EPA United States
Environmental Protection
Agency

EPA Positive Matrix Factorization (PMF) 3.0 Fundamentals & User Guide

EPA 600/R-08/108
July 2008
www.epa.gov

EPA Positive Matrix Factorization (PMF) 3.0 Fundamentals & User Guide

Gary Norris, Ram Vedantham
U.S. Environmental Protection Agency
National Exposure Research Laboratory
Research Triangle Park, NC 27711

Katie Wade, Steve Brown, Jeff Prouty
Sonoma Technology Inc.
Petaluma, CA 94954

Chuck Foley
Lockheed Martin
Systems Engineering Center
Arlington, VA 22201

U.S. Environmental Protection Agency
Office of Research and Development
Washington, DC 20460

Disclaimer

EPA through its Office of Research and Development funded and managed the research and development described here under contract 68-W-04-005 to Lockheed Martin. The User Guide has been subjected to Agency review and is cleared for official distribution by the EPA. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This User Guide is for the EPA PMF 3.0 program and the disclaimer for the software is shown below.

The United States Environmental Protection Agency through its Office of Research and Development funded and collaborated in the research described here under Contract Numbers EP-D-05-004 and 68-W-04-005 to Sonoma Technology, Inc. This software is now being subjected to external peer-review and is for evaluation purposes only. Portions of the code are Copyright©2005-2008 ExoAnalytics Inc. and Copyright©2007-2008 Bytescout.

TABLE OF CONTENTS

TABLE OF FIGURES

1.0 INTRODUCTION

1.1 Model Overview

Positive matrix factorization (PMF) is a multivariate factor analysis tool that decomposes a matrix of speciated sample data into two matrices—factor contributions and factor profiles—which then need to be interpreted by an analyst as to what source types are represented using measured source profile information, wind direction analysis, and emission inventories. The method is reviewed briefly here and described in greater detail elsewhere (Paatero and Tapper, 1994; Paatero, 1997).

A speciated data set can be viewed as a data matrix X of *i* by *j* dimensions, in which *i* number of samples and *j* chemical species were measured. The goal of multivariate receptor modeling, for example with PMF, is to identify a number of factors *p*, the species profile *f* of each source, and the amount of mass *g* contributed by each factor to each individual sample (see Equation 1-1):

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij} \tag{1-1}$$

where e_{ij} is the residual for each sample/species.

Results are constrained so that no sample can have a negative source contribution. PMF allows each data point to be individually weighed. This feature allows the analyst to adjust the influence of each data point, depending on the confidence in the measurement. For example, data below detection can be retained for use in the model, with the associated uncertainty adjusted so these data points have less influence on the solution than measurements above the detection limit. The PMF solution minimizes the object function *Q* (Equation 1-2), based upon these uncertainties (*u*).

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{u_{ij}} \right]^2 \tag{1-2}$$

Variability in the PMF solution can be estimated using a bootstrapping technique, which is a re-sampling method in which “new” data sets are generated that are consistent with the original data. Each data set is decomposed into profile and contribution matrices, and the resulting profile and contribution matrices are compared with the base run (Eberly, 2005). Instead of inspecting point estimates, this method allows the analyst to review the distribution for each species to evaluate the stability of the solution.

1.2 Multilinear Engine (ME)

Two common programs solve the PMF problem as described above. PMF2 (Paatero, 2000) was originally used. In the late 1990s, a more flexible program was developed (Paatero, 1999), known as the multilinear engine (ME). This program is currently in its second version and is referred to as ME-2. ME-2 is the underlying program used to solve the PMF problem in the program EPA PMF, the user interface that feeds the data and user specifications to ME-2. ME-2 then performs the iterations via the conjugate gradient algorithm until convergence to a minimum Q value. The minimum Q may be global or local; a user can attempt to determine which by using different starting points for the iterative process and comparing the minimum Q value reached. Output from ME-2 is then fed back through EPA PMF and formatted appropriately for users to interpret.

The differences in ME-2 and PMF2 have been examined in several studies by the application of each model to the same data set and a comparison made of the results. Overall, the studies showed similar results for the major components, but a greater uncertainty in the PMF2 results (Ramadan et al., 2003) and better source separation using ME-2 (Kim et al., 2007).

EPA PMF v1.1 uses an older version of the multilinear engine. There are some differences in how the program performs; however, results obtained from either program should be similar.

1.3 Comparison to Other Methods

Other source apportionment models include Unmix and chemical mass balance (CMB). Although both methods have aims similar to that of PMF, they have different mechanisms. Unmix uses geometrical objects called the “edges” to identify factors. An edge is identified in the hyperspace of species concentrations where the factor contribution from at least one factor is either zero or present in negligible amounts along the edge. Unmix does not allow individual weighting of data points as does PMF. Although major factors resolved by PMF and Unmix are generally the same, Unmix does not always resolve as many factors as PMF (Pekney et al., 2006; Poirot et al., 2001).

With CMB, the user must provide source profiles which the model uses to apportion mass. PMF and CMB have been compared in several studies; for example, Rizzo and Scheff (2007) compared the magnitude of source contributions resolved by each model and examined correlations between PMF- and CMB-resolved contributions. They found the major factors correlated well and were similar in magnitude; additionally, the PMF-resolved source profiles were generally similar to measured source profiles. In supplementary work, Rizzo and Scheff (2007) used information from CMB PM source profiles to influence PMF results and used CMB results to help control rotations in PMF. Jaeckels et al. (2007) used organic molecular markers with elemental carbon (EC) and organic carbon (OC) in both CMB and PMF. Good correlations were found for most factors, with some biases present in a few of the factors. They also found an additional PMF factor that did not correspond to any CMB factors.

The models discussed above are complementary and, whenever possible, should be used along with PMF to make source apportionment results more robust.

2.0 USES OF PMF

PMF has been applied to a wide range of data, including 24-hr speciated $PM_{2.5}$, size-resolved aerosol, deposition, air toxics, and volatile organic compound (VOC) data. A more complete discussion of uses of PMF is available in the "Multivariate Receptor Modeling Workbook" (2007). PMF requires a data set consisting of a suite of parameters measured across multiple samples. For example, PMF is often used on speciated $PM_{2.5}$ data sets with over 100 samples. An uncertainty data set, that assigns an uncertainty value to each species and sample, is also needed.

3.0 INSTALLING EPA PMF V3.0

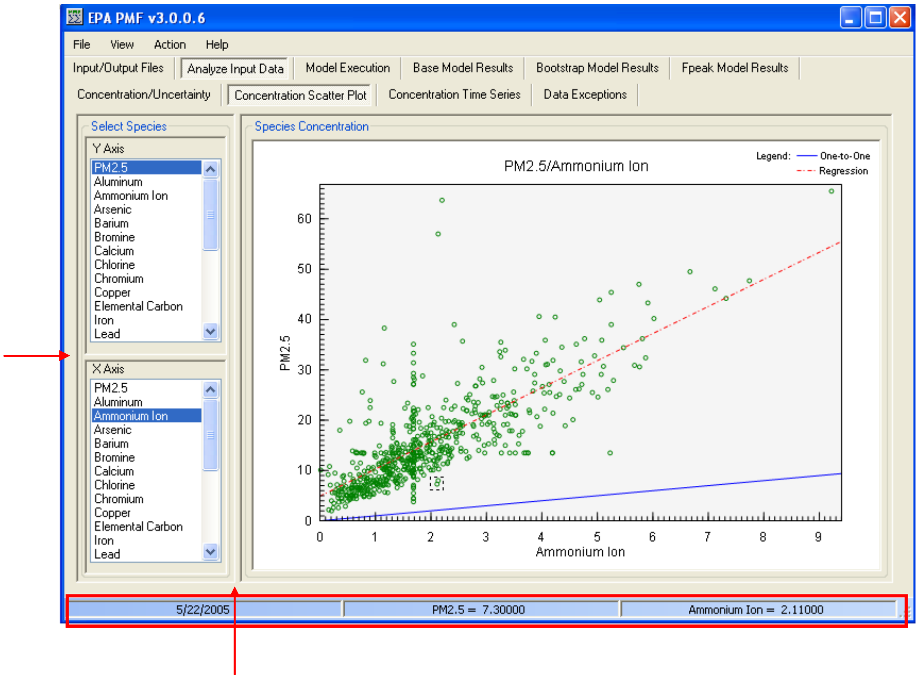
EPA PMF v3.0 can be run on a personal computer using the Windows 95 operating system or higher. The program can be obtained from EPA by e-mailing NERL_RM_Support@epa.gov. It is installed by running **EPA PMF v3.0 Setup.exe**. The installation program offers options for installation; for example, which local directory to use (the default directory is C:\Program Files\EPA PMF 3.0). Follow the installation directions on the screen. Installation problems should be reported to NERL_RM_Support@epa.gov.

A user running Windows Vista will have to disable the user account control (UAC) before running EPA PMF v3.0. EPA PMF v.3.0 can be started by double clicking **EPA PMF v3.0.exe**.

4.0 GLOBAL FEATURES

The following features are available throughout EPA PMF v3.0 where appropriate:

- **Data sorting capabilities.** Columns in tables can be sorted by left-clicking the mouse button on the heading. Clicking once will sort ascending and clicking twice will sort descending. If a column has been sorted, an arrow will appear in the header indicating the direction in which it is sorted.
- **Saving graphics.** All graphical output can be saved in a variety of formats by right-clicking on an image. Available formats are .GIF, .BMP, .PNG, and .TIFF. In the same menu, the user can choose to copy or print a graphic. When "copy" is selected, the graphic is copied to the clipboard. When "print" is selected, the graphic will automatically be sent to the local machine's default printer. When saving a graphic, a dialog box appears where the user can change the file path and file name of the output file.
- **Undocking graphs.** Any graph can be opened in a new window by right-clicking on the graph and selecting Floating Window. The user can open as many windows as required. The graphs in the floating windows do not update when model parameters and output are changed.
- **Status bar.** Most screens have a status bar across the bottom of the window that provides additional information to the user. This information changes based on the tab selected. More information is available in the discussion below of each tab. An example of the status bar on the Concentration Scatter Plot screen is shown in Figure 4-1.



5/22/2005 | PM2.5 = 7.30000 | Ammonium Ion = 2.11000

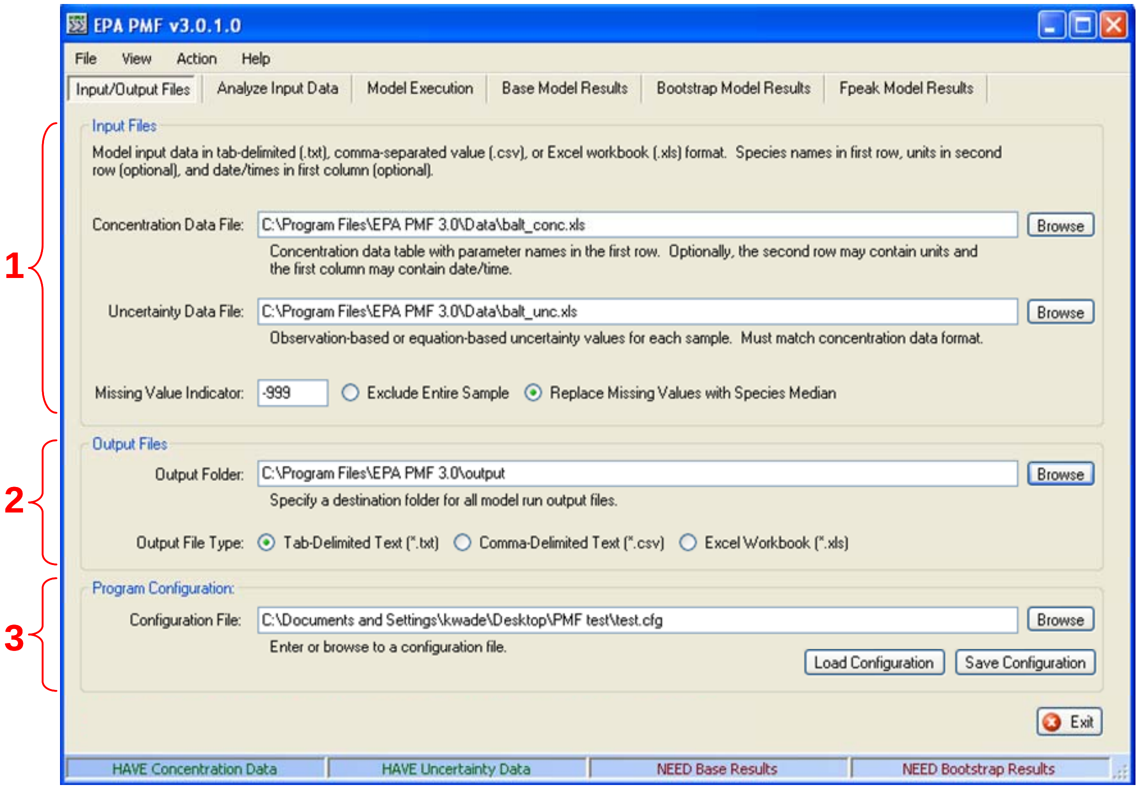
Figure 4-1.—Example of resizable sections and status bar. Red arrows indicate grey bars that enable the user to adjust height and width. The red box indicates the status bar.

- **Resizing sections within tabs.** Many tabs have multiple sections separated by a grey line (see Figure 4-1). These sections can be resized by clicking on the grey line and dragging it to the desired location.
- **Indication of selected data points.** When the user moves the cursor over a point on scatter plots and time series graphs, the point is outlined with a dashed-line square, indicating the point to which the information in the status bar refers.
- **Using arrow keys on lists/tables.** After selecting (by clicking on or tabbing to) a list or table, the keyboard arrow keys can be used to change the selected row.

5.0 GETTING STARTED

Each time the EPA PMF v3.0 program is started, a splash screen with information about the development of the software and various copyrights is displayed. The user must click the **OK** button or press the spacebar or Enter key to continue.

The first EPA PMF 3.0 screen and tab is the **Input/Output Files** screen as shown in Figure 5-1. On this screen, the user provides file location information and selects various specifications that will be used throughout the program. This screen has three sections: **Input Files** (Figure 5-1, 1), **Output Files** (Figure 5-1, 2), and **Program Configuration** (Figure 5-1, 3), each of which is described in detail below. The status bar on the **Input/Output Files** screen indicates which section of the program has been completed. Before input into the parts of this screen, the status bar displays **Need Concentration Data**, **Need Uncertainty Data**, **Need Base Results**, and **Need Bootstrap** results in red. When a task is completed, **Need** is replaced with **Have** and the color changes to green. In the Figure 5-1 example, concentration and uncertainty files have been provided to the program, so the first two items on the status bar are green, but base runs and bootstrap runs have not been completed, so the last two items are red.



Input Files

Model input data in tab-delimited (.txt), comma-separated value (.csv), or Excel workbook (.xls) format. Species names in first row, units in second row (optional), and date/times in first column (optional).

Concentration Data File: C:\Program Files\EPA PMF 3.0\Data\balt_conc.xls

Concentration data table with parameter names in the first row. Optionally, the second row may contain units and the first column may contain date/time.

Uncertainty Data File: C:\Program Files\EPA PMF 3.0\Data\balt_unc.xls

Observation-based or equation-based uncertainty values for each sample. Must match concentration data format.

Missing Value Indicator: -999 ○ Exclude Entire Sample ● Replace Missing Values with Species Median

Output Files

Output Folder: C:\Program Files\EPA PMF 3.0\output

Specify a destination folder for all model run output files.

Output File Type: ● Tab-Delimited Text (*.txt) ○ Comma-Delimited Text (*.csv) ○ Excel Workbook (*.xls)

Program Configuration:

Configuration File: C:\Documents and Settings\kwade\Desktop\PMF test\test.cfg

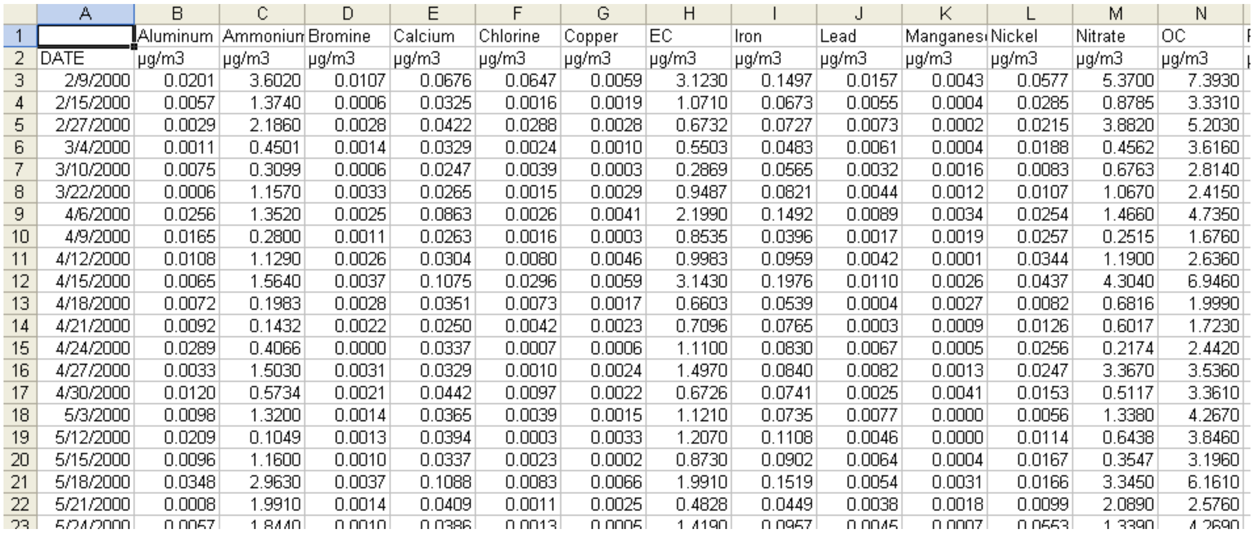
Enter or browse to a configuration file.

