


## Article

# A Novel Robust Method for Solving CMB Receptor Model Based on Enhanced Sampling Monte Carlo Simulation

Wen Hou <sup>1</sup>, Yunlei Yang <sup>2</sup>, Zheng Wang <sup>2</sup>, Muzhou Hou <sup>2,\*</sup> , Qianhong Wu <sup>3</sup> and Xiaoliang Xie <sup>4</sup><sup>1</sup> Lushan Binjiang Experimental School, Changsha 410013, China; hnmzw@csu.edu.cn<sup>2</sup> School of Mathematics and Statistics, Central South University, Changsha 410083, China; yunlei@126.com (Y.Y.); zhengwang@csu.edu.cn (Z.W.)<sup>3</sup> School of Geoscience and Info-Physics, Central South University, Changsha 410083, China; mzhomcm@sina.com<sup>4</sup> School of Mathematics and Statistics, and Mobile E-business Collaborative Innovation Center of Hunan Province, Hunan University of Commerce, Changsha 410205, China; hnucmath207@163.com

\* Correspondence: houmuzhou@sina.com; Tel.: +86-137-8708-8322

Received: 15 February 2019; Accepted: 19 March 2019; Published: 23 March 2019



**Abstract:** The traditional effective variance weighted least squares algorithms for solving CMB (Chemical Mass Balance) models have the following drawbacks: When there is collinearity among the sources or the number of species is less than the number of sources, then some negative value of contribution will appear in the results of the source apportionment or the algorithm does not converge to calculation. In this paper, a novel robust algorithm based on enhanced sampling Monte Carlo simulation and effective variance weighted least squares (ESMC-CMB) is proposed, which overcomes the above weaknesses. In the following practical instances for source apportionment, when nine species and nine sources, with no collinearity among them, are selected, EPA-CMB8.2 (U.S. Environmental Protection Agency-CMB8.2), NKCMB1.0 (NanKai University, China-CMB1.0) and ESMC-CMB can obtain similar results. When the source raise dust is added to the source profiles, or nine sources and eight species are selected, EPA-CMB8.2 and NKCMB1.0 cannot solve the model, but the proposed ESMC-CMB algorithm can achieve satisfactory results that fully verify the robustness and effectiveness of ESMC-CMB.

**Keywords:** CMB receptor model; effective variance weighted least squares algorithm; enhanced sampling Monte Carlo simulation

## 1. Introduction

Atmospheric particulate matter (PM<sub>10</sub> and PM<sub>2.5</sub>, with diameters less than 10  $\mu\text{m}$  and 2.5  $\mu\text{m}$ ) is a mixture of solid or liquid particles suspended in the air, and is an important air pollutant in urban environments [1–3]. Epidemiological studies have shown that PM<sub>2.5</sub>/PM<sub>10</sub> and an increase in respiratory symptoms, lung cancer mortality, and cardiovascular disease are closely related [4–10]. China is one of the countries with the most serious PM<sub>2.5</sub> pollution in the world. In recent years, a total of 28 provinces and cities have reported heavy PM<sub>2.5</sub> pollution phenomena; on average, each province has an annual total of nearly 20 days of heavy pollution.

At present, haze is frequent in China, affecting a wide range and having a long duration, which causes inconvenience to public life, threatens human health, and causes great concern for society and the government. Understanding and clarifying the potential sources and their contributions of PM<sub>2.5</sub> is important [4]. The work of source apportionment of PM<sub>2.5</sub> has become one of the core strategies in the prevention and control of atmospheric pollution.

The CMB (Chemical Mass Balance) air quality model [5,6] is the most important model of atmospheric particulate matter source apportionment technology [7], recommended by the United States' EPA (Environmental Protection Agency), mainly used to study the TSP (Total Suspended Particulate), PM<sub>2.5</sub>, PM<sub>10</sub>, and VOC (Volatile Organic Compounds) as well as other sources of pollutants and their contribution. CMB receptor models are established according to the principle of mass balance, and the chemical concentration of pollutants can be expressed by the sum of the product of the species richness and the source contribution.

The CMB receptor model [8,9] is composed of a set of linear equations, which indicates that the receptor concentration of each chemical element is equal to the linear sum of the product of the element content and the source contribution concentration. The basic principle of CMB model is mass conservation. It is assumed that there are several sources ( $J$ ) that contribute to atmospheric particulates in the receptor, and that: (1) compositions of source emissions are constant over the period of ambient and source sampling; (2) the number of sources or source categories is less than or equal to the number of species; (3) the chemical composition of the particulate matter emitted by the various sources is significantly different; (4) the chemical composition of the particulate matter emitted by the source class is relatively stable; (5) all sources that make an obvious contribution to the receptors have their respective emission characteristics; (6) there is no interaction between the particles emitted by the source class, so the change in the process of transmission can be ignored; and (7) measurement uncertainties are random, uncorrelated, and normally distributed. Then the total substance concentration measured on the receptor is the linear sum of the contribution of each source.

The methods for solving CMB equations mainly include: (1) trace element method [10]; (2) linear programming solution [11]; (3) ordinary weighted least squares method [12]; (4) ridge regression weighted least squares [13]; (5) partial least squares [14]; (6) neural networks [15]; and (7) effective variance weighted least squares (EVWLS) with or without an intercept [16].

At present, the most commonly used algorithm for solving CMB model is the EVWLS method [17], which is derived by minimizing the weighted sums of the squares of the differences between the measured and the calculated values of  $C_i$  and  $F_{ij}$ , and is a practical method for calculating the contribution of the source  $S_j$  and the error  $\sigma_{S_j}$ :

$$\min m^2 = \sum_{i=1}^I \frac{(C_i - \sum_{j=1}^J F_{ij} \times S_j)^2}{V_{eff,i}}, \quad (1)$$

where the effective variance is  $V_{eff,i} = \sigma_{C_i}^2 + \sum_{j=1}^J \sigma_{F_{ij}}^2 \times S_j^2$ ,  $\sigma_{S_j}$  ( $\mu\text{gm}^{-3}$  or g/g) is the uncertainty in source contribution  $S_j$  ( $\mu\text{gm}^{-3}$  or g/g),  $\sigma_{C_i}$  ( $\mu\text{gm}^{-3}$  or g/g) is the uncertainty in the ambient concentrations species  $i$ , and  $\sigma_{F_{ij}}$  is the uncertainty in the fraction of species  $i$  in the source  $j$  profile.

The EVWLS method is actually an improvement over the ordinary weighted least squares method to minimize the sum of squares of the differences between the weighted chemical composition measurements and the calculated values.

However, there are some weaknesses to the above algorithms, such as near collinear sources resulting in incorrect source contributions, and the requirement that the number of chemical species be greater than or equal to the number of sources. At the same time, most of the above algorithms are finally transformed into optimization algorithms, which are mostly NP (Non-deterministic Polynomial) problems. So, in general, we get a locally optimal value or suboptimal value instead of a globally optimal value. So, instability is a fatal drawback to these algorithms, that is to say that different runs of the same input dataset at different times using the same algorithm may produce very different outputs or exhibit high variance with the same diagnostic criteria.

