The Chemical Mass Balance Method

The relationship between particulate emissions and ambient fallout concentrations measured at a receptor (pollutant sampler) site distant from an emitting source is a complicated one. Many variables, primarily meteorological, make the direct correlation between source emissions and ambient concentrations a poor one. Each of these variables is random in nature, will vary with space and time, and may combine with other variables in a nonlinear manner. Thus, any estimation of source contribution to fallout particles based on emissions and meteorology is approximate at best. However, the chemical mass balance (CMB) receptor-oriented model is a comparatively simple "model" based on physical principles which can be used to determine the average contribution of specific sources categories to particulate fallout. This model is based on the conservation of relative aerosol chemistry from the time a chemical species is emitted from its source to the time it is measured at a receptor. That is, if *p* sources are emitting M_j mass of particles, where *m* is the total mass of the particulate collected on a fallout tray at a receptor site, the model assumes the mass on the fallout tray is a linear combination of the mass contributed from each of the sources.

The mass of a specific chemical species, m_i, is given by the following:

$$m_i = \sum_{j=1}^{p} M_{ij} = \sum_{j=1}^{p} F_{i'j} \; M_j$$

where Mij is the mass of element *I* from source *j* and FN_{ij} is the fraction of chemical species *I* in the mass from source *j* collected at the receptor. It is usually assumed that:

$$F_{i'j} = F_{ij}$$

where F_{ij} is the fraction of chemical *I* emitted by source *j* as measured at the source. The degree of validity in this assumption depends on the chemical and physical properties of the species and its potential for atmospheric modifications such as condensation, volatilization, chemical reactions, sedimentation, etc.

If we accept this equation, however, and divide both sides of Equation 1 by the total mass of the deposit collected at the receptor site, it follows that:

$$\frac{m_i}{m} = \sum_{j=1}^{p} F_{ij} \frac{M_j}{m} \quad (3)$$

or,

$$C_i = \sum_{j=1}^{p} F_{ij} S_j$$

where C_i is the concentration of the chemical component I measured at the receptor and S_j is the source contribution, i.e., the ratio of the mass contributed from source *j* to the total mass collected at the receptor site. In practice, it is this fraction of particulate pollution measured at a receptor due to source *j*, S_j, which is of primary interest in receptor modeling calculations.

If the C_i and the F_{ij} at the receptor for all *p* of the source types suspected of affecting the receptor are known, and $p < n$ (n = number of chemical species), a set of *n* simultaneous equations exists from which the source type contributions S_j may be calculated by least squares methods.

Application of the CMB Modeling Method

In a typical chemical mass balance application, EPA's Version 8.2 CMB model (EPA, 2005) is applied to selected ambient samples. The CMB receptor modeling is performed in a manner consistent with EPA's *Protocol for Applying and Validating the CMB Model* (EPA, 1987).

The CMB procedure begins with a set of linear equations which expresses the ambient concentrations of chemical species measured at an ambient receptor site as the sum of products of source compositions and source contributions. This set of equations is over-determined (more than one possible solution) because the number of chemical species exceeds the number of contributing source types. The source contributions are the unknowns in these equations. However, a unique solution cannot be found for this set of equations because measurement uncertainty precludes determination of exact values for source and receptor data. When these uncertainties are estimated for both source and receptor measurements, additional physical constraints are applied which yield a most probable solution. This solution minimizes the difference between calculated and measured receptor concentrations by using an effective variance weighting scheme. The weighting has a physical significance in that it is derived from the measurement uncertainties of both source and receptor chemical species. (Species with higher relative concentration uncertainties carry less weight in the regression than species with lower relative uncertainties.) Although the CMB solution is identical to some statistical inference methods, it is not dependent on statistical principles. The basic model equations which represent the source receptor relationship, the effective variance weighting, and the error propagation are all based on physical principles.

