the Source Apportionment of PM_{10} in Taiyun, China. *Aerosol Air Qual. Res.* 10: 177–184.

Zhang, Y., Guo, C.S., Xu, J., Tian, Y.Z., Sh, G.L. and Feng, Y.C. (2012). Potential Source Contributions and Risk Assessment of PAHs in Sediments from Taihu Lake, China: Comparison of Three Receptor Models. *Water Res.* 46: 3065–3073.

Zhang, Y.F., Xu, H., Tian, Y.Z., Shi, G.L., Zeng, F., Wu, J.H., Zhang, X.Y., Li, X., Zhu, T. and Feng, Y.C. (2011). The Study on Vertical Variability of PM_{10} and the Possible Sources on a 220 m Tower, in Tianjin, China. *Atmos. Environ.* 45: 6133–6140.

Zheng, J., Che, W., Zheng, Z., Chen, L. and Zhong, L. (2013). Analysis of Spatial and Temporal Variability of PM_{10} Concentrations Using MODIS Aerosol Optical Thickness in the Pearl River Delta Region, China. *Aerosol Air Qual. Res.* 13: 862–876.

Zheng, M., Salmon, L.G., Schauer, J.J., Zeng, L., Kiang, C.S., Zhang, Y.H. and Cass, G.R. (2005). Seasonal Trends in $PM_{2.5}$ Source Contributions in Beijing, China. *Atmos. Environ.* 22: 3967–3976.

Zheng, M., Cass, G.R., Ke, L., Wang, F., Schauer, J.J., Edgerton, E.S. and Russell, A.G. (2007). Source Apportionment of Daily Fine Particulate Matter at Jefferson Street, Atlanta, GA, during Summer and Winter. *J. Air Waste Manage. Assoc.*2: 228–242.

Received for review, January 26, 2014
Accepted, July 23, 2014