The Monte Carlo method [18], also known as stochastic simulation or statistical experiments, is based on statistical theory, according to the law of large numbers, using computer simulation

technology [19] to solve some practical problem that is difficult to figure out directly with mathematical or other methods. The Monte Carlo method uses computer programs and mathematical models [20] to simulate practical random phenomena, through simulation experiments to get experimental data, and then infers from the analysis to get the law of certain phenomena. Monte Carlo simulation [19] is a method for exploring the solution and sensitivity of a complex system by varying the parameters within the statistical constraints. It is widely used in many fields such as engineering [21], environmental science [22], statistical physics [23], biophysics [24], materials science [25], and financial engineering [26]. Many practical problems are often accompanied by many random factors. If we take these factors into account, the model will become too complex to solve. However, we can utilize the Monte Carlo method to generate a random number to simulate these complicated phenomena, and then find out the operation law. The validity of the Monte Carlo method relies on the sampling process in simulation. However, the simple Monte Carlo algorithm converges too slowly, and it is easy to converge to local extreme points.

In this paper, we explore a novel robust method for solving CMB receptor model based on enhanced sampling Monte Carlo simulation, which overcomes the shortcomings of the above algorithms. In other words, when collinearity exists in the source profiles or the number of source profiles is greater than the number of species, the ESMC-CMB (Enhanced Sampling Monte Carlo CMB) algorithm can come to the correct results for source apportionment. In general, these enhanced sampling methods can be employed to help us quickly find an optimal stable solution when the model is complex, nonlinear, or involves more than just a couple uncertain parameters.

This paper is organized as follows. Section 2 provides a literature review about the CMB model and enhanced sampling Monte Carlo simulation. In Section 3, the proposed enhanced sampling Monte Carlo CMB algorithm (ESMC-CMB) is described. Section 4 presents the related numerical experiments and a comparison with various traditional algorithms. Finally, conclusions are given in Section 5.

## 2. CMB Model and Enhanced Sampling Monte Carlo Simulation

Methods commonly used for the particulate source apportionment include receptor model, source emission inventory, and source dispersion models. The source emission inventory method determines its contribution rate by investigating and accounting for emission factors and activity levels for different source categories. The source dispersion model is a combination of meteorological conditions, emission sources, and chemical processes to assess the distribution and contribution of different source classes in three dimensions [27]. The receptor model is a commonly used model in source apportionment.

In general, due to source  $j$  with constant emission rate  $E_j$ , the source contribution  $S_j$  present at a receptor during a sampling period of length  $T$  is

$$S_j = D_j \cdot E_j, \quad (2)$$

where:

$$D_j = \int_0^T d[\vec{u}(t), \sigma(t), \vec{x}_j] dt. \quad (3)$$

$D_j$  is a dispersion factor depending on atmospheric stability ( $\sigma$ ), wind velocity ( $u$ ) and the location of source  $j$  with respect to the receptor ( $x_j$ ). All parameters in Equation (2) vary with time, so the instantaneous  $D_j$  must be integrated over time period  $T$  [27].

The CMB receptor model consists of a solution of a linear equation that represents the chemical concentration of each receptor as the product of source profile abundance and source contribution. Resource profile abundances (i.e., mass fractions of certain chemicals or other properties emitted from each source) and receptor concentrations (estimated with appropriate uncertainties) are used as input data for CMB. In order to distinguish the contribution of source types, the measured chemical and physical properties must occur in different proportions of source emissions, and the changes of these proportions between source and recipient can be neglected or approximated. The CMB model

calculates the contribution values of each source and the uncertainties of these values. The principle of the CMB receptor model is shown in Figure 1.

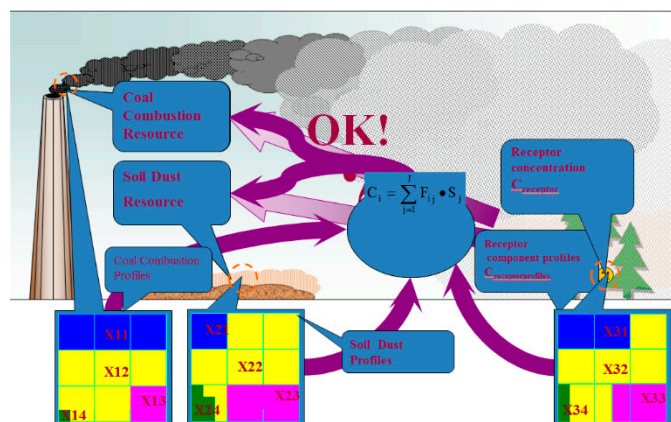


Figure 1. The principle of the Chemical Mass Balance (CMB) receptor model.

The receptor model was used to identify the source of the receptor and determine the quantitative contribution of various sources to the receptor by analyzing the chemical tracers of the source of the environmental samples and the emission sources. If there is no interaction between their emissions to cause mass removal, the total mass measured at the receptor  $C$  is a linear sum of the contributions of the individual sources  $S_j$ :

$$C = \sum_{j=1}^J D_j \cdot E_j = \sum_{j=1}^J S_j. \quad (4)$$

Similarly, the mass concentration of elemental component  $i$ ,  $C_i$ , will be

$$C_i = \sum_{j=1}^J F_{ij} \cdot S_j \quad i = 1, 2, \dots, I, \quad (5)$$

where  $F_{ij}$  is the mass fraction of source contribution  $S_j$  composed of element  $i$  at the receptor. The number of chemical species ( $I$ ) must be greater than or equal to the number of sources ( $J$ ) for a unique solution to these equations.

Equations (4) and (5) are based on material immortality and mass conservation. In Equation (5),  $C_i$  and  $S_j$  are the inputs to the model, and  $F_{ij}$  is the source contribution we need to calculate.

There are several methods to solve the CMB receptor models: (1) the tracer element method [28]; (2) an ordinary weighted least squares solution [28]; (3) a linear programming solution [29], which maximizes the sum of the source contributions; (4) a ridge regression weighted least squares solution with or without an intercept [30] that is one approach for handling the multi-collinearity; (5) a neural networks solution; and (6) an EVWLS solution, which is the most common algorithm.