HAVE Concentration Data | HAVE Uncertainty Data | NEED Base Results | NEED Bootstrap Results

Figure 5-1.— Example **Input/Output Files** screen.

5.1 Input files

Two input files are required by PMF (Figure 5-1, 1): one containing concentration values and one containing either uncertainty values or parameters for calculating uncertainty. EPA PMF will accept tab-delimited (.txt), comma-separated value (.csv), or MS Excel (.xls) files. Each file is loaded either by typing the path into the “data file” input boxes or browsing to the appropriate file. If the file includes more than one worksheet or named range, the user will be asked to select the one they want to use. The concentration file should contain parameters as columns and dates/samples as rows, with headers for each (Figure 5-2). All standard date and time conventions are accepted. Units can be included as a second heading row, but are not required. If units are supplied by the user, they will be used by the GUI for axes labels only and do not impact the model. The Baltimore example data set, included with EPA PMF v3.0 (balt_conc.xls and balt_unc.xls) is an example of input files containing units. Blank cells are not accepted; the user will be prompted to examine the data and try again. If values greater than 9000 or less than -900 are found in the data set, the program will give a warning message but will continue. If these values are not real or are missing value indicators, the user should fix the data file outside the GUI and reload the data sets.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Aluminum	Ammonium	Bromine	Calcium	Chlorine	Copper	EC	Iron	Lead	Manganese	Nickel	Nitrate	OC
2	DATE	μg/m3	μg/m3	μg/m3	μg/m3	μg/m3	μg/m3	μg/m3	μg/m3	μg/m3	μg/m3	μg/m3	μg/m3	μg/m3
3	2/9/2000	0.0201	3.6020	0.0107	0.0676	0.0647	0.0059	3.1230	0.1497	0.0157	0.0043	0.0577	5.3700	7.3930
4	2/15/2000	0.0057	1.3740	0.0006	0.0325	0.0016	0.0019	1.0710	0.0673	0.0055	0.0004	0.0285	0.8785	3.3310
5	2/27/2000	0.0029	2.1860	0.0028	0.0422	0.0288	0.0028	0.6732	0.0727	0.0073	0.0002	0.0215	3.8820	5.2030
6	3/4/2000	0.0011	0.4501	0.0014	0.0329	0.0024	0.0010	0.5503	0.0483	0.0061	0.0004	0.0188	0.4562	3.6160
7	3/10/2000	0.0075	0.3099	0.0006	0.0247	0.0039	0.0003	0.2869	0.0565	0.0032	0.0016	0.0083	0.6763	2.8140
8	3/22/2000	0.0006	1.1570	0.0033	0.0265	0.0015	0.0029	0.9487	0.0821	0.0044	0.0012	0.0107	1.0670	2.4150
9	4/6/2000	0.0256	1.3520	0.0025	0.0863	0.0026	0.0041	2.1990	0.1492	0.0089	0.0034	0.0254	1.4660	4.7350
10	4/9/2000	0.0165	0.2800	0.0011	0.0263	0.0016	0.0003	0.8535	0.0396	0.0017	0.0019	0.0257	0.2515	1.6760
11	4/12/2000	0.0108	1.1290	0.0026	0.0304	0.0080	0.0046	0.9983	0.0959	0.0042	0.0001	0.0344	1.1900	2.6360
12	4/15/2000	0.0065	1.5640	0.0037	0.1075	0.0296	0.0059	3.1430	0.1976	0.0110	0.0026	0.0437	4.3040	6.9460
13	4/18/2000	0.0072	0.1983	0.0028	0.0351	0.0073	0.0017	0.6603	0.0539	0.0004	0.0027	0.0082	0.6816	1.9990
14	4/21/2000	0.0092	0.1432	0.0022	0.0250	0.0042	0.0023	0.7096	0.0765	0.0003	0.0009	0.0126	0.6017	1.7230
15	4/24/2000	0.0289	0.4066	0.0000	0.0337	0.0007	0.0006	1.1100	0.0830	0.0067	0.0005	0.0256	0.2174	2.4420
16	4/27/2000	0.0033	1.5030	0.0031	0.0329	0.0010	0.0024	1.4970	0.0840	0.0082	0.0013	0.0247	3.3670	3.5360
17	4/30/2000	0.0120	0.5734	0.0021	0.0442	0.0097	0.0022	0.6726	0.0741	0.0025	0.0041	0.0153	0.5117	3.3610
18	5/3/2000	0.0098	1.3200	0.0014	0.0365	0.0039	0.0015	1.1210	0.0735	0.0077	0.0000	0.0056	1.3380	4.2670
19	5/12/2000	0.0209	0.1049	0.0013	0.0394	0.0003	0.0033	1.2070	0.1108	0.0046	0.0000	0.0114	0.6438	3.8460
20	5/15/2000	0.0096	1.1600	0.0010	0.0337	0.0023	0.0002	0.8730	0.0902	0.0064	0.0004	0.0167	0.3547	3.1960
21	5/18/2000	0.0348	2.9630	0.0037	0.1088	0.0083	0.0066	1.9910	0.1519	0.0054	0.0031	0.0166	3.3450	6.1610
22	5/21/2000	0.0008	1.9910	0.0014	0.0409	0.0011	0.0025	0.4828	0.0449	0.0038	0.0018	0.0099	2.0890	2.5760
23	5/24/2000	0.0057	1.8440	0.0010	0.0386	0.0013	0.0005	1.4190	0.0957	0.0045	0.0007	0.0553	1.3390	4.2690

Figure 5-2.— Example formatting of input concentration file.

The user must also provide an uncertainty file to give the model an estimate of the confidence the user has in each value. The uncertainties provided should encompass errors such as sampling and analytical errors. For some data sets, the analytical laboratory or reporting agency provides an uncertainty estimate for each value. However, uncertainties are not always reported and, when they are not available, must be estimated by the user. A discussion of calculating uncertainties is provided in Reff et al. (2007).

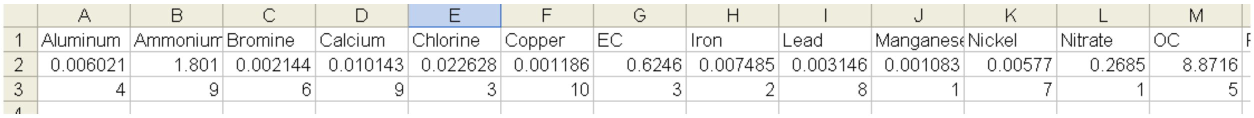
EPA PMF v3.0 accepts two types of uncertainty file: sample-specific and equation-based. The sample-specific uncertainty file provides an estimate of the uncertainty for each sample of each species. It should have the same dimensions as the concentration file, however, the uncertainty file should not include units. If the concentration file contains a row of units, the uncertainty file will have one less row than the concentration file. The user will be notified if the column and row headers do not match, but the program will continue. If the headers are different due to naming conventions but actually have the same order, the user should proceed. If not, the user should correct the problem outside the GUI and reload the files. Negative values and zero are not permitted as uncertainties; EPA PMF will provide an error message and the user will have to remove these values outside EPA PMF and reload the uncertainty file.

The equation-based uncertainty file provides species-specific parameters that EPA PMF v3.0 uses to calculate uncertainties for each sample. This file should have one column for each species, with species names as the column header (Figure 5-3). The first row under the species name is the detection limit; the second row is the error fraction. The error fraction should be the percent uncertainty x 100. Zeroes or negatives are not permitted for either the detection limit or the percent uncertainty. If the concentration is less than or equal to the method detection limit (MDL) provided, the uncertainty is calculated using the following equation (Polissar et al., 1998).

$$Unc = \frac{5}{6} \times MDL \tag{1-1}$$

If the concentration is greater than the MDL provided, the calculation is

$$Unc = \sqrt{(Error\ Fraction \times concentration)^2 + (MDL)^2} \tag{1-2}$$



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Aluminum	Ammonium	Bromine	Calcium	Chlorine	Copper	EC	Iron	Lead	Manganese	Nickel	Nitrate	OC
2	0.006021	1.801	0.002144	0.010143	0.022628	0.001186	0.6246	0.007485	0.003146	0.001083	0.00577	0.2685	8.8716
3	4	9	6	9	3	10	3	2	8	1	7	1	5

Figure 5-3.— Example of an equation-based uncertainty file.

The user can specify a “Missing Value Indicator” in the “Input Files” box on the **Input/Output Files** screen, which can be any numeric value. The user should use caution not to choose a numeric indicator that could potentially be a real concentration. The GUI will either remove the entire sample or replace the species concentration with the median concentration of that species and the uncertainty with four times the median concentration. For example, if the user specifies “-999” as the missing value indicator, and chooses to replace the species with the median, the GUI will find all instances of “-999” in the data file and replace them with the species-specific median. The GUI will also replace all associated uncertainty values with a high uncertainty of four times the species-specific median. If all samples of a species are missing, that species is automatically categorized as “bad” and excluded from further analysis.

Whenever new input files are provided by the user, the GUI clears all output displays from previous runs. The user should take care to save all relevant graphics before providing new data sets to the GUI.

5.2 Output Files

The user defines the output directory (“Output Folder”) and chooses the EPA PMF output file types (“Output File Type” radio buttons): tab-delimited text, **.txt**; comma-separated variable, **.csv**; or MS Excel, **.xls** in “Output Files” (Figure 2, 2). Five output files are automatically created by EPA PMF during base runs and are saved in the output folder selected by the user (if MS Excel output is designated by the user, the files are represented as separate tabs in ***_base.xls**):

- ***_diag** contains a record of the user inputs and model diagnostic information,
- ***_contrib** contains the contributions for each base run,
- ***_profile** contains the profiles for each base run,
- ***_resid** contains the residuals (regular and scaled by the uncertainty) for each base run, and
- ***_strength** contains the factor strength for each base run,

where * is the user-specified output file name prefix. The content of these output files are described in detail in Section 6.3.3. Additional files are created and saved after bootstrapping (***_profile_boot**) and Fpeak (***_fpeak**) have been performed. The file, ***_profile_boot**, contains the number of bootstrap runs mapped to each base run, each bootstrap profile that was mapped to the base profile, and all bootstrapping statistics generated by the GUI. The file, ***_fpeak** contains the profiles and contributions of each fpeak run.

5.3 Configuration files

EPA PMF saves user preferences in a configuration file (Figure 2, 3). The details saved include input files, output file location, qualifier, file type, species categorization, and all run specifications from the **Model Execution** screen (see Figure 9). Previous model output is not saved in the configuration file. To save or load a configuration file, the user can click on “Browse” to browse to the correct path or type in a path and name. The user should then select “Load Configuration from File” to open a configuration or “Save Current Configuration to File” to save the current settings to a configuration file.

6.0 BASIC OPERATIONS

6.1 Suggested Order of Operations

The GUI is designed to give the user as much flexibility as possible when running the model. However, certain steps must be completed before other steps are possible. The order of operations is based on how the tabs are arranged (from left to right) in the program (Figure 5); the sections in this User's Guide also follow this order. To begin using the program, the user must provide input files before other operations are available. The first time PMF is performed on the data set, the user should look at the data via the **Analyze Input Data** screen. This step is usually followed by "**Model Execution**" and "**Base Model Results**"; these steps should be repeated as needed until the user reaches a reasonable solution. Once a solution is chosen, the user should perform bootstrap runs in the "**Model Execution**" screen; the results are output to the "**Bootstrap Model Results**" screen. Advanced users may wish to initiate Fpeak runs, again from the "**Model Execution**" screen, with results presented in the "**Fpeak Model Results**" screen. Each of these operations is explained in detail below.

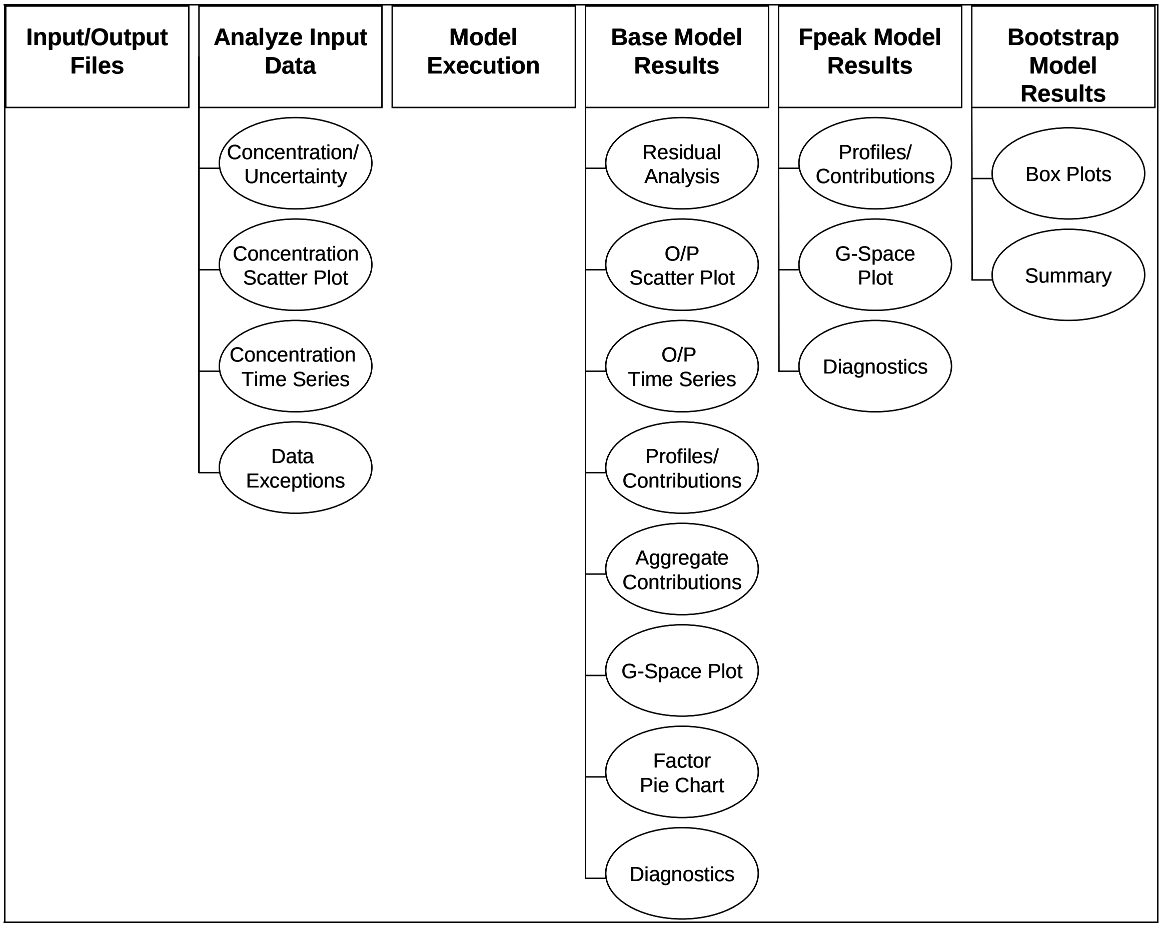


Figure 6-1.— Flow chart of tabs within EPA PMF v3.0.

6.2 Analyze Input Data

Several tools are available to help the user analyze the concentration and uncertainty data before running the model. These tools help the user decide whether certain species should be excluded or down-weighted (for example, due to increased uncertainty or a low signal-to-noise ratio), or if certain samples should be excluded (for example, due to an outlier event). All changes and deletions should be reported with the final solution. The four sub-screens of the **Analyze Input Data** screen and their uses are described below.