The CMB provides a source contribution estimate (SCE) and associated standard error uncertainty (STD ERR) for each source category. The model produces these estimates by making an effective variance weighted least squares fit between the chemical composition of the ambient sample and the composition of the sources. It estimates what amounts of each source (the SCEs) will collectively best explain the chemical composition of the ambient sample.

There are five basic data types necessary for CMB modeling:

- Source category names;
- Chemical composition or profile to be associated with each source category;
- Uncertainty in the chemical composition of each source category;
- Chemical composition of the fallout particles sampled at a receptor; and
- Uncertainty in the receptor chemical composition.
- The ability of the CMB model to achieve a proposed set of apportionment goals is determined before the data is input into the computer. In other words, the chemical composition of the source profiles and ambient aerosol are established before the model is applied. At the time of data input, the only options available are the selection of source profiles and the source category names to associate with the profiles.

There are four major steps involved in applying the CMB receptor model to an existing database:

- Determine the appropriateness of the application;
- Form the input data files;
- Select the optimum model solution for each receptor sample; and
- Validate the model results.

The appropriateness of a data set for CMB modeling must be determined before the CMB model is applied. There are no quantitative rules that can be used. However, the EPA suggests using the following criteria as a guide (EPA, 1987):

- Although the model can be applied to a single sample, an adequate number of samples need to be available and included to represent the area or time period for which conclusions are to be drawn.

- Species appropriate to the problem must be included in the database and with precision and accuracy's adequate to achieve source apportionment goals.
- Source categories must not be collinear and their chemical compositions must represent the range of variability expected from a number of individual emitters in the same source type category.
- Source profiles must be representative of the emissions as they would arrive at the receptor.
- The number of source categories in a single application must be less than the number of species included in the regression.

Once it is determined that application of the CMB model is appropriate, it can be applied at varying levels of complexity. The EPA arbitrarily separates these into three levels. Level I uses existing data or data that can easily be obtained from analyses of existing samples. Level II involves additional analyses on existing samples or the acquisition of additional samples. Level III is a comprehensive CMB analysis and includes the acquisition of new data from both ambient and source sampling.

The process of CMB analysis consists of selecting the optimum solution to the effective variance least squares regression using the following seven steps:

- Assessment of the general applicability of the CMB model to the situation under study;
- Configuration of the model with appropriate sources, source profiles, and chemical species concentrations at receptor sites;
- Examination of model statistics and diagnostics;
- Determination of agreement with model assumptions;
- Identification of problems, changing the model configuration, and rerunning;
- Testing of the consistency and stability of model results; and
- Evaluation of the validity of model results.

Although there is a degree of subjectivity in this selection process, much of the subjectivity is removed if the fitting protocols and goodness-of-fit statistical criteria recommended by the EPA are used. The first step is to include all the sources or representatives of all source categories and all defined key species in the initial CMB analysis. Examination of the statistical goodness-of-fit criteria resulting from this initial analysis is used to evaluate the quality of the source contribution estimates. Based on this examination, a different set of sources and species is selected and evaluated. This stepwise procedure continues until, based on the following criteria, an optimum fit is obtained:

- Percent mass explained is close to 100%;
- R-square is close to 1;
- Chi-square is minimized;
- T-statistic is greater than 2;
- Source uncertainty clusters are minimized;
- Calculated-to-measured species ratios are close to 1;
- Ratios of R/U are close to 0; and
- Degrees of freedom are maximized. These criteria are defined and described in Table 1.