At present, the most commonly used algorithm to solve the CMB model is the effective variance least squares method, because this method is a practical method to calculate the error  $\sigma_{S_j}$  of source contribution  $S_j$ . The effective variance least squares method is actually an improvement on the ordinary weighted least squares method, which minimizes the sum of squares of the difference between the measured and calculated values of the weighted chemical components:

$$\min m^2 = \sum_{i=1}^I \frac{(C_i - \sum_{j=1}^J F_{ij} \times S_j)^2}{V_{eff,i}}, \quad (6)$$

where the effective variance is  $V_{eff,i} = \sigma_{C_i}^2 + \sum_{j=1}^J \sigma_{F_{ij}}^2 \times S_j^2$ ,  $\sigma_{S_j}$  ( $\mu\text{gm}^{-3}$  or  $\text{g/g}$ ) is the uncertainty in source contribution  $S_j$  ( $\mu\text{gm}^{-3}$  or  $\text{g/g}$ ),  $\sigma_{C_i}$  ( $\mu\text{gm}^{-3}$  or  $\text{g/g}$ ) is the uncertainty (i.e., measurement errors) in the ambient concentrations species  $i$ , and  $\sigma_{F_{ij}}$  is the uncertainty (i.e., measurement errors) in the fraction of species  $i$  in the source  $j$  profile.

The matrix form of the CMB model is as follows:

$$C_{i \times 1} = F_{i \times j} S_{j \times 1}. \quad (7)$$

The steps of EVWLS iterative algorithm for solving the CMB model (Equation (7)) are as follows:

1. Set the initial estimate of the source contributions equal to zero:

$$S_j^{k=0} = 0 \quad j = 1, 2, \dots, J. \quad (8)$$

2. Calculate the diagonal components  $V_{eff,i}$  of the effective variance matrix. All off-diagonal components of this matrix are equal to zero:

$$V_{eff,i}^k = \sigma_{C_i}^2 + \sum_{j=1}^J (S_j^k)^2 \times \sigma_{F_{ij}}^2. \quad (9)$$

3. Calculate the  $K + 1$  value of  $S_j$ :

$$S_j^{k+1} = (F^T (V_e^k)^{-1} F)^{-1} F^T (V_e^k)^{-1} C. \quad (10)$$

4. If the result of Equation (10) is greater than 1%, the previous iteration is executed; if less than 1%, the iteration is terminated.

If  $|S_j^{k+1} - S_j^k| / S_j^{k+1} > 0.01$ , go to step 2. If  $|S_j^{k+1} - S_j^k| / S_j^{k+1} \leq 0.01$ , go to step 5.

5. Calculate the value of  $\sigma_{S_j}$  in the  $K + 1$  step iteration, then

$$\sigma_{S_j} = \left[ (F^T (V_e^{k+1})^{-1} F_{jj})^{-1} \right]^{1/2} \quad j = 1, 2, \dots, J, \quad (11)$$

where  $C = (C_1, \dots, C_I)^T$  is a column vector with  $C_i$  as the  $i$ th component;  $S = (S_1, \dots, S_J)^T$  is a column vector with  $S_j$  as the  $j$ th component;  $F$  is an  $I \times J$  matrix of  $F_{ij}$ , the source composition matrix;  $\sigma_{C_i}$  is one standard deviation uncertainty of the  $C_i$  measurement;  $\sigma_{F_{ij}}$  is one standard deviation uncertainty of the  $F_{ij}$  measurement; and  $V_e$  is diagonal matrix of effective variances.

The above algorithm shows that the input parameters of the model are: the measured values of the concentration spectrum of the chemical components of the receptor  $C_i$  and the standard deviation  $\sigma_{C_i}$  of  $C_i$ , the measured values  $F_{ij}$  of the source chemical composition spectrum and the standard deviation  $\sigma_{F_{ij}}$  of  $F_{ij}$ . The output parameters of the model are: the calculated source contribution values of  $S_j$  and the standard deviation  $\sigma_{S_j}$  of  $S_j$ , the calculated source contribution values of chemical composition  $S_{ij}$ , and the standard deviation  $\sigma_{S_{ij}}$  of  $S_{ij}$ .

In the actual work of source apportionment, there are two commonly used software tools, EPA-CMB8.2 (V8.2, EPA, Washington, USA, 2004) and NKCMB1.0 (V1.0, Nankai University, Tianjin, China, 2005), which are the concrete implementation of above effective variance least squares algorithm for solving the CMB model.

The CMB receptor model is one of the standard methods used by the U.S. Environmental Protection Agency (EPA) to assess air quality. The practical tool software EPA-CMB8.2 based on the CMB model and the effective variance least squares algorithm is recommended by the EPA. NKCMB1.0 is a practical software tool for PM<sub>2.5</sub> source apportionment, developed by the Key Laboratory of Urban Air Particulate Pollution Prevention and Control, Nankai University, Tianjin China, based on the CMB receptor mathematical model and the corresponding effective variance least squares algorithm. NKCMB1.0 is more suitable for source analysis and calculation in China's more complex air quality environment.

As a stochastic method, Monte Carlo modeling can be used to describe and analyze complex problems by computer simulation sampling based on probability theory combined with certain statistical methods. Although the method emerged in the 1940s, it was limited to defense-related nuclear technology because it required sufficient computing resources to analyze the neutron behavior in matter [20]. With the rapid development of high-speed computers, the Monte Carlo simulation method is more and more widely used [19,20].

The basic idea of the Monte Carlo method is to establish an appropriate probability model or stochastic process so that its parameters (such as the probability of events, the mathematical expectation of random variables) are equal to the solution of the problem. Then repeated random sampling test of the model or process are carried out. With the statistical analysis to the results, the final calculation of the parameters, the approximate solution is obtained.

For example, in a Monte Carlo Simulation problem we represent the quantity we want to know as the expected value of a random variable  $Y$ , such as  $\mu = E(Y)$ . Then we generate values  $Y_1, \dots, Y_n$  randomly and independently from the distribution of  $Y$  and get their average:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad (12)$$

as the estimate of  $\mu$ .

However, the convergence speed of the above simple sampling Monte Carlo method is too slow; for a large dimension sampling space, the time to complete the sampling calculation is intolerable.

This paper will explore a new, enhanced sampling method to accelerate the convergence of the algorithm from the following aspects.

Firstly, in the process of solving the receptor CMB model, if the diagnostic indicator  $PM = \sum_{j=1}^J \eta_j = \sum_{j=1}^J S_j/C < \lambda$ , the results did not meet the requirements. So we could sample in the following space  $PM = \sum_{j=1}^J \eta_j = \sum_{j=1}^J S_j/C \geq \lambda$ , for which the dimensions of the sample space will be reduced to some extent, and in the following experiment,  $\lambda = 0.7$  will be selected. In the new sampling space, the Gibbs sampling method will be used.

Gibbs sampling [31–33] or a Gibbs sampler is a MCMC (Markov chain Monte Carlo) algorithm for obtaining a sequence of observations that are approximated from a specified multivariate probability distribution. Like other MCMC algorithms, Gibbs sampling from Markov chain can be regarded as a special case of the Metropolis-Hastings algorithm; its sampling distribution can be deduced from the properties of the Markov chain and probability transition matrix, and it finally converges to joint distribution. The name of the algorithm originated from Josiah Willard Gibbs and was proposed by brothers Stewart and Donald Gemman in 1984 [31–33]. Gibbs sampling is suitable for multivariate distribution, where conditional distribution is easier to sample than edge distribution. At the same time, in order to accelerate the convergence speed of the simulation process, in this paper we adopt the enhanced Gibbs sampling algorithm from [34], called the enhanced sampling algorithm for short.