6.2.1 Concentration/Uncertainty

Input data statistics and concentration/uncertainty scatter plots are presented in the **Concentration/Uncertainty** screen, as shown in Figure 6-2. The following statistics are calculated for each species and displayed in a table on the left of the screen (Figure 6-2, 1):

- Signal-to-noise ratio (S/N) –indicates whether the variability in the measurements is real or within the noise of the data. In EPA PMF v3.0, it is calculated as

$$\left(\frac{S}{N}\right)_j = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij}-s_{ij}\right)^2}{\sum_{i=1}^{n} s_{ij}^2}} \tag{6-1}$$

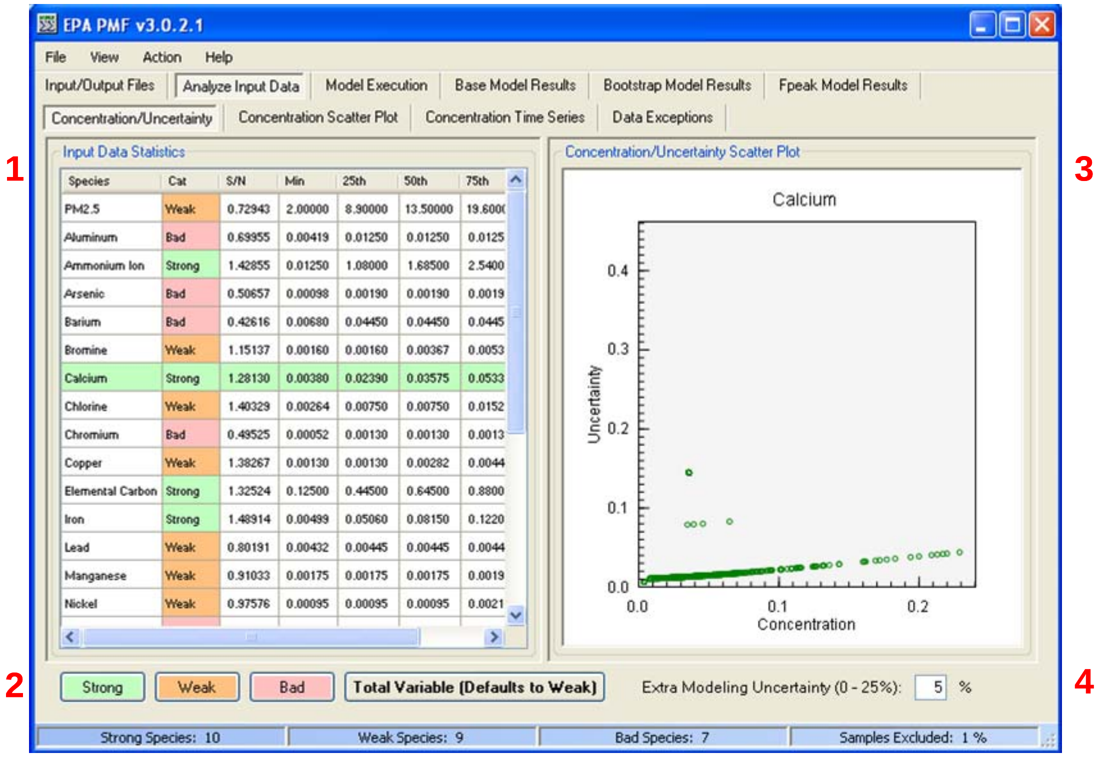
- Minimum (Min) – minimum concentration value
- 25th percentile (25th)
- Median – 50th percentile
- 75th percentile (75th)
- Maximum (Max) – maximum value reported

Based on these statistics, and knowledge of the data set, the user can categorize a species as “Strong”, “Weak”, or “Bad” by selecting the species in the **Input Data Statistics** table (Figure 6, 1) and selecting the appropriate button under the table (Figure 6, 2). Guidelines for using signal-to-noise ratios to determine a species categorization are presented in Paatero and Hopke (2003); they suggest categorizing a species as “bad” if the signal-to-noise ratio is less than 0.2 and “weak” if the signal-to-noise ratio is greater than 0.2 but less than 2. Species with low signal-to-noise ratio and/or a high percentage of data below detection will likely not provide enough variability in concentrations to meaningfully contribute to factor identification and will contribute to the noise in the results. The default value for all species is “Strong”. A categorization of “Weak” triples the provided uncertainty, and a categorization of “Bad” excludes the species from the rest of the analysis. If a species is marked “Weak”, the row is highlighted orange; if a species is marked “Bad”, the row is highlighted pink. Other than the statistics presented in the GUI, the user should consider other supplementary information that may be available: is the species present in sources in the area; is the species chemically distinct; how many samples are missing or below detection; known problems with the collection or analysis of the species, and is the species reactive or not conserved? A discussion of these considerations is provided in Reff et al. (2007).

Concentration/uncertainty scatter plots are displayed on the right of the screen (Figure 6, 3). The species to be plotted is selected in the **Input Data Statistics** table either by clicking on the species row using the mouse or scrolling up and down through the species. Only one species can be displayed at a time. The X axis is the concentration and Y axis is the uncertainty. The graph title is the name of the species plotted. If a user changes a species categorization to “Weak”, the concentration/uncertainty scatter plot for that species will be updated to three times the original uncertainty and the data points will be changed to orange squares. If a user changes a species categorization to “Bad”, the graph for that species will not be displayed.

The user can also add “Extra Modeling Uncertainty (0-25%)”, which is applied to all species, by entering a value in the box in the lower right corner of the screen (Figure 6, 4). This value encompasses various errors not considered measurement or lab errors (which are included in the user-provided uncertainty files). Some issues that could cause modeling errors include variation of source profiles, and chemical transformations in the atmosphere. The model uses the “Extra Modeling Uncertainty” variable to calculate “sigma”, which corresponds to total uncertainty (modeling uncertainty plus species/sample-specific uncertainty). If the user specifies extra modeling uncertainty, all concentration/uncertainty graphs will be updated to reflect the increase in uncertainty.

As shown in equation 1-2, the uncertainty values are an important piece of information in the PMF model. Any changes to the uncertainty should be documented by the user and reported with the final solution.



Species	Cat	S/N	Min	25th	50th	75th
PM2.5	Weak	0.72943	2.00000	8.90000	13.50000	19.600(
Aluminum	Bad	0.69955	0.00419	0.01250	0.01250	0.0125
Ammonium Ion	Strong	1.42855	0.01250	1.08000	1.68500	2.5400
Arsenic	Bad	0.50657	0.00098	0.00190	0.00190	0.0019
Barium	Bad	0.42616	0.00680	0.04450	0.04450	0.0445
Bromine	Weak	1.15137	0.00160	0.00160	0.00367	0.0053
Calcium	Strong	1.28130	0.00380	0.02390	0.03575	0.0533
Chlorine	Weak	1.40329	0.00264	0.00750	0.00750	0.0152
Chromium	Bad	0.49525	0.00052	0.00130	0.00130	0.0013
Copper	Weak	1.38267	0.00130	0.00130	0.00282	0.0044
Elemental Carbon	Strong	1.32524	0.12500	0.44500	0.64500	0.8800
Iron	Strong	1.48914	0.00499	0.05060	0.08150	0.1220
Lead	Weak	0.80191	0.00432	0.00445	0.00445	0.0044
Manganese	Weak	0.91033	0.00175	0.00175	0.00175	0.0019
Nickel	Weak	0.97576	0.00095	0.00095	0.00095	0.0021

Extra Modeling Uncertainty (0 - 25%): 5 %

Strong Species: 10 | Weak Species: 9 | Bad Species: 7 | Samples Excluded: 1 %

Figure 6-2.—Example “Concentration/Uncertainty” screen.

Also on this screen, the user can specify a “Total Variable” (Figure 6, 2) that will be used by the GUI in the post-processing of results. For example, if the data used are $PM_{2.5}$ components, the total variable would be $PM_{2.5}$ mass. The user specifies the total variable by selecting the species and pressing the “Total Variable” button beneath the **Input Data Statistics** table. Because a total variable should not have a large influence on the solution, it should be given a high uncertainty. Therefore, when a species is selected as a total variable, its categorization is automatically “Weak”. If the user has already adjusted the uncertainty of the total variable outside the GUI and wishes to categorize it as “Strong”, the default characterization can be overridden by selecting “Strong” for the variable after selecting “Total Variable”. A species designated “Bad” cannot be selected as a total variable, and a total variable cannot be made “Bad”.

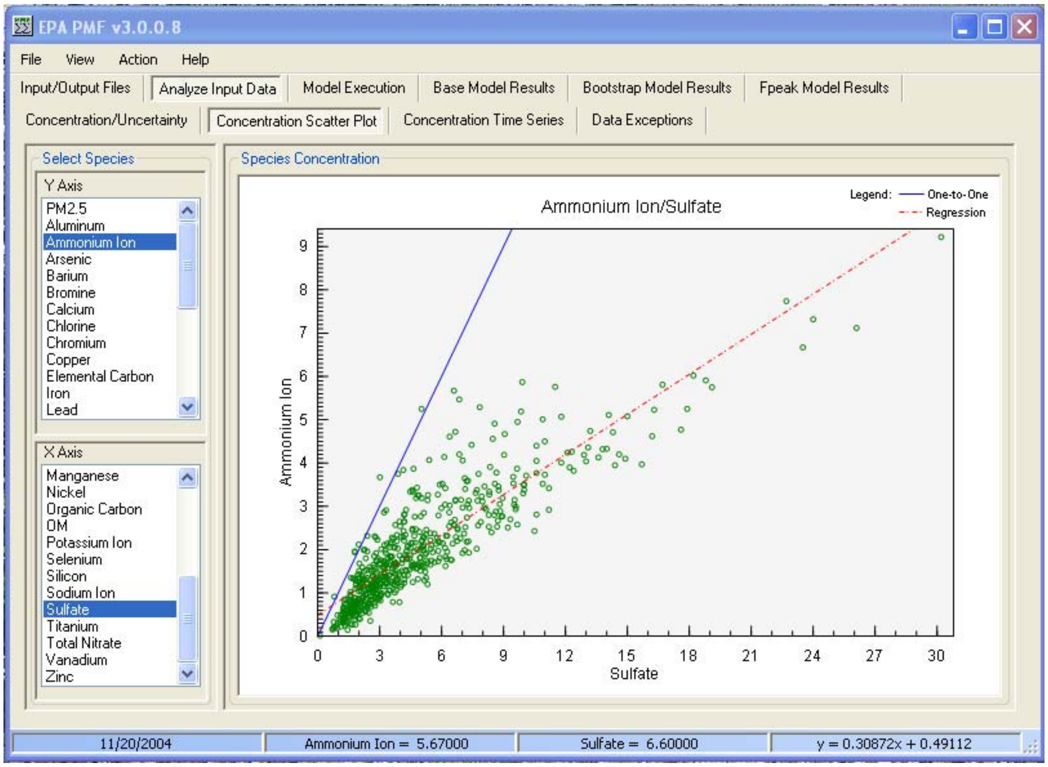
The status bar in the **Concentration/Uncertainty** screen displays the number of species of each category as well as the percentage of samples excluded by the user. Hot keys can be used to assign strong (Alt-S), weak (Alt-W), bad (Alt-B) and total variable (Alt-T).

6.2.2 Concentration Scatter Plots

Scatter plots between species are a useful pre-PMF analysis tool. A good correlation between species indicates a similar source or source type. A bifurcated line indicates multiple sources. The user should examine scatter plots to look for expected relationships, for example between soil components, as well as to look for other relationships that might indicate sources or source categories.

The **Concentration Scatter Plot** screen shows scatter plots between two user-specified species (Figure 7). The user selects the species for each axis in the appropriate “Y Axis” or “X Axis” list. Only one species can be selected for each axis. A one-to-one line (in blue) and linear regression line (dashed, red) are provided on the plot. Axis labels are the species names and units (if provided) and the plot title is “Y Axis Species/X Axis Species”.

The status bar on this screen shows the date, x-value, y-value, and regression equation for individual data points as the user mouses over them.



11/20/2004 | Ammonium Ion = 5.67000 | Sulfate = 6.60000 | y = 0.30872x + 0.49112

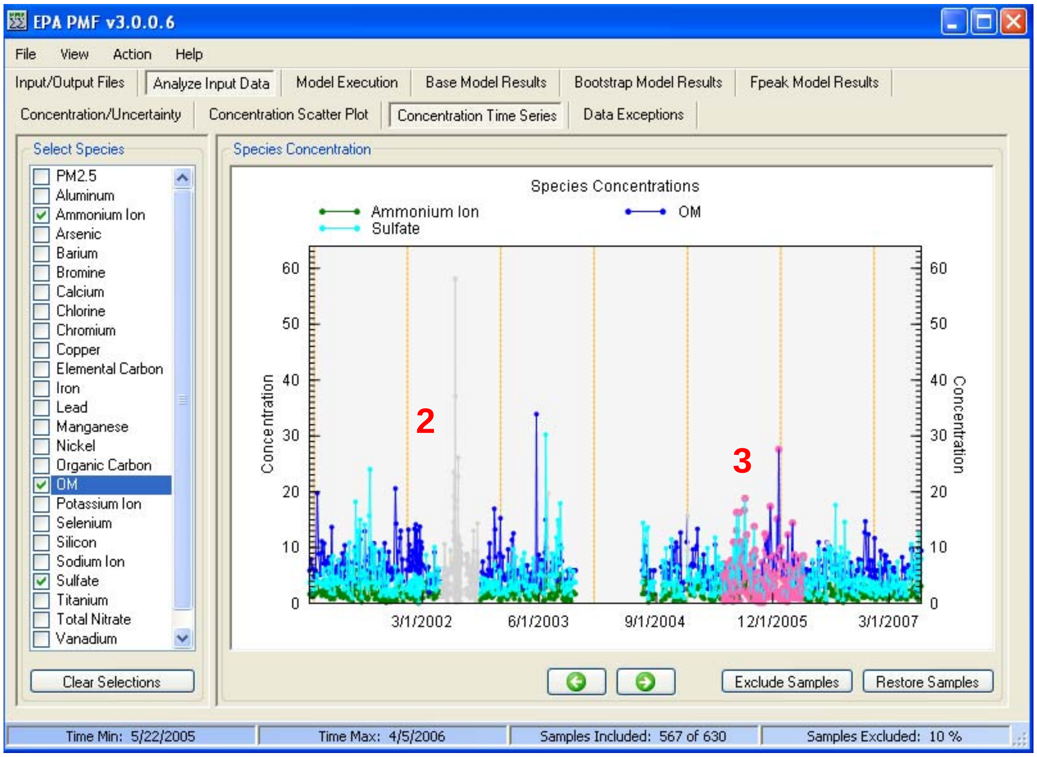
Figure 6-3.—Example concentration scatter plot.

6.2.3 Concentration Time Series

Time series of species concentrations (Figure 8) are useful to determine whether expected temporal patterns are present in the data and if there are any unusual events. By overlaying multiple species, the user can see if any unusual events are present across a group of species that may indicate a shared source. The user should also examine time series for extreme events that should be excluded from modeling (for example, elevated potassium concentrations on the Fourth of July).

The user can select up to 10 species in the **Concentration Time Series** list by checking the box next to the species name (Figure 8, 1). The selected species will be displayed in varying colors on the plot. To clear all species from the plot, the user should select “Clear Selections” below the list. Vertical orange lines denote January 1 of each year (if appropriate) for reference. A legend is provided at the top of the graph with species names and units (if available). The legend automatically updates with each selection. The arrow buttons below the plot, or the right and left arrow keys on the keyboard, can be used to scroll through samples. If a group of samples is selected, the arrows will move the first selected sample forward/backward by one sample. Samples can be removed from analysis by selecting individual data points with a single mouse click or dragging the mouse over a range of dates and pressing the “Exclude Samples” button below the plot. If a sample has been removed, it is grayed out for all species (example in Figure 8, 2) and can be included again by selecting the data point/range on any time series graph and pressing “Include Sample”. The sample will be highlighted in pink on the plot if it has been selected (example in Figure 8, 3). If a sample is removed from analysis, it will not be included in the statistics or plots generated by EPA PMF or in any model output. It is not removed from the original user input files. Hot keys can be used to restore (Alt-R) or exclude (Alt-E) selected samples.

The status bar on this screen shows the minimum and maximum sample dates for the selected range, the number of samples included out of the total number of samples, and the percent of samples excluded by the user.



□ PM2.5
□ Aluminum
☑ Ammonium Ion
□ Arsenic
□ Barium
□ Bromine
□ Calcium
□ Chlorine
□ Chromium
□ Copper
□ Elemental Carbon
□ Iron
□ Lead
□ Manganese
□ Nickel
□ Organic Carbon
☑ OM
□ Potassium Ion
□ Selenium
□ Silicon
□ Sodium Ion
☑ Sulfate
□ Titanium
□ Total Nitrate
□ Vanadium

Time Min: 5/22/2005 | Time Max: 4/5/2006 | Samples Included: 567 of 630 | Samples Excluded: 10 %

Figure 6-4.—Example of **Concentration Time Series** screen with a range of samples excluded (grey, 2), and with a range of samples selected (pink, 3).

6.2.4 Data Exceptions

Changes made by the GUI to the input data are detailed in the **Data Exceptions** screen. These changes include designating a species “Weak” or “Bad”, excluding a sample via the **Concentration Time Series** screen, or excluding a sample using the “Missing Value Indicator”.

6.3 Base Runs

Model runs are referred to as “base runs” by EPA PMF as they are the basis for advanced analyses using bootstrapping or Fpeak. Each set of base runs uses the same model input and a seed value as the starting point for iterations. If a random seed is used, the base runs will have different starting points and may converge to different solutions (local minima). A user can test if the solution found is a local or global minima by using many random seeds and examining whether the minima is constant. If a specific seed is supplied by the user, that seed will be used in a pseudo-random number generator to generate seeds for each run. Each run in a set of base runs will have a different seed, but if the base runs are re-run, the same seeds will be generated in the same order.

6.3.1 Initiating a Base Run

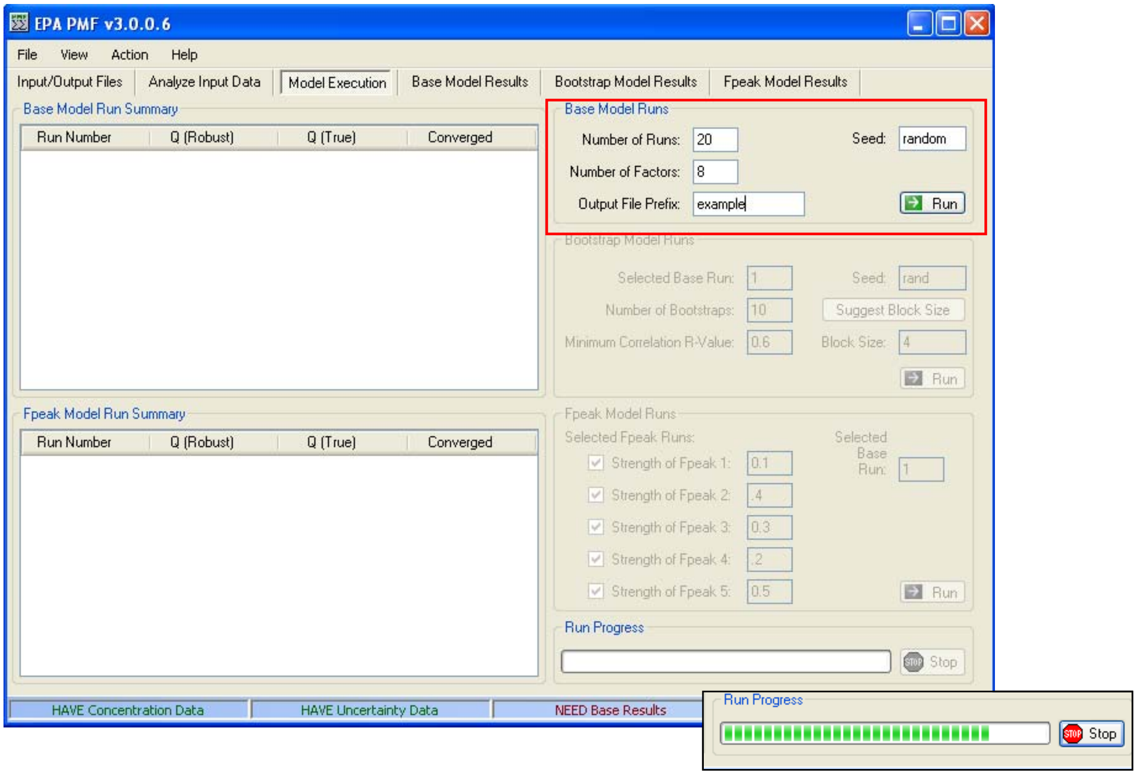
Base runs are initiated on the **Model Execution** screen. Inputs for the base runs are provided in **Base Model Runs** (Figure 9, red box). The user must specify several parameters that determine how the model is run:

- “Number of Runs” – the number of base runs to be performed, this number must be an integer between 1 and 999. The recommended number of runs is 20.
- “Number of Factors” – the number of factors the model should fit; this number must be an integer between 1 and 999. The number of factors to be chosen will depend on the user’s understanding of the sources impacting the airshed, number of samples, sampling frequency, and species characteristics.