Output/Statistic	Abbreviation	EPA Target	Explanation
Std. Error	STD ERR	<< SCE	The standard error of the SCE.
T-statistic	T-STAT	> 2.0	The ratio of the value of the SCE to the uncertainty in the SCE. A T-STAT greater than 2 means that the SCE has a relative uncertainty of less than 50%. T-STAT = SCE/STD ERR
R-square	R-SQUARE	0.80 to 1.00	A measure of the variance of the ambient concentration explained by the calculated concentration. The target range is 0.8 to 1.0, where an r-square of 1.0 is perfect.
Chi-square	CHI-SQUARE	0.0 to 4.0	A term that compares the difference between the calculated and measured ambient concentrations to the uncertainty of the difference. A perfect fit has a chi-square of 0, and a chi-square less than 2 usually indicates a good fit. The target range is 0.0 to 4.0.
Percent Mass Explained	% MASS	100% " 20%	The ratio of the total calculated to measured mass. The target range is 80% to 120%. % MASS = M_c/M_m H 100
Degrees of Freedom	DF	> 5	The difference between the number of fitting species and the number of fitting sources. This value must exceed 1 and should be greater than 5.
Uncertainty/Similarity Clusters	U/S CLUSTERS	None	A list of sources that were not sufficiently resolved by the CMB analysis. No clustering is preferred.
Ratio of Calculated to Measured	RATIO C/M	0.5 to 2.0	The ratio of the calculated to measured concentration of an ambient species. Ideally, this value should be 1.0, but the target range is 0.5 to 2.0. RATIO C/M = C_i/M_i for each species *I*.
Ratio of Residual to Uncertainty	RATIO R/U	2.0 to 2.0	The ratio of the residual (calculated minus measured) to the uncertainty of the residual (square root of the sum of squares of the uncertainties). Target range is -2.0 to 2.0

The model provides three primary outputs: the contribution estimates to ambient concentrations of the sources or source Categories which are included in the fit (SCE), the standard errors of these source contribution estimates (STD ERR), and the species concentrations calculated from the fit (CALC).

The model provides three statistical measures which can be used to evaluate how well the model's calculated species concentrations match the ambient measurements for these species. These statistics are the percent of total mass explained by the fit (% MASS), R-SQUARE, and CHI-SQUARE. It is generally desirable to obtain a good fit of the data based on these three measures while obtaining SCEs with low STD ERR relative to the size of the SCE.

The model provides four diagnostics to help identify data responsible for a poor fit so that improved data might be obtained or included to rectify the situation. These are the uncertainty/similarity clusters (U/S CLUSTERS), the ratio of calculated to measured species concentrations (RATIO C/M), the ratio of the residual (calculated minus measured) to the uncertainty of this difference (RATIO R/U), and the portion of a calculated species concentration that is attributed by the model to each source (SSCONT). The latter diagnostic is not included on the standard CMB printout.

There are four main error categories that can impact model performance: incorrect ambient data, incorrect source profiles, incorrect source list, and profile uncertainty/incorrect collinearity. The existence of these errors can be inferred from the diagnostics and indicators listed above. Possible corrective actions include evaluating ambient and source data, reanalyzing samples, including different sources in the source list, deleting sources from the source list, compositing collinear source profiles, analyzing samples for additional species, etc. After corrective action has been taken, the fit of the measured species data is reevaluated.

When statistically sound and physically reasonable fits have been obtained for the ambient samples of interest, the stability of the CMB model results are assessed. This includes the evaluation of the sensitivity of the model's results to errors in the sources, source profiles, and the ambient data. The final step in the application of the CMB model is validation. In this step, the model results are evaluated for their consistency with available related data (e.g. meteorological, spatial, emissions, and particle size data). Comparisons are made with the results of other receptor and/or dispersion models, if available.

When the summary statistics and diagnostics are generally within target ranges, when there are no significant deviations from model assumptions, when the sensitivity tests uncover no unacceptable instability or consistency problems, and when the results are consistent with available related data, the CMB analysis is considered complete and valid.

Using the fitting parameters in Table 1 and the EPA guidelines, this modeling procedure will generally result in optimized source contributions. The resulting fit is only one of many possible solutions, but it should be the most probable solution. The existence of several different solutions with similar fitting parameters suggests similar probabilities of correctness for each set of source contributions. In such a case, the SCEs of the major sources will likely be quite similar.