In order to overcome the shortcomings of the effective variance algorithm for solving the CMB model, in this paper, the EVWLS (effective variance weighted least square) algorithm will be combined with the Monte Carlo simulation algorithm of enhanced sampling to obtain a novel robust ESMC-CMB

algorithm for solving the CMB receptor model. The algorithm is programmed by using MATLAB (V8.5, Mathworks, Natick USA, 2015) and implemented through numerical experiments with a real background. By comparing with the results of EPA-CMB 8.2 and NKCMB 1.0, the accuracy, robustness, and superiority of ESMC-CMB algorithm are fully verified.

### 3. Solving CMB Model Based on Enhanced Sampling Monte Carlo Simulation

For the CMB model with consideration of random error:

$$C_i = \sum_{j=1}^J F_{ij} \cdot S_j + \varepsilon_i, i = 1, 2, \dots, I, \quad (13)$$

where  $C_i$  is the ambient concentration of species  $i$ ,  $S_j$  is the source contribution of source  $j$ ,  $F_{ij}$  is the fraction of species  $i$  in source  $j$ ,  $\varepsilon_i$  is for errors. The number of chemical species ( $I$ ) must be equal to or greater than the number of sources ( $J$ ) for a unique solution to these equations. Equation (13) is solved by an effective variance weighted least squares approach: minimizing  $\chi^2$ , where

$$\chi^2 = \sum_{i=1}^I \left[ \frac{(C_i - \sum_{j=1}^J F_{ij} S_j)^2}{\sigma_{C_i}^2 + \sum_{j=1}^J \alpha_{F_{ij}}^2 S_j^2} \right]. \quad (14)$$

In the CMB model, uncertainties in the source contribution are estimated as

$$\sigma_{S_j} = \left( \sum_{i=1}^I \frac{F_{ij}^2}{\sigma_{C_i}^2 + \sum_{j=1}^J \alpha_{F_{ij}}^2 S_j^2} \right)^{-1/2}, \quad (15)$$

where  $\sigma_{S_j}$  ( $\mu\text{gm}^{-3}$  or g/g) is the uncertainty in source contribution  $S_j$  ( $\mu\text{gm}^{-3}$  or g/g),  $\sigma_{C_i}$  ( $\mu\text{gm}^{-3}$  or g/g) is the uncertainty in the ambient concentrations species  $i$ , and  $\sigma_{F_{ij}}$  is the uncertainty in the fraction of species  $i$  in the source  $j$  profile. Uncertainties in input variables are propagated by inversely weighting the EV (effective variance).

In this paper a new method for solving CMB receptor model based on the enhanced sampling Monte Carlo simulation was proposed as follows:

$$\left\{ \begin{array}{l} \min \chi^2 = \sum_{i=1}^I \left[ \frac{(C_i - \sum_{j=1}^J F_{ij} S_j)^2}{\sigma_{C_i}^2 + \sum_{j=1}^J \alpha_{F_{ij}}^2 S_j^2} \right] \\ \text{st.} \left\{ \begin{array}{l} \text{Generate random inputs : } S_j \\ \text{with Enhanced Gibbs sampler} \\ \sum_{j=1}^J S_j \leq C \\ S_j \geq 0 \\ PM = \sum_{j=1}^J \eta_j = \sum_{j=1}^J S_j / C \geq \lambda \\ i = 1, 2, \dots, I, j = 1, 2, \dots, J \end{array} \right. \\ \sigma_{S_j} = \left( \sum_{i=1}^I \frac{F_{ij}^2}{\sigma_{C_i}^2 + \sum_{j=1}^J \alpha_{F_{ij}}^2 S_j^2} \right)^{-1/2} \end{array} \right. . \quad (16)$$

Then we can get the following ESMC-CMB algorithm:

**Algorithm ESMC-CMB:** Given the initial receptor and source profile data  $C_i$ ,  $\sigma_{C_i}$ ,  $F_{ij}$ ,  $\sigma_{F_{ij}}$ ,  $i = 1, 2, \dots, I, j = 1, 2, \dots, J$ , the number of source and receptor components  $I$ , the number of source  $J$ ,  $obj = 10^{100}$ , the number of simulation times  $N$ ,  $n = 0$ .



Step 1: Generate random variables with the enhanced sampling Monte Carlo method proposed in this paper:  $S_j \geq 0, j = 1, 2, \dots, J$ .

Step 2: If  $\sum_{j=1}^J S_j \geq C$ , go to step 1.

Step 3:  $n = n + 1$ , Calculate  $\chi^2$ , if  $\chi^2 < obj$ , then  $obj = \chi^2$   $objS_j = S_j$ .

Step 4: if  $n < N$  then step 1.

Step 5: Calculate  $\chi^2$ ,  $\eta_j = \frac{objS_j}{C}$ ,  $\sigma_{S_j}$ .

#### 4. Application to a Realistic Case

This realistic case focuses on the dataset from a city in China. The profiles of the receptor and source component are shown in Tables 1 and 2.

Table 1. Receptor component profiles.

Ele.	Conc.	STDE	Ele.	Conc.	STDE
TOT	111.8677	54.19443	Co	0.000505	0.000458
Na	0.381248	0.149582	Ni	0.006908	0.00752
Mg	0.201556	0.094942	Cu	0.055663	0.076044
Al	2.647172	2.03143	Zn	0.237994	0.184731
Si	2.435858	1.56244	Pb	0.111147	0.091934
P	0.061124	0.039434	OC	20.2725	12.6826
K	1.372987	0.862706	EC	3.855547	2.132063
Ca	2.912185	1.292981	Cl	0.26934	0.560002
Ti	0.014792	0.00704	NO <sub>3</sub>	4.703921	5.350789
Cr	0.018382	0.012077	SO <sub>4</sub>	17.27229	7.314421
Mn	0.041736	0.035401	NH <sub>4</sub>	9.960722	5.706486
Fe	4.122549	6.704566			

Note: Ele. = Elements, Conc. = Concentration ( $\mu\text{g}/\text{m}^3$ ), STDE = Standard Deviation.

EPA-CMB8.2 and NKCMB1.0 software can be used to solve the CMB model when the number of sources or source categories is less than or equal to the number of species. So, firstly, we select nine sources (Soil Dust, Construction Dust, Coal Combustion, Cooking Smoke, Biomass Burning, Industrial Processes, NO<sub>3</sub><sup>−</sup>, SO<sub>4</sub><sup>2−</sup>, Vehicular Emissions) and nine components (Al, Si, K, Ca, Fe, OC, EC, NO<sub>3</sub>, SO<sub>4</sub>), and use EPA-CMB8.2 and NKCMB1.0 to calculate source apportionment with the data in Tables 1 and 2; the results are shown in Figures 2 and 3.