- “Output File Prefix” – the prefix that will be used as the first part of any output file, this prefix can be any character or string of characters. If this prefix is not changed when a new run is initiated, previous files will be overwritten with no prompt.
- “Seed” – the starting point for each iteration by ME-2. The default seed is “random”, which tells the GUI to randomly choose a starting point for each run. To reproduce results, fix the seed, number of runs and factors (for example - Seed = 25, Number of Runs = 20, Number of Factors = 7).

After the above parameters are specified, the user should press the “Run” button in **Base Model Runs** to initiate the base runs. Once runs are initiated, the “Run Progress” box in the lower right corner of the screen activates. Base runs can be terminated at any time by pressing the “Stop” button in the “Run Progress” box. The progress bar in this box also fills whenever runs are being performed. No information about the runs will be saved or displayed if the runs are stopped.

The status bar on the **Model Execution** screen displays the same information as on the **Input/Output Files** screen.



Run Number	Q (Robust)	Q (True)	Converged

Number of Runs: 20
Seed: random
Number of Factors: 8
Output File Prefix: example

Selected Base Run: 1
Seed: rand
Number of Bootstraps: 10
Minimum Correlation R-Value: 0.6
Block Size: 4

Run Number	Q (Robust)	Q (True)	Converged

☑ Strength of Fpeak 1: 0.1
☑ Strength of Fpeak 2: .4
☑ Strength of Fpeak 3: 0.3
☑ Strength of Fpeak 4: .2
☑ Strength of Fpeak 5: 0.5
Selected Base Run: 1

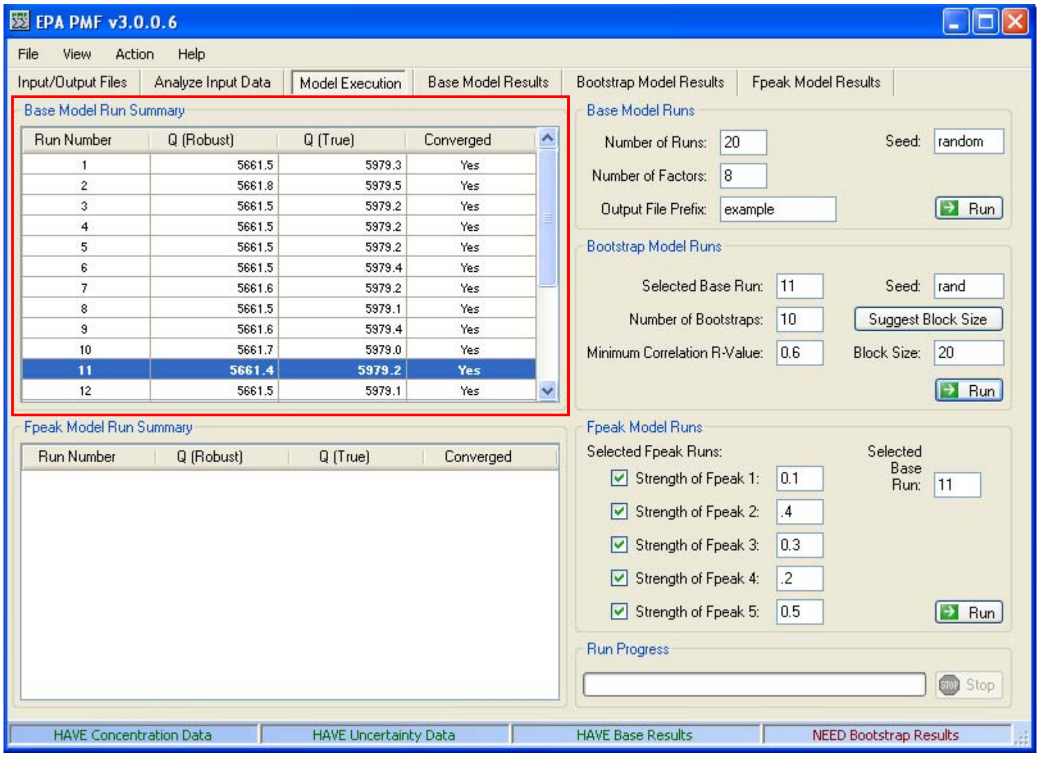
HAVE Concentration Data
HAVE Uncertainty Data
NEED Base Results

Figure 6-5.—Example **Model Execution** screen before base runs and active “Run Progress” box.

6.3.2 Base Run Summary

When the base runs are completed, a summary of each run appears on the left of the **Model Execution** screen in the **Base Model Run Summary** table (Figure 10, red box). The Q values are goodness-of-fit parameters calculated using Equation 1-2 and are an assessment of how well the model fit the input data. The lowest Q(robust) value is boldfaced and is automatically highlighted by the GUI. This summary includes the Q(robust) and Q(true) for each run, as well as whether the run converged. Q(robust) is the goodness-of-fit parameter calculated excluding outliers, defined as samples for which the scaled residual is greater than 4 and the Q(true) is calculated including all points. The theoretical Q is not calculated by EPA PMF but can be approximated by the user as *nm – p(n+m)*, where *n* is the number of species, *m* is the number of samples in the data set, and *p* is the number of factors fitted by the model. Solutions where Q(true) is greater than 1.5 times Q(robust) indicate that peak events may be disproportionately influencing the model.

Only converged solutions should be investigated further using the tools available in EPA PMF 3.0. Non-convergence implies that the model did not find any minima. Several things could cause non-convergence, including uncertainties that are too low, specified incorrectly, or inappropriate input parameters.



Base Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged
1	5661.5	5979.3	Yes
2	5661.8	5979.5	Yes
3	5661.5	5979.2	Yes
4	5661.5	5979.2	Yes
5	5661.5	5979.2	Yes
6	5661.5	5979.4	Yes
7	5661.6	5979.2	Yes
8	5661.5	5979.1	Yes
9	5661.6	5979.4	Yes
10	5661.7	5979.0	Yes
11	5661.4	5979.2	Yes
12	5661.5	5979.1	Yes

Base Model Runs

Number of Runs: 20 Seed: random
Number of Factors: 8
Output File Prefix: example

Bootstrap Model Runs

Selected Base Run: 11 Seed: rand
Number of Bootstraps: 10
Minimum Correlation R-Value: 0.6 Block Size: 20

Fpeak Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged

Fpeak Model Runs

Selected Fpeak Runs:
☑ Strength of Fpeak 1: 0.1
☑ Strength of Fpeak 2: .4
☑ Strength of Fpeak 3: 0.3
☑ Strength of Fpeak 4: .2
☑ Strength of Fpeak 5: 0.5

Selected Base Run: 11

Run Progress

HAVE Concentration Data | HAVE Uncertainty Data | HAVE Base Results | NEED Bootstrap Results

Figure 6-6.—Example **Model Execution** screen after base runs have been run.

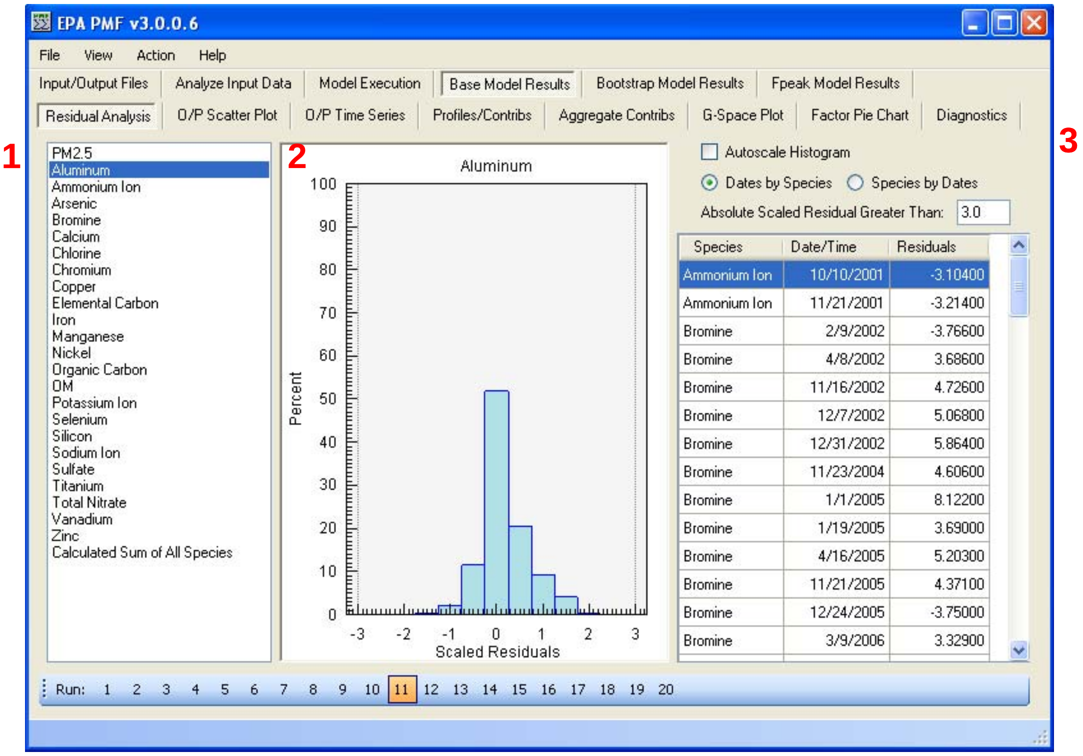
6.3.3 Base Run Results

Details of the base run results are provided in the sub-screens of the **Base Model Results** screen. A run is chosen either by highlighting it in the **Base Model Run Summary** table on the **Model Execution** screen, or by selecting the run number at the bottom of the **Base Model Results** screen. Selecting a run on one screen will select the same run on all screens. Additionally, selecting a species on the **Residual Analysis**, **O/P Scatter Plots**, or **O/P Time Series** sub-screens will select the same species throughout the program.

Residual Analysis

The **Residual Analysis** screen (Figure 11) displays the scaled residuals in several formats. At the left of the screen (Figure 11, 1), the user can select a species which will be displayed in the histogram in the center of the screen (Figure 11, 2). The histogram shows the percent of all scaled residuals in a given bin. Each bin is equal to 0.5. These plots are useful to determine how well the model fit each species. If a species has many large scaled residuals or displays a non-normal curve, it may be an indication of a poor fit. The species in Figure 11 (Aluminum) is well-modeled; all residuals are between +3 and -3 and they are normally distributed (from O/P Scatter Plot Screen, below). Grey lines are provided for reference at +3 and -3. The user can use the “Autoscale Histogram” box (Figure 11, 3) to adjust the y-axis of the histogram. If the box is checked, the Y axis will be set to the maximum value + 10% for each species. If the box is unchecked, the Y axis maximum is fixed at 100%. Checking the “Autoscale Histogram” function is helpful when examining individual species and the shape of their distributions; leaving the “Autoscale Histogram” function unchecked is helpful when comparing species.

The screen also displays the samples with scaled residuals that are greater than a user-specified value (Figure 11, 3). The default value is 3.0. The residuals can be displayed as Dates by Species or Species by Dates by choosing the appropriate option above the table. When a species is selected in the list on the left (Figure 11, 1), the table on the right (Figure 11, 3) automatically scrolls to that species.



☐ Autoscale Histogram

● Dates by Species ○ Species by Dates

Absolute Scaled Residual Greater Than: 3.0

Species	Date/Time	Residuals
Ammonium Ion	10/10/2001	-3.10400
Ammonium Ion	11/21/2001	-3.21400
Bromine	2/9/2002	-3.76600
Bromine	4/8/2002	3.68600
Bromine	11/16/2002	4.72600
Bromine	12/7/2002	5.06800
Bromine	12/31/2002	5.86400
Bromine	11/23/2004	4.60600
Bromine	1/1/2005	8.12200
Bromine	1/19/2005	3.69000
Bromine	4/16/2005	5.20300
Bromine	11/21/2005	4.37100
Bromine	12/24/2005	-3.75000
Bromine	3/9/2006	3.32900

Figure 6-7.—Example **Residual Analysis** screen.

An additional residual calculation comparing the residuals between base runs is performed for the base runs but is not displayed on the **Residual Analysis** screen. These results are recorded in a diagnostic file and can be viewed through the GUI by selecting the **Diagnostics** tab. First, the sum of the squared difference between the scaled residuals for each pair of base runs is calculated for each variable as follows:

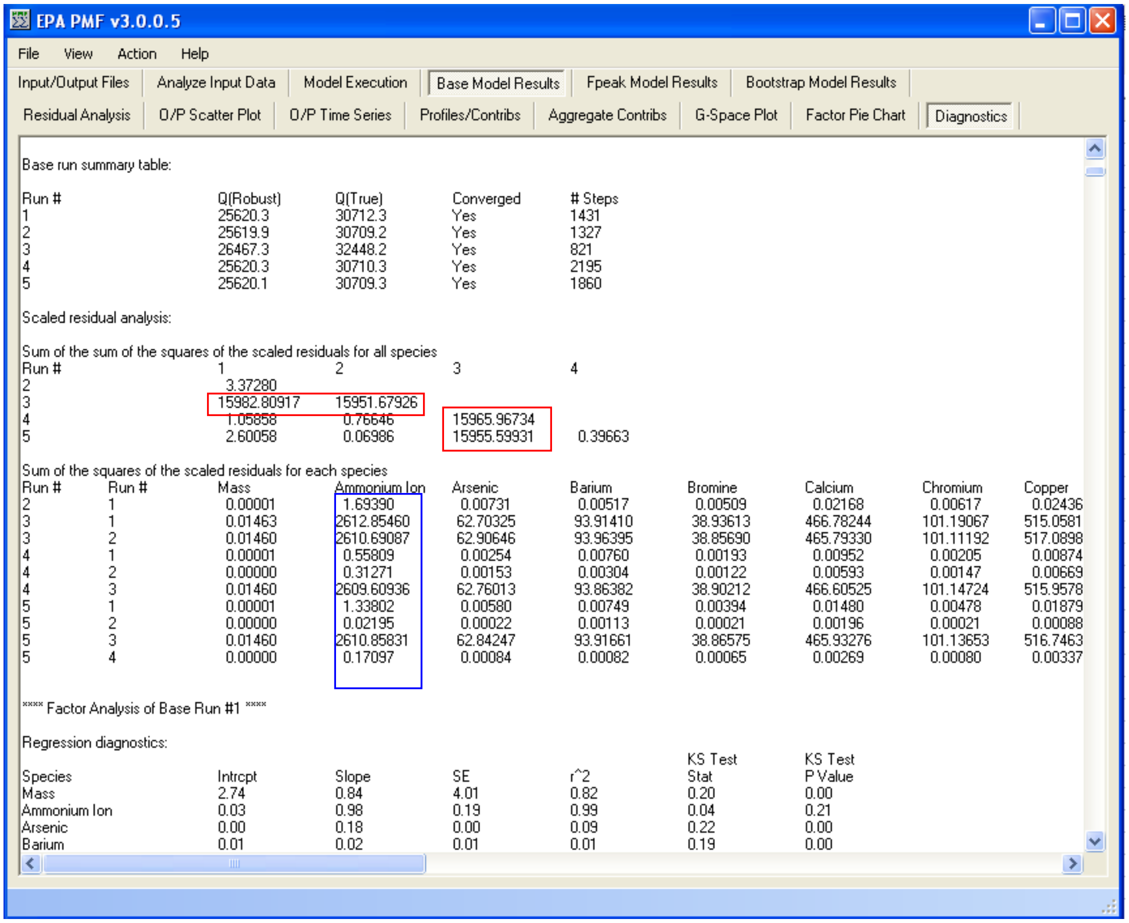
$$d_{jkl} = \sum_{i} \left(r_{ijk} - r_{ijl} \right)^2 \qquad \textbf{(6-2)}$$

where *r* is the scaled residual, *j* is the variable, *i* is the sample, and *k* and *l* are two different runs.