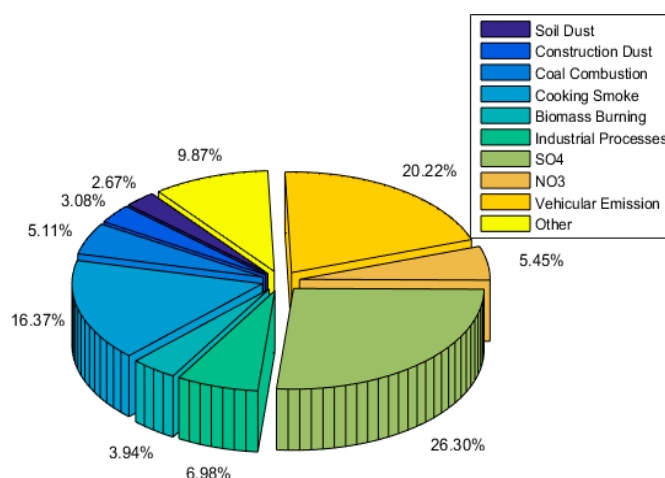


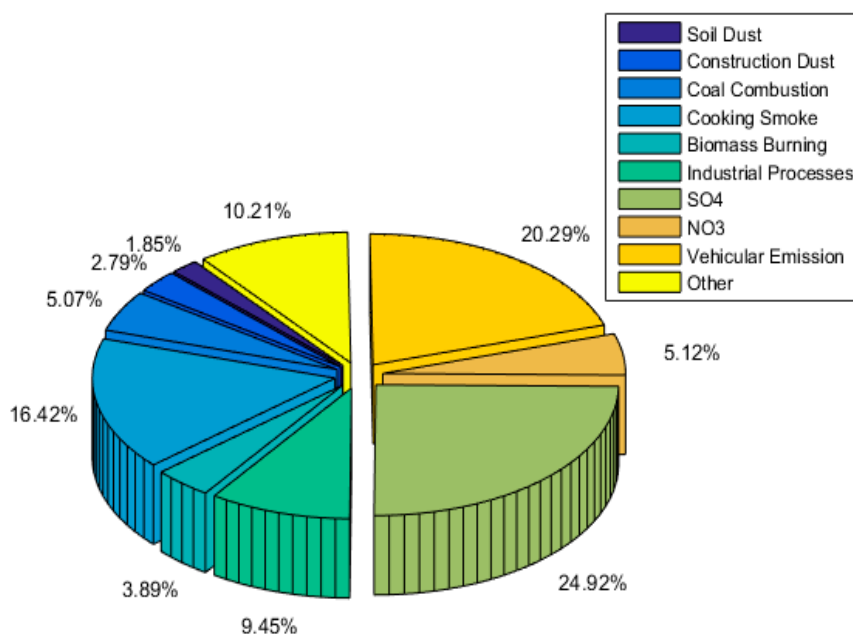
Figure 2. The result using EPA-CMB8.2 with nine sources and nine species.



Table 2. Source component profiles.

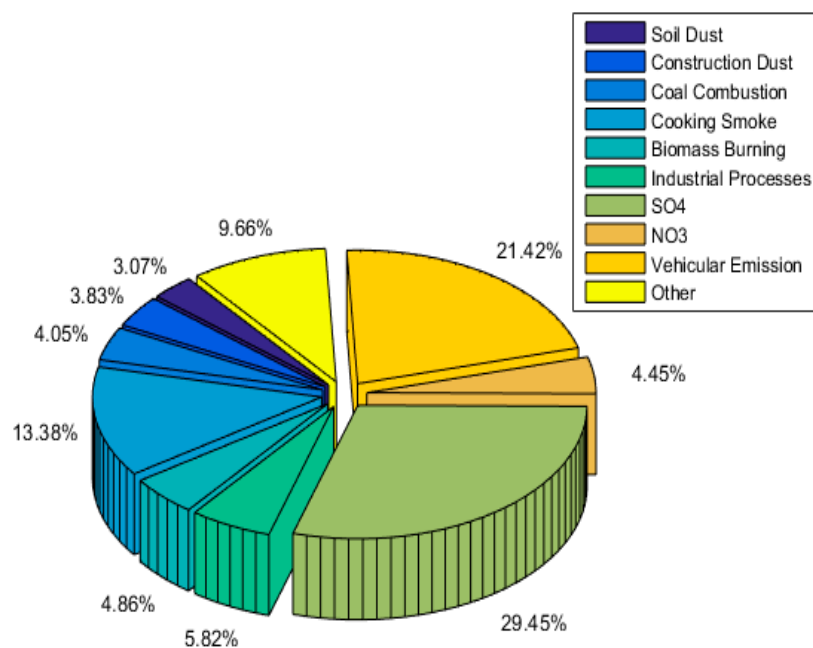
Ele.	Raise Dust		Soil Dust		Construction Dust		Coal Combustion		Cooking Smoke		Biomass Burning		Industrial Processes		NO <sub>3</sub> <sup>−</sup>		SO <sub>4</sub> <sup>2−</sup>		Vehicular Emission	
	Conc.	STDE	Conc.	STDE	Conc.	STDE	Conc.	STDE	Conc.	STDE	Conc.	STDE	Conc.	STDE	Conc.	STDE	Conc.	STDE	Conc.	STDE
Na	0.004722	0.001744	0.007309	0.004183	0.002478	0.000735	0.006365	0.004774	0.008617	0.005742	0.002959	0.002666	0.008600	0.000100	0	0.000001	0	0.000001	0.009363	0.005199
Mg	0.007276	0.001945	0.014675	0.006636	0.008546	0.002448	0.011922	0.008209	0.017405	0.012709	0.004460	0.003831	0.015600	0.000400	0	0.000001	0	0.000001	0.010941	0.006701
Al	0.088236	0.011353	0.118910	0.038482	0.069371	0.031392	0.239006	0.182281	0.012367	0.007308	0.031969	0.026745	0.005300	0.000100	0	0.000001	0	0.000001	0.010639	0.007285
Si	0.137211	0.033887	0.232882	0.076667	0.098363	0.022080	0.081033	0.081762	0.013954	0.014697	0.051409	0.047943	0.013100	0.001300	0	0.000001	0	0.000001	0.012261	0.006434
P	0.00081	0.000228	0.000874	0.000383	0.000264	0.000115	0.000311	0.000262	0.000321	0.000191	0.000126	0.000092	0.000000	0.000001	0	0.000001	0	0.000001	0.002077	0.000802
K	0.013932	0.003267	0.018596	0.006236	0.017332	0.002317	0.008941	0.007336	0.013851	0.015402	0.104925	0.065980	0.033000	0.000700	0	0.000001	0	0.000001	0.012172	0.005053
Ca	0.108035	0.028816	0.125479	0.085679	0.274893	0.043775	0.056683	0.086205	0.012212	0.007046	0.008350	0.007519	0.092000	0.001900	0	0.000001	0	0.000001	0.013024	0.006604
Ti	0.002224	0.000548	0.003509	0.001219	0.002643	0.001152	0.045129	0.030700	0.007512	0.006108	0.001460	0.001747	0.000400	0.000040	0	0.000001	0	0.000001	0.006982	0.003594
Cr	0.000138	0.000050	0.000322	0.000198	0.000084	0.000024	0.000795	0.000887	0.000449	0.000214	0.000872	0.001734	0.000300	0.000030	0	0.000001	0	0.000001	0.001887	0.002440
Mn	0.000501	0.000169	0.000722	0.000303	0.000322	0.000132	0.000193	0.000169	0.000187	0.000162	0.000079	0.000116	0.009800	0.000100	0	0.000001	0	0.000001	0.000901	0.001107
Fe	0.030867	0.009165	0.038558	0.016161	0.013179	0.007315	0.053666	0.032867	0.019319	0.014766	0.010202	0.008648	0.367000	0.000200	0	0.000001	0	0.000001	0.034335	0.022235
Co	0.000011	0.000003	0.000026	0.000012	0.000006	0.000004	0.000013	0.000017	0.000003	0.000003	0.000006	0.000013	0.000300	0.000030	0	0.000001	0	0.000001	0.000028	0.000033
Ni	0.000046	0.000019	0.000135	0.000082	0.000032	0.000004	0.000568	0.000988	0.000211	0.000131	0.000278	0.000550	0.000100	0.000100	0	0.000001	0	0.000001	0.001459	0.001336
Cu	0.000123	0.000048	0.000249	0.000149	0.000070	0.000019	0.000294	0.000233	0.000407	0.000332	0.000178	0.000147	0.000400	0.000100	0	0.000001	0	0.000001	0.001014	0.001473
Zn	0.000579	0.000181	0.000838	0.000476	0.000155	0.000050	0.000649	0.000530	0.001357	0.000918	0.000543	0.000496	0.009800	0.000900	0	0.000001	0	0.000001	0.000952	0.000729
Pb	0.000225	0.000127	0.000121	0.000065	0.000035	0.000006	0.000117	0.000088	0.000115	0.000085	0.000051	0.000036	0.003200	0.000300	0	0.000001	0	0.000001	0.000207	0.000262
OC	0.040941	0.007951	0.024068	0.005945	0.040341	0.007325	0.121996	0.115939	0.642280	0.409048	0.397684	0.092230	0.008200	0.000800	0	0.000001	0	0.000001	0.345982	0.172765
EC	0.006195	0.000620	0.000056	0.000006	0.001326	0.000133	0.013801	0.001380	0.018633	0.001863	0.042355	0.004236	0.004800	0.000480	0	0.000001	0	0.000001	0.147948	0.079996
Cl	0.002261	0.001599	0.004345	0.007539	0.001112	0.001547	0.005714	0.004869	0.008058	0.004448	0.169234	0.084954	0.007200	0.002300	0	0.000001	0	0.000001	0.004652	0.003824
NO <sub>3</sub>	0.006385	0.001638	0.006381	0.004647	0.001362	0.000347	0.006629	0.006358	0.010071	0.008292	0.004203	0.004376	0.000000	0.000001	0.794872	0.079487	0	0.000001	0.009028	0.004308
SO <sub>4</sub>	0.045996	0.014860	0.015446	0.006891	0.024699	0.004715	0.052591	0.047759	0.026123	0.020926	0.021658	0.015497	0.016600	0.002300	0	0.000001	0.727273	0.072727	0.013842	0.010331
NH <sub>4</sub>	0.001191	0.000957	0.001302	0.000726	0.000190	0.000137	0.013185	0.014081	0.010247	0.014778	0.084053	0.051916	0.000000	0.000001	0.205128	0.020513	0.272727	0.027273	0.009194	0.007708

Note: Ele. = Elements, Conc. = Concentration (%), STDE = Standard Deviation.



**Figure 3.** The result using NKCM1.0 with nine sources and nine species.

With the same selection of the source profiles and receptor components and the same dataset, we use our proposed ESMC-CMB algorithm to calculate the source apportionment, and the results are shown in Figure 4. Table 3 shows the numerical comparison of the contribution rates of the above three algorithms.



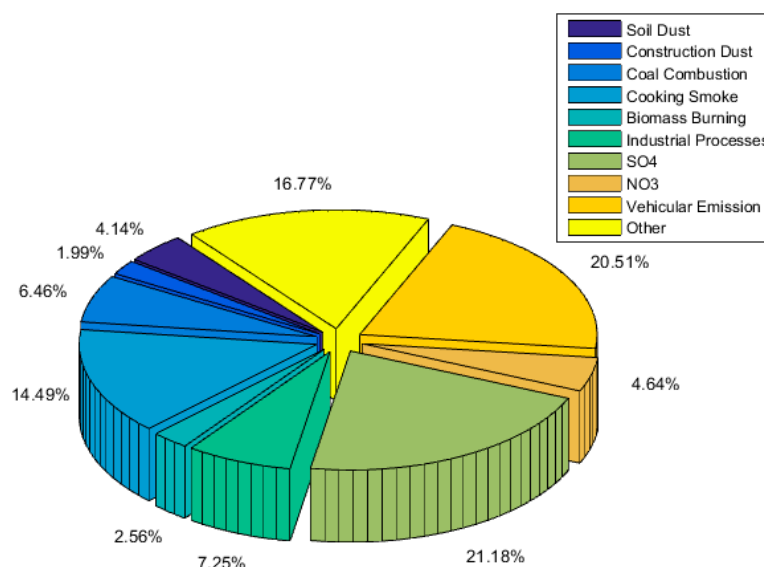
**Figure 4.** The result using proposed ESMC-CMB algorithm (with nine sources and nine species).