The *d* values for each species are then summed for each pair of runs:

$$D_{kl} = \sum_{j} D_{jkl} \qquad \textbf{(6-3)}$$

The *D* values are reported in a matrix of base run pairs. The user should examine this matrix for large variations, indicating that two runs resulted in truly different solutions rather than merely being rotations of each other. If different solutions are seen, the user can then examine the *d* values, which will indicate the individual species that are fitted differently across the runs. Figure 12 shows an example where run 3 (red boxes) is clearly a different solution than the other runs. Examining the results for each species shows that ammonium ion (blue box) has high *D* values that are a result of the model reaching a different solution in run 3.



```
Base run summary table:

Run #                 Q(Robust)    Q(True)      Converged    # Steps
1                     25620.3      30712.3      Yes          1431
2                     25619.9      30709.2      Yes          1327
3                     26467.3      32448.2      Yes          821
4                     25620.3      30710.3      Yes          2195
5                     25620.1      30709.3      Yes          1860

Scaled residual analysis:

Sum of the sum of the squares of the scaled residuals for all species
Run #                 1            2            3            4
2                      3.37280
3                     15982.80917  15951.67926
4                      1.05858      0.76646     15965.96734
5                      2.60058      0.06986     15955.59931   0.39663

Sum of the squares of the scaled residuals for each species
Run #     Run #       Mass         Ammonium Ion Arsenic      Barium       Bromine      Calcium      Chromium     Copper
2         1            0.00001      1.69390      0.00731      0.00517      0.00509      0.02168      0.00617      0.02436
3         1            0.01463     2612.85460   62.70325     93.91410     38.93613     466.78244    101.19067    515.0581
3         2            0.01460     2610.69087   62.90646     93.96395     38.85690     465.79330    101.11192    517.0898
4         1            0.00001      0.55809      0.00254      0.00760      0.00193      0.00952      0.00205      0.00874
4         2            0.00000      0.31271      0.00153      0.00304      0.00122      0.00593      0.00147      0.00669
4         3            0.01460     2609.60936   62.76013     93.86382     38.90212     466.60525    101.14724    515.9578
5         1            0.00001      1.33802      0.00580      0.00749      0.00394      0.01480      0.00478      0.01879
5         2            0.00000      0.02195      0.00022      0.00113      0.00021      0.00196      0.00021      0.00088
5         3            0.01460     2610.85831   62.84247     93.91661     38.86575     465.93276    101.13653    516.7463
5         4            0.00000      0.17097      0.00084      0.00082      0.00065      0.00269      0.00080      0.00337

**** Factor Analysis of Base Run #1 ****

Regression diagnostics:
                                                                         KS Test      KS Test
Species               Intrcpt      Slope        SE           r^2         Stat         P Value
Mass                  2.74         0.84         4.01         0.82        0.20         0.00
Ammonium Ion          0.03         0.98         0.19         0.99        0.04         0.21
Arsenic               0.00         0.18         0.00         0.09        0.22         0.00
Barium                0.01         0.02         0.01         0.01        0.19         0.00
```

Figure 6-8.—Example of a residual analysis.

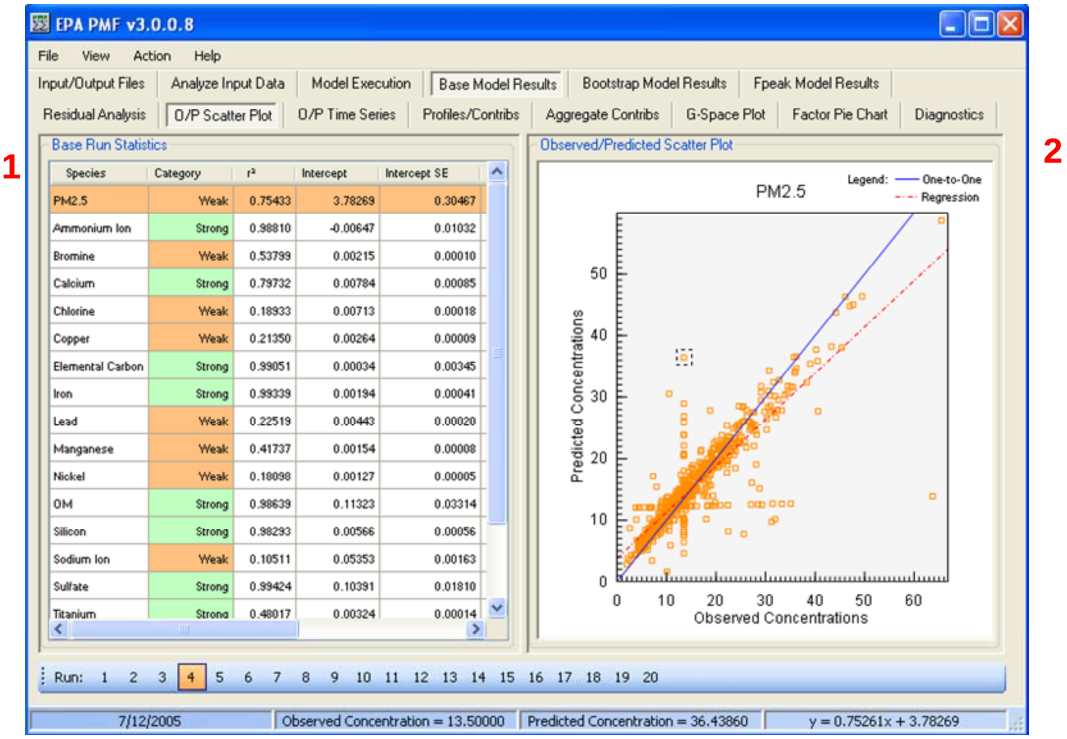
Observed/Predicted (O/P) Scatter Plot

A comparison between observed (input data) values and predicted (modeled) values is useful to determine if the model fits the individual species well. Species that do not have a good correlation between observed and predicted values should be evaluated by the user to determine if they should be down weighted or excluded from the model.

A table in the **O/P Scatter Plot** screen (Figure 13, 1) shows **Base Run Statistics** for each species. These numbers are calculated using the observed and predicted concentrations to indicate how well each species was fit by the model. The statistics shown are the coefficient of determination (r^2), Intercept, Intercept SE (standard error), Slope, Slope SE, and SE. The table also indicates if the residuals are normally distributed, as determined by a Kolmogorov-Smirnoff test. If the test indicates that the residuals are not normally distributed, the user should visually inspect the histogram for outlying residuals. If not all statistics are visible, the user can use the scroll bars at the bottom and side of the table to display additional statistics. These statistics are also provided in the *_diag output file.

The **O/P Scatter Plot** (Figure 13, 2), shows the observed (X axis) and predicted (Y axis) concentrations for the selected species. A blue 1:1 line is provided on this plot for reference (a perfect fit would line up exactly on this line), and the regression line is shown as a dotted red line.

The status bar on this screen displays the date, x-value, y-value, and regression equation between predicted and observed data as data points are moused-over.



Base Run Statistics

Species	Category	r²	Intercept	Intercept SE
PM2.5	Weak	0.75433	3.78269	0.30467
Ammonium Ion	Strong	0.98810	-0.00647	0.01032
Bromine	Weak	0.53799	0.00215	0.00010
Calcium	Strong	0.79732	0.00784	0.00085
Chlorine	Weak	0.18933	0.00713	0.00018
Copper	Weak	0.21350	0.00264	0.00009
Elemental Carbon	Strong	0.99051	0.00034	0.00345
Iron	Strong	0.99339	0.00194	0.00041
Lead	Weak	0.22519	0.00443	0.00020
Manganese	Weak	0.41737	0.00154	0.00008
Nickel	Weak	0.18098	0.00127	0.00005
OM	Strong	0.98639	0.11323	0.03314
Silicon	Strong	0.98293	0.00566	0.00056
Sodium Ion	Weak	0.10511	0.05353	0.00163
Sulfate	Strong	0.99424	0.10391	0.01810
Titanium	Strong	0.48017	0.00324	0.00014

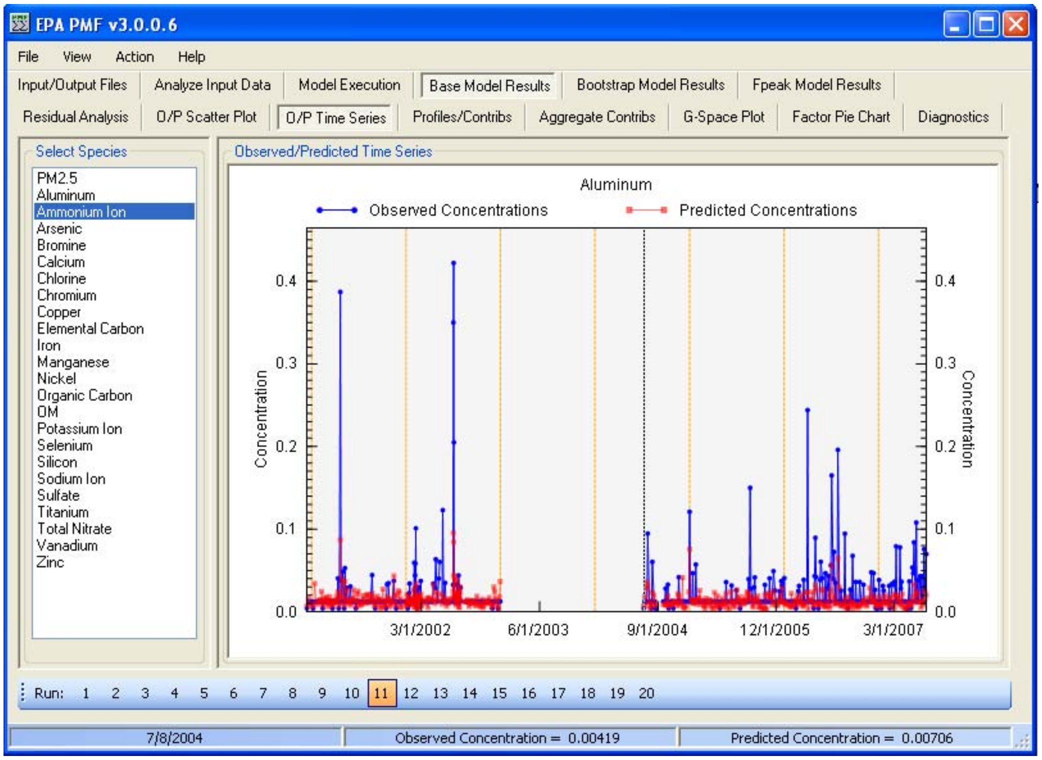
7/12/2005 | Observed Concentration = 13.50000 | Predicted Concentration = 36.43860 | y = 0.75261x + 3.78269

Figure 6-9.—Example **O/P Scatter Plot** screen.

O/P Time Series

The data displayed on the **O/P Scatter Plot** screen are the same data displayed as a time series on the **O/P Time Series** screen (Figure 14). A dotted black vertical reference line is provided at the date closest to the position of the mouse. When a species is selected by the user, the observed (user-input) data for that species are displayed in blue and the predicted (modeled) data are displayed in red. The user can view this screen to determine when the model is fitting the observed data well. If specific samples are not being modeled well across species, it might be advisable to exclude those samples and rerun the model (see Section 6.2.3).

The status bar on this screen displays the sample date, observed concentration, and predicted concentration of the moused-over data point.



7/8/2004 | Observed Concentration = 0.00419 | Predicted Concentration = 0.00706

Figure 6-10.—Example **O/P Time Series** screen.

Profiles/Contribs

The factors resolved by PMF are displayed under the "**Profiles/Contribs**" tab. Two graphs are shown for each factor, one displaying the factor profile and the other displaying the contribution per sample of each factor (Figure 15). The profile graph, displayed on top (Figure 15, 1), shows the mass of each species apportioned to the factor as a pale blue bar and the percent of each species apportioned to the factor as a red box. The mass bar corresponds to the left Y axis, which is a logarithmic scale. The percent of species corresponds to the right Y axis. The bottom graph shows the contribution of each factor to the total mass by sample (Figure 15, 2). Orange reference lines delineate years. This graph is normalized so that the average of all contributions for each factor is 1. If a total variable is selected, the user can select "Mass Units" in the bottom left corner of the screen to display the contributions in the same units as the total mass. If this option is selected, the GUI multiplies the contributions by the mass of the total variable in that factor. If no mass from the total variable is apportioned to the factor, the graph is not shown and the GUI instead displays "Total Variable" mass is 0 for this run/factor".

Two sets of buttons across the bottom of this screen allow the user to easily compare runs and factors. Beginning in the bottom left corner, each run can be chosen by clicking on the appropriate run number. The user can quickly compare runs to assess the stability of the solution or determine what, if any, individual species or factors are varying between runs. To the right of the run numbers are the factor numbers, which allow the user to switch between the factors resolved by PMF.

The status bar on this screen displays the date and contributions of data points as they are moused-over on the Factor Contributions plot.

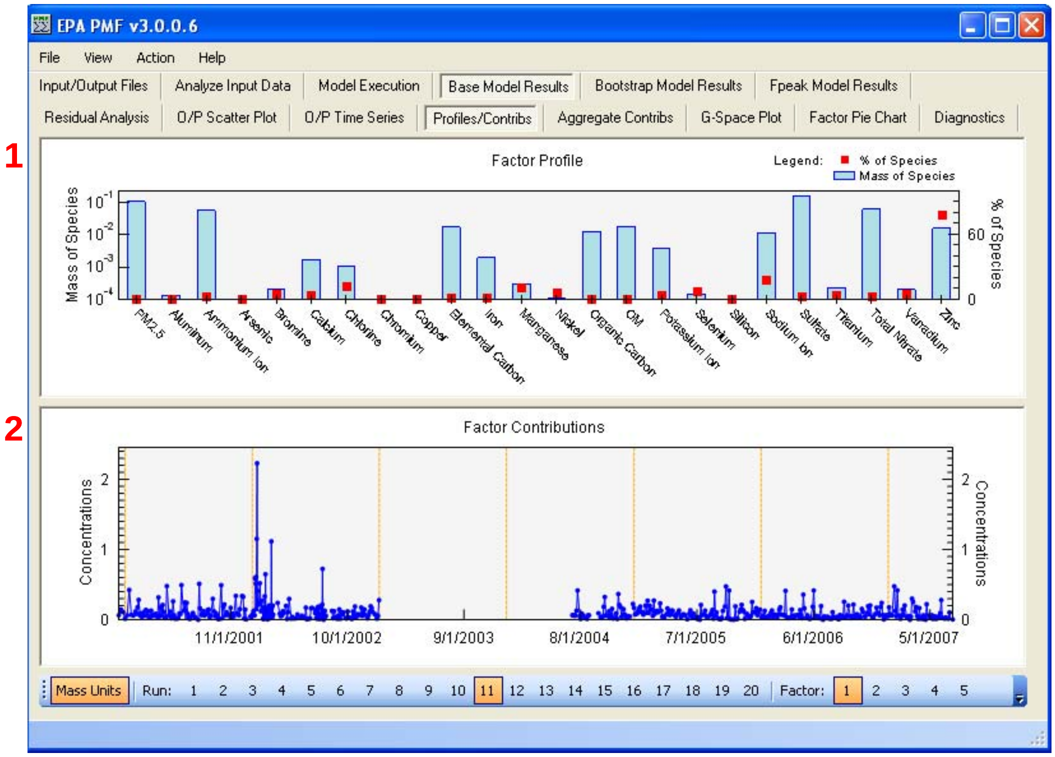


Figure 6-11.—Example **Profiles/Contrib** screen.

Aggregate Contribs

Three box plot graphs are displayed on each of the three sub-screens under the **Aggregate Contribs** screen. A box plot displays information about the distribution of data; in these plots, the box represents the interquartile range (25th-75th percent of contributions), the red line represents the median concentration, and the lines (or "whiskers") extend above and below the box to the 95th and 5th percentile of contributions, respectively. Dashed blue lines connect the median values of each box. The X axis gives the grouping and the number of data points represented by each box. A particularly large or small box could be caused by too few data points; if this is the case, the box should not be used in comparisons with other boxes.

The **Aggregate Contribs** screen is shown in Figure 16. The top graphic displays box plots for the selected factor by year (Figure 16, 1); the middle graphic by season (Figure 16, 2; winter is December–February, spring is March–May, summer is June–August, and fall is September–November); and the bottom graphic by day of week (Figure 16, 3; weekday is Monday–Friday, weekend is Saturday and Sunday).