**Table 3.** A numerical comparison of EPA-CMB8.2, NKCMB1.0, and ESMC-CMB.

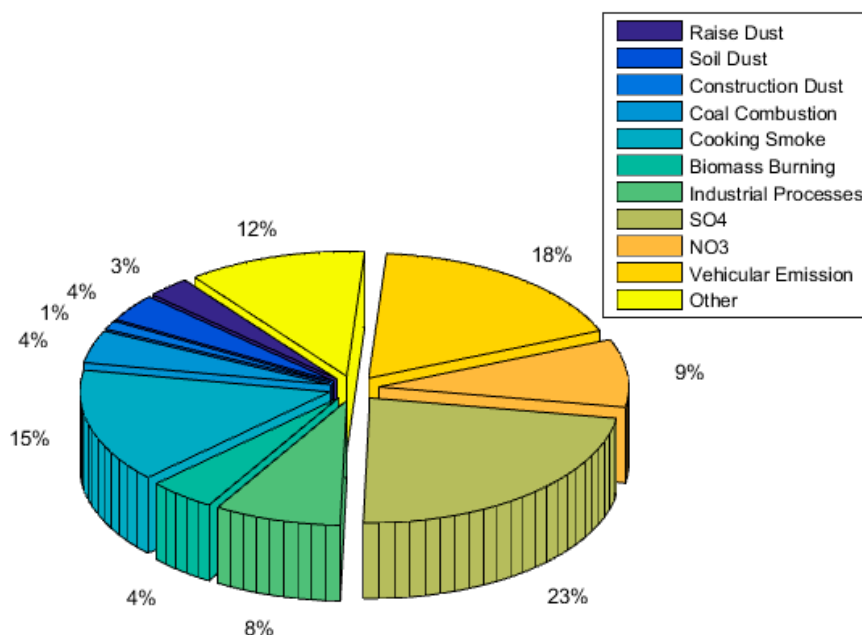
Algorithms Source Contribution	EPA-CMB8.2	NKCMB1.0	ESMC-CMB
Soil Dust	0.026698964	0.018464739	0.030721789
Construction Dust	0.030752527	0.027916899	0.038269989
Coal Combustion	0.051126227	0.050733701	0.04052869
Cooking Smoke	0.163715722	0.164181048	0.13384989
Biomass Burning	0.039397644	0.038893606	0.048647818
Industrial Processes	0.069811171	0.094471832	0.058153522
SO <sub>4</sub> <sup>2−</sup>	0.263039684	0.249173003	0.294499399
NO <sub>3</sub> <sup>−</sup>	0.054531719	0.051178676	0.04448467
Vehicular Emissions	0.202244424	0.202902245	0.214208926
Other	0.098681918	0.102084251	0.096635306

From the results of Figures 2–4 and Table 3, we can see that the results of source apportionment calculated with the above three algorithms are very close, and the correctness of the ESMC-CMB algorithm is verified.

If eight species such as Al, Si, K, Ca, Fe, OC, EC, and NO<sub>3</sub><sup>−</sup> are selected, then the software EPA-CMB8.2 and NKCMB1.0 cannot be used because the number of species is less than the number of sources, but the proposed algorithm ESMC-CBM can calculate the results in Figure 5.

**Figure 5.** The result using ESMC-CMB with nine sources and eight species.

As there is strong collinearity between the sources Raise Dust (RD) and Soil Dust, if RD is added to the source profiles (Soil Dust, Construction Dust, Coal Combustion, Cooking Smoke, Biomass Burning, Industrial Processes, NO<sub>3</sub><sup>−</sup>, SO<sub>4</sub><sup>2−</sup>, Vehicular Emissions) to participate in the calculation using EPA-CMB8.2 and NKCMB1.0, some values of source contribution will be negative, so correct results cannot be obtained. However, using our proposed ESMC-CMB algorithm, we can get the correct value of the source apportionment as shown in Figure 6.



**Figure 6.** The result using ESMC-CMB with 10 sources and nine species including RD (Raise Dust) collinear with Soil Dust.

A comparison of the above results is given in Table 4. As can be seen clearly from Table 4, in the practical instances for source apportionment, when nine species and nine sources, with no collinearity among them, are selected, EPA-CMB8.2, NKCMB1.0, and ESMC-CMB can obtain similar results. However, because there is strong collinearity between source Raise Dust (RD) and Soil Dust, when the source Raise Dust is added to the source profiles, or nine sources and eight species are selected, EPA-CMB8.2 and NKCMB1.0 cannot solve the model, but the proposed ESMC-CMB algorithm can come to a satisfactory results, which fully verify the robustness and effectiveness of ESMC-CMB.

**Table 4.** A comparison of NKCMB1.0 and MC-CMB.

Algorithms \ Conditions	EPA-CMB8.2	NKCMB1.0	ESMC-CMB
Number of sources $\leq$ number of species and existing no collinearity	Having results	Having results	Having results
Number of sources $>$ number of species	No results	No results	Having results
The collinearity exist in sources	No results	No results	Having results

## 5. Conclusions

In this paper, a new robust algorithm for a CMB receptor model based on enhanced sampling Monte Carlo simulation and the effective variance weighted least squares is proposed. Because of the weaknesses of the traditional algorithms and software for CMB receptor source apportionment model such as collinearity and the requirement that the number of chemical species be greater than or equal to the number of sources, in many cases, software such as EPA-CMB8.2 and NKCMB1.0 cannot obtain results for the source apportionment or some values of the source contribution are negative. However, the proposed robust novel ESMC-CMB algorithm can overcome the above weaknesses and achieve satisfactory results. In the realistic source apportionment experiments, firstly, we selected nine sources (Soil Dust, Construction Dust, Coal Combustion, Cooking Smoke, Biomass Burning, Industrial Processes,  $\text{NO}_3^-$ ,  $\text{SO}_4^{2-}$ , Vehicular Emissions) with no collinearity among them and nine species (Al, Si, K, Ca, Fe, OC, EC,  $\text{NO}_3$ ,  $\text{SO}_4$ ), and used the EPA-CMB8.2, NKCMB1.0, and ESMC-CMB algorithms to calculate source contributions, and got similar results, but when we selected eight species and

nine sources or added Raise Dust to the source profiles, because of the collinearity with Soil Dust, EPA-CMB8.2 and NKCMB1.0 could not obtain correct results; however, the proposed ESMC-CMB algorithm can calculate the right results for source apportionment. This has fully demonstrated the robustness and effectiveness of the ESMC-CMB algorithm.

Although the ESMC-CMB algorithm has many advantages, there is often missing data in the actual problem. How to further improve the ESMC-CMB algorithm in the case of missing data is the next area of research to tackle.

Due to the limitations of the CMB model, in the realistic study of air pollution, the results of source analysis from the ESMC-CMB algorithm should be referred to the calculation results of other models, such as PMF (Positive Matrix Factorization) and CMAQ (Community Multiscale Air Quality), to obtain more reasonable results.

**Author Contributions:** Conceptualization, M.H.; Data curation, W.H.; Formal analysis, M.H., W.H., and Y.Y.; Methodology, W.H. and Z.W.; Writing, Q.W. and X.X.