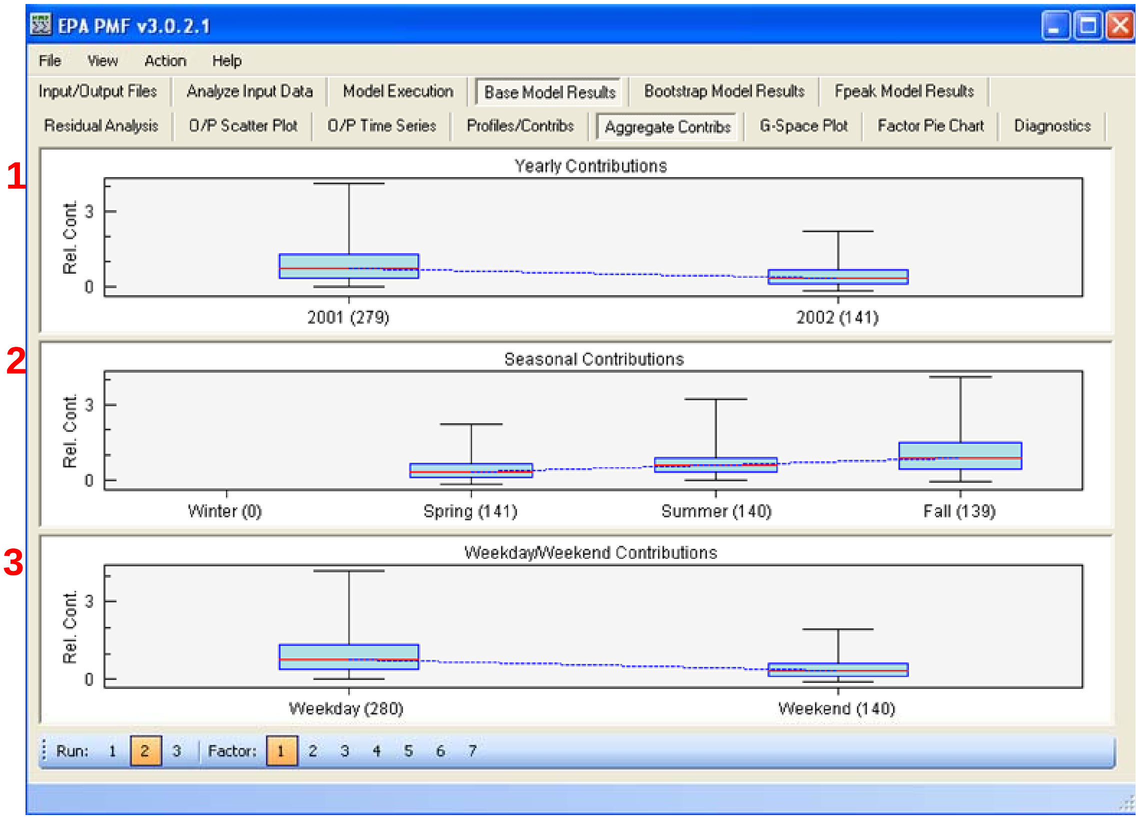
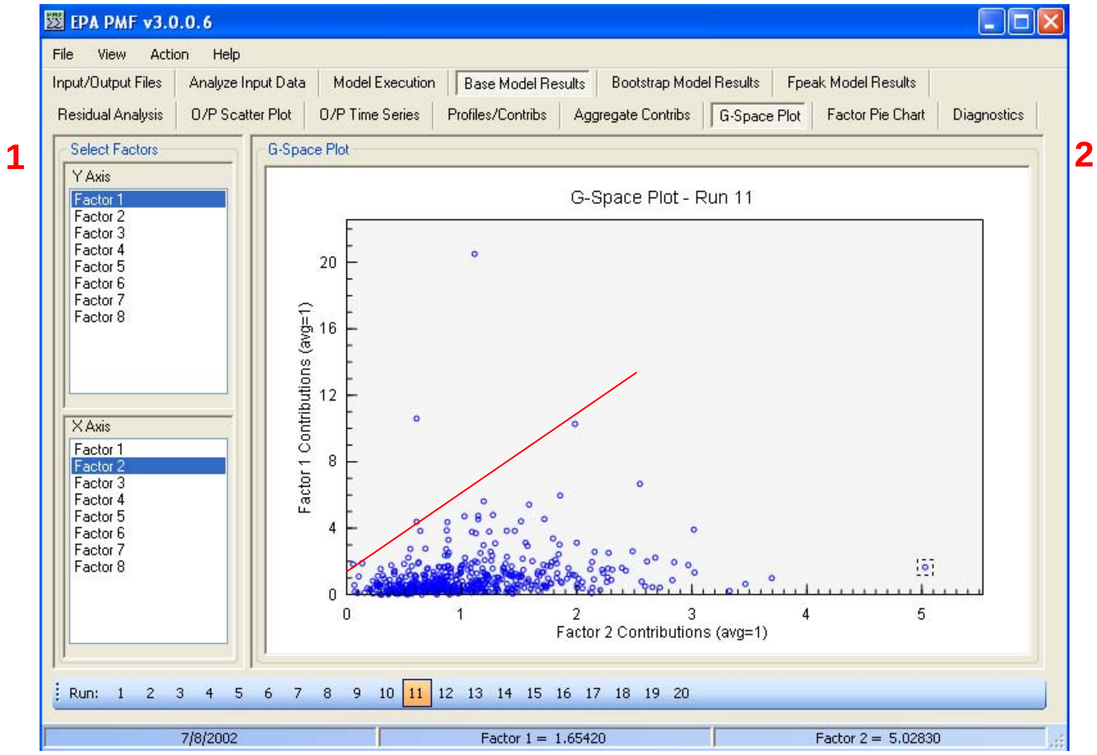


Figure 16. Example Aggregate Contibs screen

G-Space Plot

A **G-Space Plot** screen (**Figure 6-13**) shows scatter plots of one factor versus another factor which can be used to determine if the solution has filled the solution space or if it has some rotational ambiguity. If a solution fills the solution space, the edges of the scatter plot will correspond to the axes. The user selects one factor for the Y axis and one factor for the X axis from lists on the left of the screen (Figure 6-13, 1). A scatter plot of these factors is then shown on the right of this screen (Figure 6-13, 2). The plots in Figure 6-13 are an example of an unrealistic rotation of a factor, which appears as oblique edges on a G-Space plot (red line added for reference). In EPA PMF, the user can explore different rotations via the Fpeak option, which is explained in detail in Section 6.5 (Paatero et al., 2005).

The status bar on this screen displays the date, x-value, and y-value of moused-over data points.

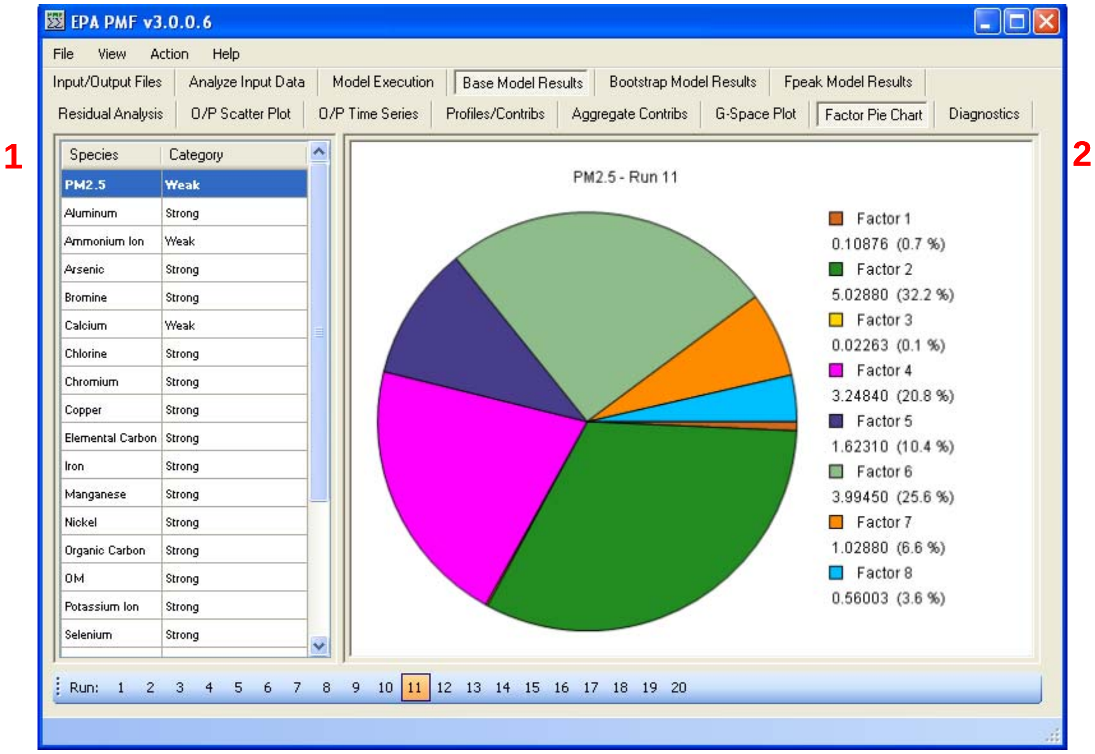


7/8/2002 | Factor 1 = 1.65420 | Factor 2 = 5.02830

Figure 6-13.—Example **G-Space Plot** screen with a red line indicating an edge.

Factor Pie Chart

The **Factor Pie Chart** screen displays the distribution of each species among the factors resolved by PMF. The species of interest is selected from the table on the left of the screen (Figure 18, 1). The categorization of that species is also displayed for reference. If a total variable was chosen by the user under the Analyze Input Data tab, that variable is boldfaced in the table. The pie chart for the selected species appears on the right side of the screen (Figure 18, 2). If the user has specified a total variable, the distribution of this variable across the factors will be of particular importance. The user may also want to examine the distribution of certain key species, such as toxic species, across factors.



Species	Category
PM2.5	**Weak**
Aluminum	Strong
Ammonium Ion	Weak
Arsenic	Strong
Bromine	Strong
Calcium	Weak
Chlorine	Strong
Chromium	Strong
Copper	Strong
Elemental Carbon	Strong
Iron	Strong
Manganese	Strong
Nickel	Strong
Organic Carbon	Strong
OM	Strong
Potassium Ion	Strong
Selenium	Strong

Figure 6-14.—Example **Factor Pie Chart** screen.

Diagnostics

The **Diagnostics** screen displays the ***_diag** file. It is updated as the ***_diag** file is updated.

Output files

After the base runs are completed, the GUI creates four output files (or one Excel file with four worksheets) that contain all of the data used for the on-screen display of results. These files are saved to the directory specified on the Input/Output Files screen, using the prefix specified on the Model Execution screen:

- ***_diag** contains a record of the user inputs and model diagnostic information (identical to the Diagnostics screen).
- ***_contrib** contains the contributions for each base run used to generate the contribution graphs on the Profile/Contribs tab. Contributions are sorted by run number. Normalized contributions are shown first, followed by contributions in mass units if a total variable is specified.
- ***_profile** contains the profiles for each base run used to generate the profile graphs on the Profile/Contribs tab. Profiles are sorted by run number. Profiles in mass units are written first, followed by profiles in percent of species and concentration fraction of species total if a total mass variable is specified.
- ***_resid** contains the residuals (regular and scaled by the uncertainty) for each base run, used to generate the graphs and tables on the Residual Analysis screen.

- ***_strength** contains the factor strength for each base run.

6.4 Bootstrap Runs

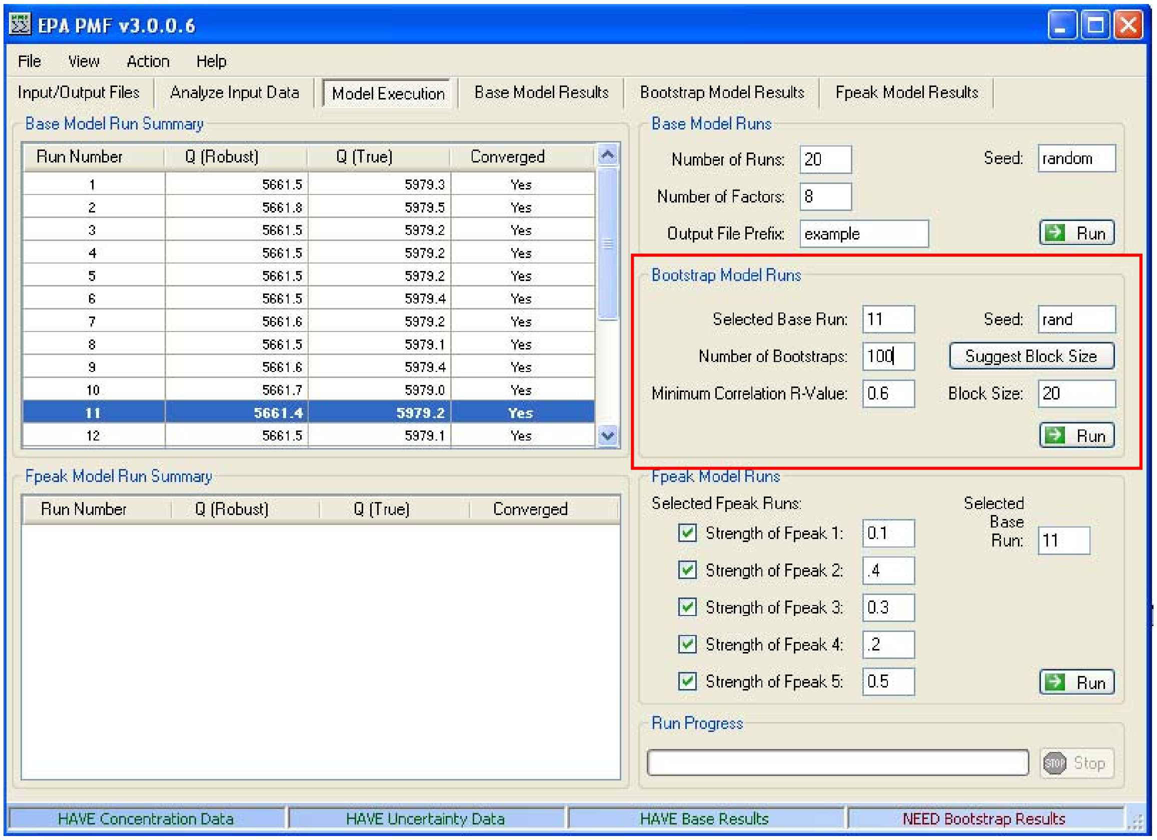
After the user has found a solution believed to be the local minima, bootstrapping is performed to estimate the stability and uncertainty of that solution. EPA PMF performs bootstrapping by randomly selecting non-overlapping blocks of samples (consecutive samples, block size supplied by user) and creating a new input data file of the selected samples, with the same dimensions as the original data set. PMF is then run on the new data set, and each bootstrap factor is mapped to a base run factor by comparing the contributions of each factor. The bootstrap factor is assigned to the base factor with which the bootstrap factor has the highest correlation, above a user-specified threshold. If no base factors have a correlation above the threshold for a given bootstrap factor, that factor is considered “unmapped”. If more than one bootstrap factor from the same run are correlated best with the same base factor, they will all be mapped to that base factor. This process is repeated for as many bootstrap runs as the user specifies. EPA PMF then summarizes all the bootstrapping runs. The user should examine the Q values and factor identifications for stability and the interquartile ranges around the profiles. These bootstrapping statistics should be reported with the final solution.

6.4.1 Initiating Bootstrap Runs

In EPA PMF v3.0, bootstrapping is initiated in the **Model Execution** screen, **Bootstrap Model Runs** (**Figure 19**, red box). As with the base runs, the user must specify several parameters for bootstrap runs:

- Selected Base Run – the base run to be used to map each bootstrap run. The base run can be designated by either selecting a run in the **Base Model Run Summary** table or manually entering a run number in the “Selected Base Run” text box under **Bootstrap Model Runs**.
- Number of Bootstraps – the number of bootstrap runs to be performed. For a final analysis, it is recommended that at least 100 bootstrap runs be performed to ensure the robustness of the statistics; for preliminary analysis, fewer bootstrap runs may be performed to quickly gauge the stability of a solution.
- Minimum Correlation R-Value – the minimum Pearson correlation coefficient that will be used in the assignment of a bootstrap run factor to a base run factor. The default value is 0.6. If a large number of factors are unmapped, the user may want to lower the R-value. This change should be reported with the final solution.
- Seed – similar to base runs, the number used in a pseudo-random number generator to generate the starting point for each iteration performed by ME-2. The default seed is “random.”
- Block Size – the number of samples that will be selected in each step of resampling. For example, a block size of three means that, for each sample chosen for a bootstrap data set, three samples will be selected for the bootstrap data set. Blocks are non-overlapping. The default block size is calculated according to Politis and White (2003) but can be overridden by the user. If the default has been overridden, the user can press the “Suggest Block Size” button to restore the default value.

After all input parameters are entered, the bootstrap runs are initiated by pressing the “Run” button inside the **Bootstrap Model Runs** box. As with the base runs, the user can interrupt the runs by pressing the “Stop” button in the lower right corner of the **Model Execution** screen. No outputs will be saved if the run is interrupted.



Base Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged
1	5661.5	5979.3	Yes
2	5661.8	5979.5	Yes
3	5661.5	5979.2	Yes
4	5661.5	5979.2	Yes
5	5661.5	5979.2	Yes
6	5661.5	5979.4	Yes
7	5661.6	5979.2	Yes
8	5661.5	5979.1	Yes
9	5661.6	5979.4	Yes
10	5661.7	5979.0	Yes
11	5661.4	5979.2	Yes
12	5661.5	5979.1	Yes

Fpeak Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged

Base Model Runs

Number of Runs: 20
Seed: random
Number of Factors: 8
Output File Prefix: example

Bootstrap Model Runs

Selected Base Run: 11
Seed: rand
Number of Bootstraps: 100
Minimum Correlation R-Value: 0.6
Block Size: 20

Fpeak Model Runs

Selected Fpeak Runs:
Selected Base Run: 11

☑ Strength of Fpeak 1: 0.1
☑ Strength of Fpeak 2: .4
☑ Strength of Fpeak 3: 0.3
☑ Strength of Fpeak 4: .2
☑ Strength of Fpeak 5: 0.5

Run Progress

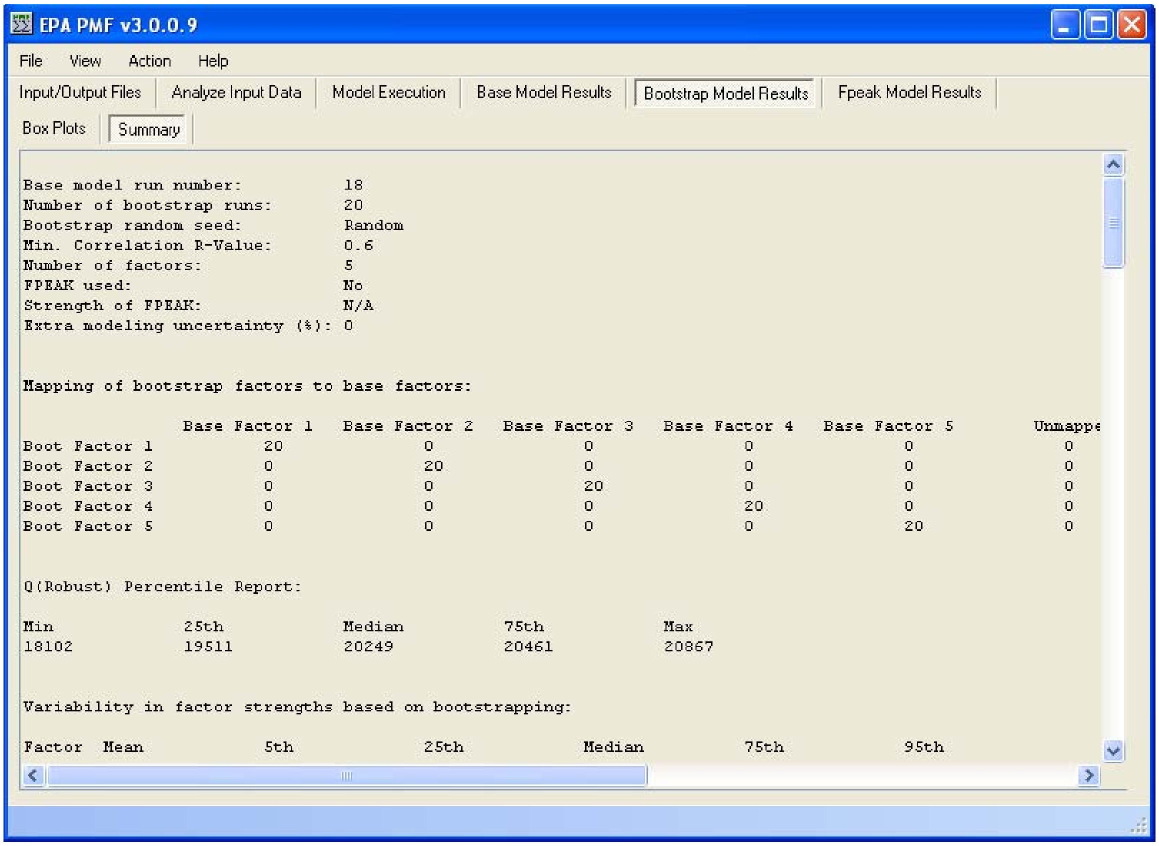
HAVE Concentration Data | HAVE Uncertainty Data | HAVE Base Results | NEED Bootstrap Results

Figure 6-15.—Example **Model Execution** screen, highlighting **Bootstrap Model Runs**.

6.4.2 Summary of Bootstrap Runs

Bootstrapping results are displayed in the **Bootstrap Model Results** screen in the **Box Plots** and **Summary** sub-screens. The first eight lines in the **Summary** screen (Figure 20) contain all the input parameters for bootstrapping, as specified by the user in the **Model Execution** screen. The **Summary** screen also includes several tables that summarize the bootstrap runs. The first table is a matrix of how many bootstrap factors were matched to each base factor. The next table shows the minimum, maximum, median, and 25th and 75th percentiles of the Q(robust) values. The variability in factor strengths is given as the mean, 5th percentile, 25th percentile, median, 75th percentile, and 95th percentile of factor strengths. The rest of the summary is the variability in each factor profile, also given as the mean, 5th percentile, 25th percentile, median, 75th percentile, and 95th percentiles. The base run of each profile is included as the first column for reference, as is a column indicating if the base run profile is within the inter-quartile range of the bootstrap run profiles.

EPA PMF also calculates the Discrete Difference Percentiles (DDP) associated with the bootstrap runs and reports these values in the **Summary** screen. This method estimates the 90th and 95th percentile confidence intervals around the base run profile, reported as percentages. The DDP is calculated by taking the 90th and 95th percentile of the absolute differences between the base run and the bootstrap runs for each species in each profile and expressing it as a percentage of the base run value. If the DDP percent is greater than 999, a “+” is displayed on screen. The original value is saved in the output files. If the base run value for a species is zero, it is not possible to calculate the DDP; in these cases, an asterisk, “*”, is displayed.



```
Base model run number:          18
Number of bootstrap runs:       20
Bootstrap random seed:          Random
Min. Correlation R-Value:       0.6
Number of factors:              5
FPEAK used:                     No
Strength of FPEAK:              N/A
Extra modeling uncertainty (%): 0


Mapping of bootstrap factors to base factors:

                Base Factor 1   Base Factor 2   Base Factor 3   Base Factor 4   Base Factor 5      Unmappe
Boot Factor 1        20              0               0               0               0               0
Boot Factor 2        0               20              0               0               0               0
Boot Factor 3        0               0               20              0               0               0
Boot Factor 4        0               0               0               20              0               0
Boot Factor 5        0               0               0               0               20              0


Q(Robust) Percentile Report:

Min             25th            Median          75th            Max
18102           19511           20249           20461           20867


Variability in factor strengths based on bootstrapping:

Factor  Mean            5th             25th            Median          75th            95th
```

Figure 6-16.—Example of bootstrap **Summary** screen.

6.4.3 Bootstrap Results

The variability in bootstrap runs is shown graphically in the **Box Plots** screen (Figure 21). Two graphs are presented here: the variability in the percentage of each species (Figure 21, 1) and the variability in the concentration of each species (Figure 21, 2), which corresponds to the Variability in Factor Profiles table in the **Summary** screen. In both box plots, the box shows the interquartile range (25th–75th percentile) of the bootstrap runs. The horizontal green line represents the median bootstrap run and the red crosses represent values outside the interquartile range. The base run is shown as a blue box for reference. Values outside of the interquartile range are shown as red crosses . At the bottom of this screen, the base run numbers are grayed out and not selectable; however, the base run used for bootstrapping is highlighted in orange. The user can select the factor they want to view by clicking on the factor number across the bottom of the screen. Selecting "U" displays the summary of unmapped factors. These graphs are left blank if there are no unmapped factors.

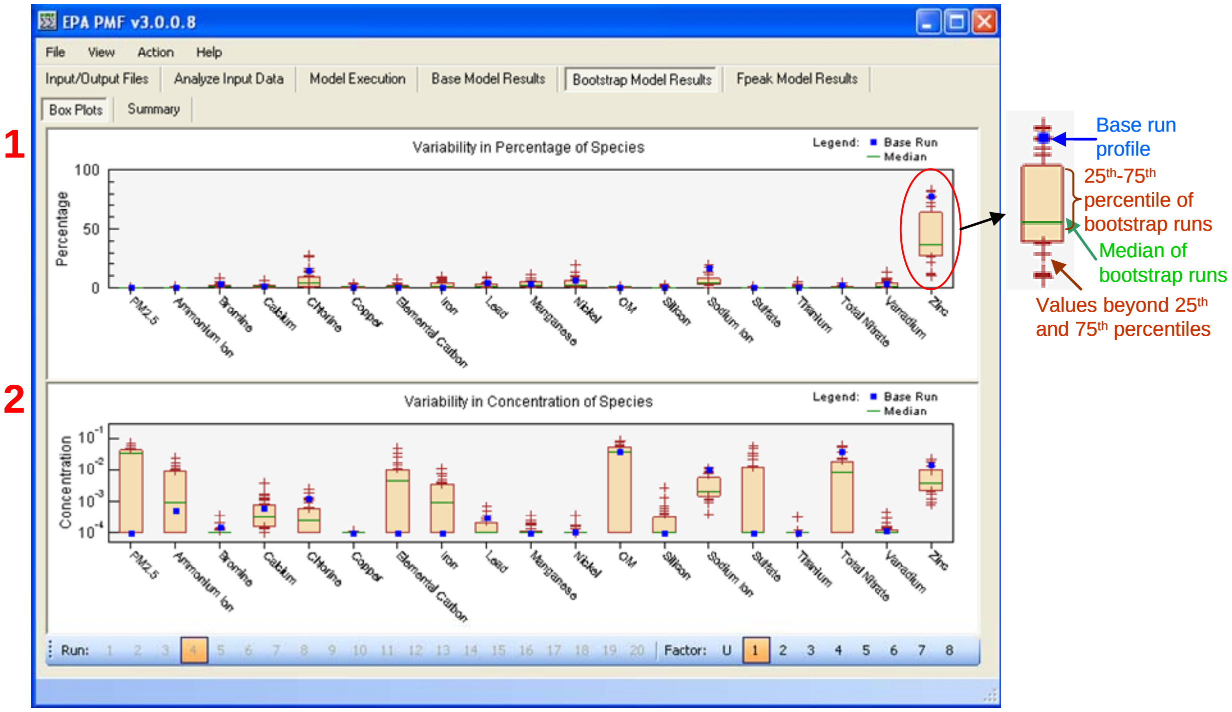

EPA PMF v3.0.0.8
File View Action Help
Input/Output Files | Analyze Input Data | Model Execution | Base Model Results | Bootstrap Model Results | Fpeak Model Results
Box Plots | Summary
1
Variability in Percentage of Species
Legend: ■ Base Run — Median
Percentage
100
50
0
PM2.5 Ammonium Ion Bromine Calcium Chlorine Copper Elemental Carbon Iron Lead Manganese Nickel OM Silicon Sodium Ion Sulfate Titanium Total Nitrate Vanadium Zinc
2
Variability in Concentration of Species
Legend: ■ Base Run — Median
Concentration
10^{-1}
10^{-2}
10^{-3}
10^{-4}
PM2.5 Ammonium Ion Bromine Calcium Chlorine Copper Elemental Carbon Iron Lead Manganese Nickel OM Silicon Sodium Ion Sulfate Titanium Total Nitrate Vanadium Zinc
Run: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 | Factor: U 1 2 3 4 5 6 7 8
Base run profile
25th-75th percentile of bootstrap runs
Median of bootstrap runs
Values beyond 25th and 75th percentiles

Figure 6-17.—Example of bootstrap **Box Plots** screen.

6.5 Fpeak Runs

A pair of factor matrices (*G* and *F*) that can be transformed to another pair of matrices (*G** and *F**) with the same Q-value is said to be "rotated". The transformation takes place as follows:

$$G^* = GT \text{ and } F^* = T^{-1}F \tag{6-4}$$

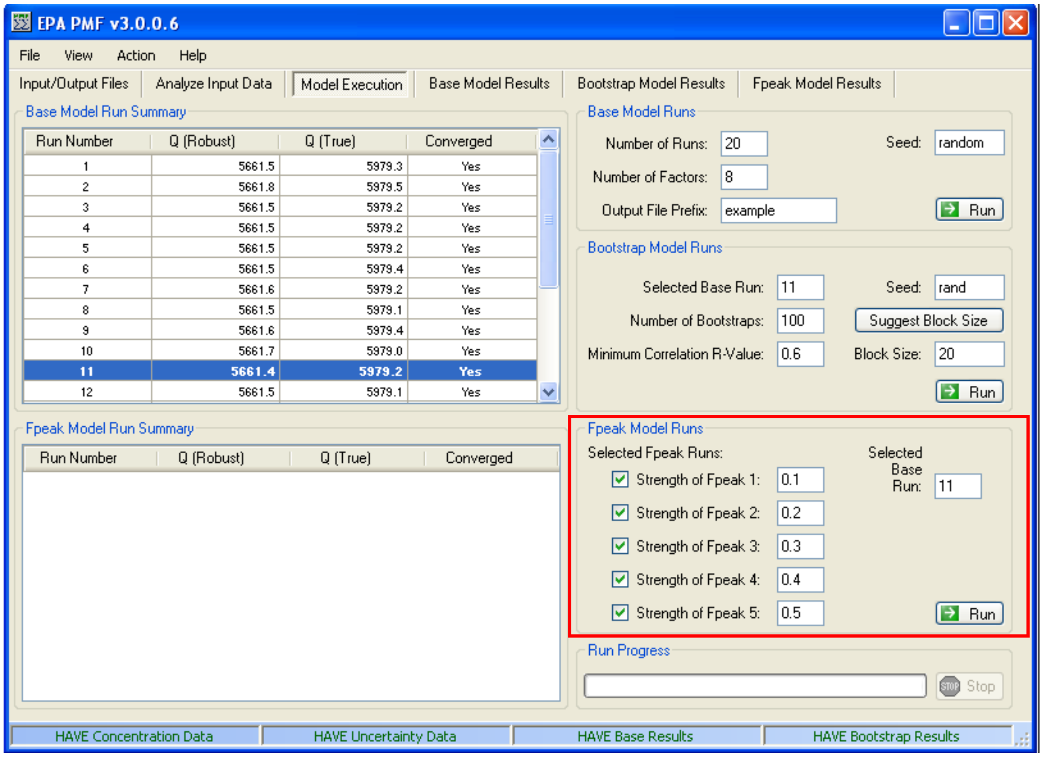
The *T* matrix is a *p* x *p*, non-singular matrix. In PMF, this is not strictly a rotation but rather a linear transformation of the *G* and *F* matrices. Due to the non-negativity constraints in PMF, a rotation (i.e., a specific T matrix) is only possible if none of the elements of the new matrices are less than zero. If no rotation is possible, the solution is unique.

For some solutions, the non-negativity constraint is enough to ensure that there is little rotational ambiguity in a solution. If there are a sufficient number of 0 values in the profiles (F-matrix) and contributions (G-matrix) of a solution, the solution will not rotate away from the "real" solution. However, in many cases, the non-negativity constraint is not sufficient to prevent rotation away from the "real" solution. To help determine if an incorrect rotation has occurred, the user should inspect the G-space plots (see Figure 6-13) for each pair of factors in the original solution. An improperly rotated solution will have oblique edges that do not correspond to the axes (see red line in Figure 6-12). It is not necessary for all G-space plots to have these edges for a solution to be rotated, i.e., only some factors may be incorrectly rotated in the solution. In these cases, to accurately use the results of the model, the solution must be rotated back to the real solution. In EPA PMF, using Fpeak makes this possible.

Before using Fpeak, the user should perform multiple base runs with no rotational forcing and choose one run as a starting point. Using this base-case run, the user should use several values of Fpeak to evaluate different rotations. It should be noted that a Fpeak rotation is not always required.

6.5.1 Initiating Fpeak Runs

In EPA PMF, Fpeak runs are initiated on the **Model Execution** screen in **Fpeak Model Runs** (Figure 22, red square). The Base Model Run with the lowest Q(robust) is automatically selected by the program as the base run for Fpeak runs; this can be overwritten by the user in the Selected Base Run box. The user can perform up to five Fpeak runs by checking the appropriate number of boxes and entering the desired strength of each Fpeak run. While there are no limits on the values that can be entered as Fpeak strengths, generally values between -5 and 5 should be explored first. Positive Fpeak values sharpen the F matrix and smear the G matrix and negative Fpeak values smear the F matrix and sharpen the G matrix. More details on positive and negative Fpeak values can be found in the Paatero, 2000 reference document. The Fpeak strengths in ME-2 are not the same as those in PMF2; values of around 5 times the PMF2 values are needed to produce comparable results in ME-2. Additionally, an Fpeak value of 0 is not allowed; EPA PMF will give the user an error message if 0 is entered in any Fpeak strength box. Fpeak runs begin when the user presses the “Run” button in **Fpeak Model Runs**. Base run and bootstrap run results will not be lost when Fpeak is run.



Base Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged
1	5661.5	5979.3	Yes
2	5661.8	5979.5	Yes
3	5661.5	5979.2	Yes
4	5661.5	5979.2	Yes
5	5661.5	5979.2	Yes
6	5661.5	5979.4	Yes
7	5661.6	5979.2	Yes
8	5661.5	5979.1	Yes
9	5661.6	5979.4	Yes
10	5661.7	5979.0	Yes
11	5661.4	5979.2	Yes
12	5661.5	5979.1	Yes

Fpeak Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged

Base Model Runs

Number of Runs: 20
Seed: random
Number of Factors: 8
Output File Prefix: example

Bootstrap Model Runs

Selected Base Run: 11
Seed: rand
Number of Bootstraps: 100
Minimum Correlation R-Value: 0.6
Block Size: 20

Fpeak Model Runs

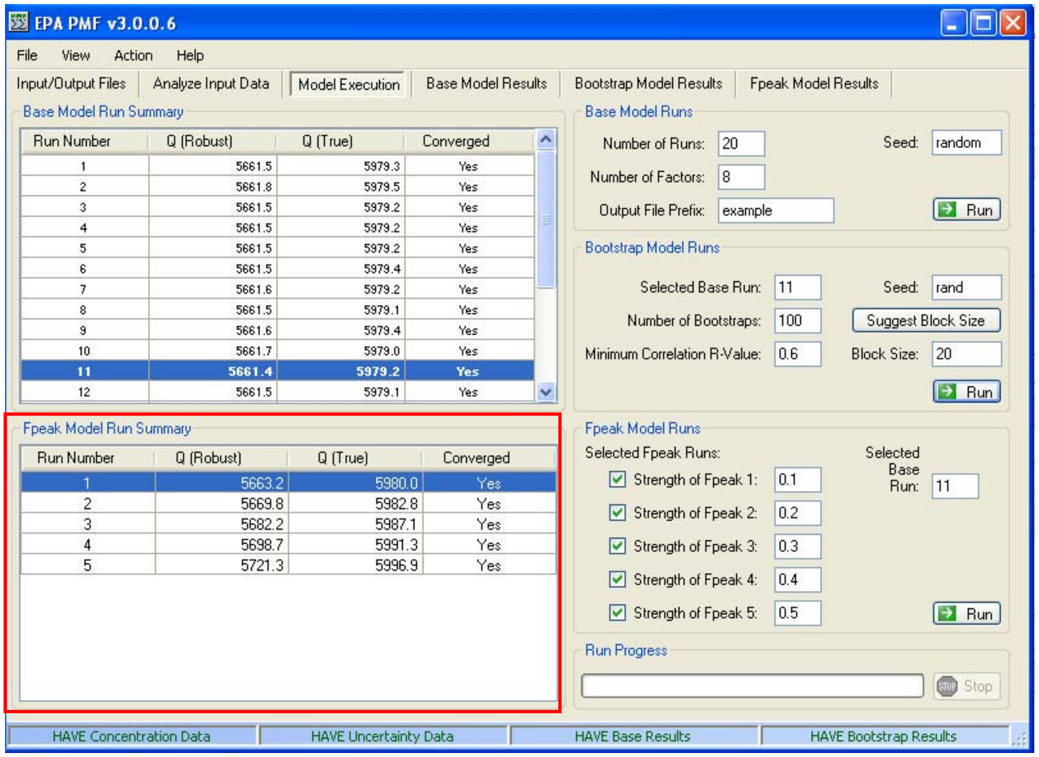
Selected Fpeak Runs:
☑ Strength of Fpeak 1: 0.1
☑ Strength of Fpeak 2: 0.2
☑ Strength of Fpeak 3: 0.3
☑ Strength of Fpeak 4: 0.4
☑ Strength of Fpeak 5: 0.5
Selected Base Run: 11

Run Progress

Figure 6-18.—Fpeak Model Runs highlighted in the **Model Execution** screen.