**Funding:** This study was funded by the Natural Science Foundation of China under Grants 61375063, 61271355, 11301549, and 11271378.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Shen, G.; Wang, W.; Yang, Y.; Zhu, C.; Min, Y.; Xue, M.; Ding, J.; Li, W.; Wang, B.; Shen, H.; et al. Emission factors and particulate matter size distribution of polycyclic aromatic hydrocarbons from residential coal combustions in rural Northern China. *Atmos. Environ.* **2010**, *44*, 5237–5243. [\[CrossRef\]](#)
- Kong, S.; Ji, Y.; Lu, B.; Chen, L.; Han, B.; Li, Z.; Bai, Z. Characterization of PM<sub>10</sub> source profiles for fugitive dust in Fushun—a city famous for coal. *Atmos. Environ.* **2011**, *45*, 5351–5365. [\[CrossRef\]](#)
- Zheng, J.; Che, W.; Zheng, Z.; Chen, L.; Zhong, L. Analysis of Spatial and Temporal Variability of PM<sub>10</sub> Concentrations Using MODIS Aerosol Optical Thickness in the Pearl River Delta Region, China. *Aerosol Air Qual. Res.* **2013**, *13*, 862–876. [\[CrossRef\]](#)
- Zheng, M.; Salmon, L.G.; Schauer, J.J.; Zeng, L.; Kiang, C.S.; Zhang, Y.; Cass, G.R. Seasonal trends in PM<sub>2.5</sub> source contributions in Beijing, China. *Atmos. Environ.* **2005**, *39*, 3967–3976. [\[CrossRef\]](#)
- Friedlander, S.K. Chemical element balances and identification of air pollution sources. *Environ. Sci. Technol.* **1973**, *7*, 235–240. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cooper, J.A.; Watson, J.G., Jr. Receptor oriented methods of air particulate source apportionment. *J. Air Pollut. Control Assoc.* **1980**, *30*, 1116–1125. [\[CrossRef\]](#)
- Gordon, G.E. Receptor models. *Environ. Sci. Technol.* **1988**, *22*, 1132–1142. [\[CrossRef\]](#) [\[PubMed\]](#)
- Watson, J.G. Overview of receptor model principles. *J. Air Pollut. Control Assoc.* **1984**, *34*, 619–623. [\[CrossRef\]](#)
- Hidy, G.M.; Venkataraman, C. The chemical mass balance method for estimating atmospheric particle sources in Southern California. *Chem. Eng. Commun.* **1996**, *151*, 187–209. [\[CrossRef\]](#)
- Miller, M.; Friedlander, S.; Hidy, G. A chemical element balance for the Pasadena aerosol. *J. Colloid Interface Sci.* **1972**, *39*, 165–176. [\[CrossRef\]](#)
- Houglund, E. Chemical element balance by linear programming. In Proceedings of the 73rd Annual Meeting of the Air Pollution Control Association, Atlanta, GA, USA, 19–24 June 1983.
- Gartrell, G.; Friedlander, S. Relating particulate pollution to sources: The 1972 California aerosol characterization study. *Atmos. Environ.* **1975**, *9*, 279–299. [\[CrossRef\]](#)
- Watson, J.G.; Robinson, N.F.; Chow, J.C.; Henry, R.C.; Kim, B.; Nguyen, Q.T.; Meyer, E.L.; Pace, T.G. *Receptor Model Technical Series, Vol. III (1989 Revision) CMB7 User's Manual*; US Environmental Protection Agency: Washington, DC, USA, 1990.
- Geladi, P.; Kowalski, B.R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17. [\[CrossRef\]](#)
- Song, X.-H.; Hopke, P.K. Solving the chemical mass balance problem using an artificial neural network. *Environ. Sci. Technol.* **1996**, *30*, 531–535. [\[CrossRef\]](#)
- Watson, J.G.; Cooper, J.A.; Huntzicker, J.J. The effective variance weighting for least squares calculations applied to the mass balance receptor model. *Atmos. Environ.* **1984**, *18*, 1347–1355. [\[CrossRef\]](#)

17. Shi, G.L.; Zeng, F.; Li, X.; Feng, Y.C.; Wang, Y.Q.; Liu, G.X.; Zhu, T. Estimated contributions and uncertainties of PCA/MLR–CMB results: Source apportionment for synthetic and ambient datasets. *Atmos. Environ.* **2011**, *45*, 2811–2819. [[CrossRef](#)]
18. Mahadevan, S. Monte carlo simulation. In *Mechanical Engineering-New York And Basel-Marcel Dekker*; Marcel Dekker Inc.: New York, NY, USA, 1997; pp. 123–146.
19. Brémaud, P. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*; Springer Science & Business Media: Berlin, Germany, 2013; Volume 31.
20. Sobol, I.M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* **2001**, *55*, 271–280. [[CrossRef](#)]
21. Smith, A. *Sequential Monte Carlo Methods in Practice*; Springer Science & Business Media: Berlin, Germany, 2013.
22. Hanna, S.R.; Chang, J.C.; Fernau, M.E. Monte Carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. *Atmos. Environ.* **1998**, *32*, 3619–3628. [[CrossRef](#)]
23. Landau, D.P.; Binder, K. *A Guide to Monte Carlo Simulations in Statistical Physics*; Cambridge University Press: Cambridge, UK, 2014.
24. Friedland, W.; Dingfelder, M.; Kunderát, P.; Jacob, P. Track structures, DNA targets and radiation effects in the biophysical Monte Carlo simulation code PARTRAC. *Mutat. Res./Fund. Mol. Mech. Mutagen.* **2011**, *711*, 28–40. [[CrossRef](#)] [[PubMed](#)]
25. Ohno, K.; Esfarjani, K.; Kawazoe, Y. *Computational Materials Science: From AB Initio to Monte Carlo Methods*; Springer Science & Business Media: Berlin, Germany, 2012; Volume 129.
26. Glasserman, P. *Monte Carlo Methods in Financial Engineering*; Springer Science & Business Media: Berlin, Germany, 2003; Volume 53.
27. Watson, J.G. Chemical Element Balance Receptor Model Methodology for Assessing the Sources of Fine and Total Suspended Particulate Matter in Portland, Oregon. Ph.D. Thesis, Department of Environmental Science, Oregon Graduate Center, Oregon City, OR, USA, 1979.
28. Christensen, W.F.; Gunst, R.F. Measurement error models in chemical mass balance analysis of air quality data. *Atmos. Environ.* **2004**, *38*, 733–744. [[CrossRef](#)]
29. Cheng, M.; Hopke, P.K. *Linear Programming Procedure and Regression Diagnostics for least-Squares Solution Using CMB Receptor Model*, in *Receptor Methods for Source Apportionment—Real World Issues and Applications*; Air Pollution Control Association: Pittsburgh, PA, USA, 1986.
30. Gleser, L.J. Some thoughts on chemical mass balance models. *Chemom. Intell. Lab. Syst.* **1997**, *37*, 15–22. [[CrossRef](#)]
31. Yue, K.; Wu, H.; Liu, W.; Zhu, Y. Representing and processing lineages over uncertain data based on the Bayesian network. *Appl. Soft Comput.* **2015**, *37*, 345–362. [[CrossRef](#)]
32. Kozumi, H.; Kobayashi, G. Gibbs sampling methods for Bayesian quantile regression. *J. Stat. Comput. Simul.* **2011**, *81*, 1565–1578. [[CrossRef](#)]
33. Gilks, W.R.; Wild, P. Adaptive rejection sampling for Gibbs sampling. *Appl. Stat.* **1992**, *41*, 337–348. [[CrossRef](#)]
34. Arroyo, D.; Emery, X.; Peláez, M. An enhanced Gibbs sampler algorithm for non-conditional simulation of Gaussian random vectors. *Comput. Geosci.* **2012**, *46*, 138–148. [[CrossRef](#)]