6.5.2 Fpeak Results

A summary of the Fpeak results, with the same information as the **Base Model Run Summary** table, is shown in the **Fpeak Model Run Summary** table (Figure 23).



Base Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged
1	5661.5	5979.3	Yes
2	5661.8	5979.5	Yes
3	5661.5	5979.2	Yes
4	5661.5	5979.2	Yes
5	5661.5	5979.2	Yes
6	5661.5	5979.4	Yes
7	5661.6	5979.2	Yes
8	5661.5	5979.1	Yes
9	5661.6	5979.4	Yes
10	5661.7	5979.0	Yes
11	5661.4	5979.2	Yes
12	5661.5	5979.1	Yes

Base Model Runs

Number of Runs: 20
Seed: random
Number of Factors: 8
Output File Prefix: example

Bootstrap Model Runs

Selected Base Run: 11
Seed: rand
Number of Bootstraps: 100
Minimum Correlation R-Value: 0.6
Block Size: 20

Fpeak Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged
1	5663.2	5980.0	Yes
2	5669.8	5982.8	Yes
3	5682.2	5987.1	Yes
4	5698.7	5991.3	Yes
5	5721.3	5996.9	Yes

Fpeak Model Runs

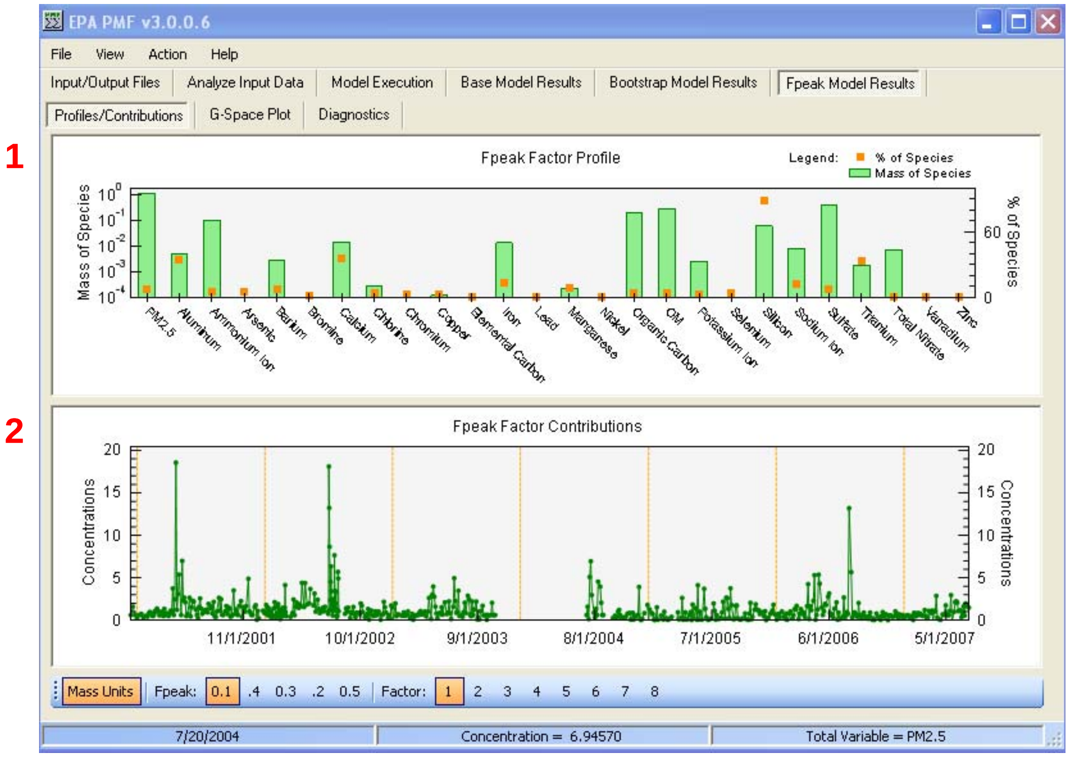
Selected Fpeak Runs:
☑ Strength of Fpeak 1: 0.1
☑ Strength of Fpeak 2: 0.2
☑ Strength of Fpeak 3: 0.3
☑ Strength of Fpeak 4: 0.4
☑ Strength of Fpeak 5: 0.5
Selected Base Run: 11

Run Progress

Figure 6-19.—Fpeak Model Run Summary highlighted in the **Model Execution** screen.

The results of the Fpeak runs are displayed in the **Fpeak Model Results** screen. There are three sub-screens: **Profiles/Contributions** (Figure 24), **G-Space Plots** (Figure 25), and **Diagnostics**. These screens correspond to the names of the sub-screens in the **Base Model Results** screen, which should be used as a reference when evaluating the Fpeak runs.

The **Profiles/Contributions** sub-screen presents profile (Figure 24, 1) and contribution (Figure 24, 2) plots by Fpeak value and factor. In the profile graph, the mass of species (left Y axis) is a green bar and the percent of species (right Y axis) is an orange box. The Fpeak values are in the same order as entered on the **Model Execution** screen. The factors are in the same order as those in **Base Model Results** (see Figure 17, 1). These graphs should be compared among Fpeak values and with the corresponding **Base Model** G-space plot (see Figure 17, 2) to look for deviations (i.e., increases or decreases in a particular species in a factor). The user can select an Fpeak value and factor number by clicking on the desired number at the bottom of the screen.

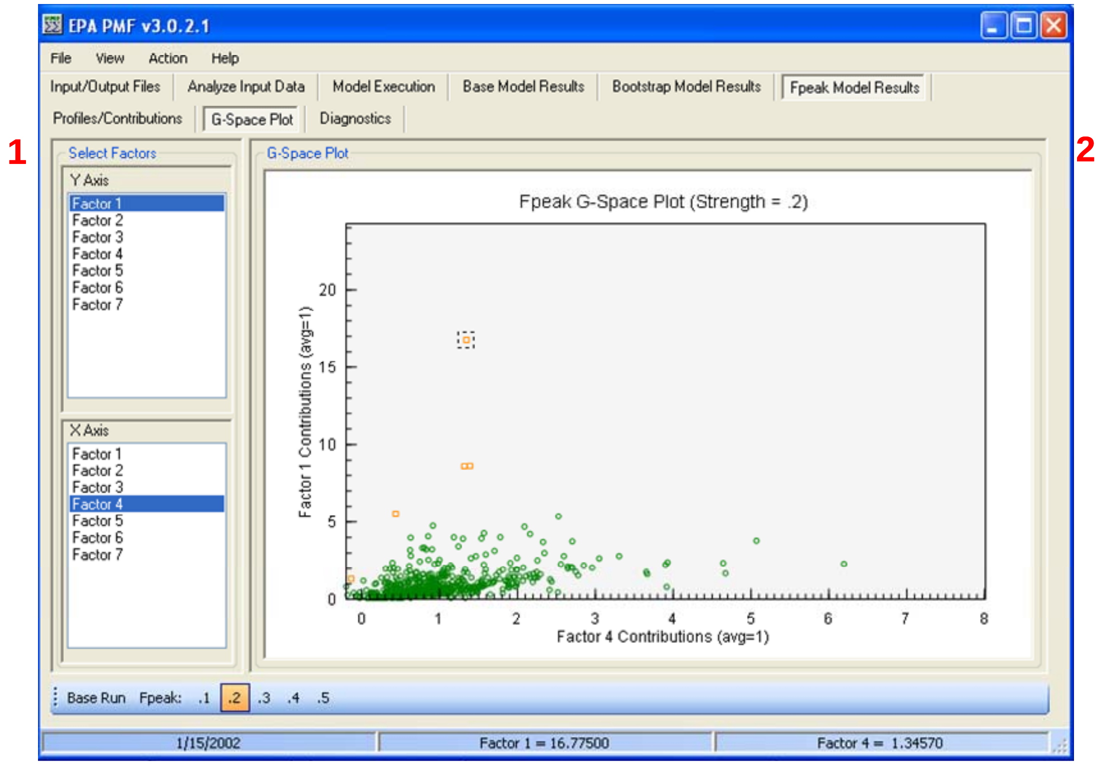


7/20/2004 | Concentration = 6.94570 | Total Variable = PM2.5

Figure 6-20.—Example of Fpeak **Profiles/Contributions** screen.

The status bar in the **Profiles/Contributions** sub-screen displays the date and contribution of data points closest to the mouse position on the contribution graph.

As in the **Base Model Results** screen, the **G-Space Plot** graphic in the **Fpeak Model Results** screen is a scatter plot of factors. The user assigns a factor to the X and Y axes by selecting the desired factor from the lists on the left of the screen (Figure 25, 1). The Fpeak value to display (or the base run) is selected at the bottom of the screen. Once an Fpeak value is selected in either the **Profiles/Contributions** sub-screen or the **G-Space Plot** sub screen, it is automatically selected in both screens. The user can also select points in any G-space plot by clicking on that point. The point selected will turn orange and the date and x and y values will be stored to the ***_Fpeak_diag** file. This feature helps the user identify and track rotations. For example, if a G-space plot appears rotated, the user can mark the edge points. Using *a priori* information, such as meteorological conditions or emissions information, the user can determine if these edge points should be 0 (i.e., the contribution from that factor should be 0 for given samples).



1/15/2002 | Factor 1 = 16.77500 | Factor 4 = 1.34570

Figure 6-21.—Example **G-Space Plot** sub-screen in **Fpeak Model Results**.

The status bar on the **G-Space Plot** sub-screen displays the date, x-value, and y-value of data points closest to the mouse position.

The **Diagnostics** screen summarizes the Fpeak input parameters and output for reference. All of the information on this screen is saved in ***_Fpeak_diag**.

6.5.3 Evaluating Fpeak Runs

Fpeak runs should be viewed by the user as a means of exploring the full space of the chosen PMF solution. Several aspects of the solution should be evaluated to understand how Fpeak changes the PMF solution. The user should first examine the Q values of the Fpeak runs (available in the **Fpeak Model Run Summary** on the **Model Execution** screen) to evaluate the increase from the base run Q value. In a pure rotation, the Q value would not change because the rotation is simply a linear transformation of the original solution. However, due to the non-negativity constraints of PMF, pure rotations are not usually possible and the rotations induced by Fpeak are approximate rotations, which do change the Q value. In general, change in the Q value due to Fpeak rotations by a factor of 10, for small data sets or by a factor of 100 for large data sets can be viewed as acceptable. As discussed in Section 6.3.3, G-space plots of the base run can be used to identify possible rotations in the solution. Corresponding G-Space plots of Fpeak solution factors should be examined to see if any edges viewed in the base runs are more or less evident in the Fpeak runs (Figure 26 is an example of a G-space plot with no edges). Additionally, profiles and contributions should be examined for species/samples that deviate from the base run to ensure that they are reasonable.

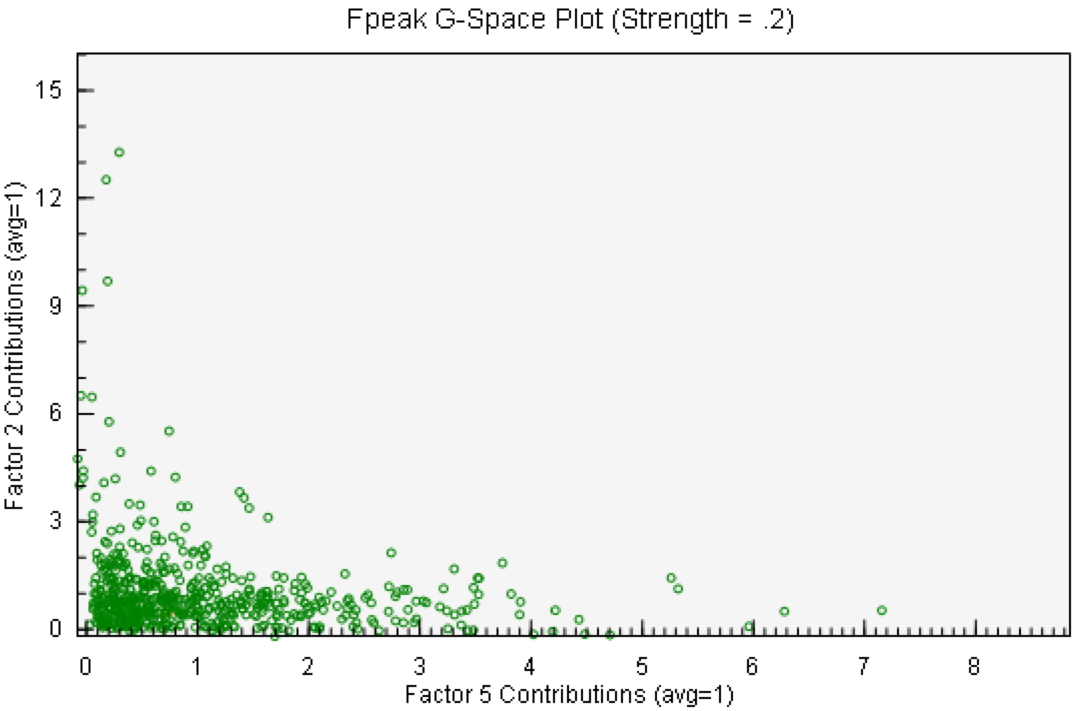


Figure 6-22.—Example of G-Space plot illustrating independence between factors.

7.0 TROUBLESHOOTING

Common problems in EPA PMF v3.0, including the error message generated by the GUI and the action the user should take to correct the problem, are detailed in Table 7-1. If a problem cannot be resolved using the following information, send an email to NERL_RM_Support@epa.gov.

Table 7-1. *Common problems in EPA PMF v3.0.*

Problem	Error Message	Action
Cannot run base runs	Access to the path 'C:\Program Files\EPA PMF 3.0\PMFData.txt' is denied. Please close all output files.	Turn off User Access Controls in Microsoft Vista
Column headers of concentration and uncertainty files do not match	Species names in uncertainty file do not match those in concentration file. Do you wish to continue?	If the names are correct, continue. If the columns are in a different order, correct and retry.
Number of columns in concentration file is not the same as in uncertainty file	Number of species in uncertainty file do not match the number of species in concentration file.	Select "OK", examine input files. The same number of columns, in the same order, should be included in the concentration and uncertainty files. If named ranges are used, check that the ranges are defined correctly.
Number of rows in concentration file is not the same as in uncertainty file	Dates/times in uncertainty file do not match those in concentration file.	Select "OK", examine input files. The same number of rows, sorted by the date/time should be included in the concentration and uncertainty files. If named ranges are used, check that the ranges are defined correctly.
Blank cells are included in concentration file	Null concentration values are not permitted. Please check your data file.	Select "OK", remove blank cells from input file before trying again.
Blank cells, zero values, or negative values are included in uncertainty file	Null, zero, and negative uncertainty values are not permitted. Please check your data file.	Select "OK", remove inappropriate cells from input file before trying again.
Cannot save output files because one is open	The process cannot access the file '*file path and name*' because it is being used by another process. Please close all output files.	Close file and select "Retry" or select "Cancel" to change the file path and name.
Missing key	The multilinear engine (ME2) cannot find an authorization key file (ME2KEY.KEY) in the program folder.	Copy key to same folder as EPA PMF v3.0.exe.
Invalid key	The multilinear engine (ME2) reports that the authorization key file in the program folder is invalid.	Check that the key is correctly named and in the correct location.
Inappropriate user input	*When "run" is pressed, a message box will indicate the input that is incorrect as well as the type of input that would be appropriate. For example:* Random Seed must be the word 'Random' or a 32-bit integer.	Correct the input and try again.

Problem	Error Message	Action
Specified configuration file cannot be found	User specified configuration file not found	Verify path for configuration file and that file exists
No configuration file specified	Please enter or browse to a valid configuration file	Enter the full path of a configuration file (or browse to it) before pressing the "Load Configuration" button
Species already selected for total variable is declared bad	The Total Variable may not be declared as Bad	Unselect the species as total variable (or select another species as the total variable) before declaring it bad
Species that has been declared bad is selected as the total variable	Species declared as Bad may not be used as a Total Variable	Change species categorization to "Strong" or "Weak" before declaring it a total variable.
Cannot print output	EPAPMF-Printing Error	Check printer settings
Too many species are selected for time series graphs	Up to ten species may be displayed. Please remove some selections.	Unselect some species to keep the total number of species selected at or below 10.
Too many samples are excluded via the time series graphs	No more than 50% of the input samples may be excluded.	Select fewer samples to exclude or exclude samples prior to bringing data into EPA PMF
Data for a species is all missing value indicators	All samples for species *X* are missing. This species will be marked as BAD.	If this is correct, no action is necessary. Otherwise, check input files.
Dates/times in input data are not in chronological order	Please note that input concentration dates are not in sort order and so are unsuitable for display in the factor contributions time series graphs.	If species are not in order on purpose, no action is needed but contribution graphs will not be interpretable. Otherwise, sort data before bringing it into EPA PMF.

8.0 REFERENCES

Brown S.G., Wade K.S., and Hafner H.R. (2007) Multivariate receptor modeling workbook. Prepared for the U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC, by Sonoma Technology, Inc., Petaluma, CA, STI-906207.01-3216, August.

Eberly S. (2005) EPA PMF 1.1 user's guide. Prepared by the U.S. Environmental Protection Agency, National Exposure Research Laboratory, Research Triangle Park, NC, June.

Jaeckels J.M., Bae M.-S., and Schauer J.J. (2007) Positive matrix factorization (PMF) analysis of molecular marker measurements to quantify the sources of organic aerosols. *Environ. Sci. Technol.* **41** (16), 5763-5769.

Kim M., Deshpande S.R., and Crist K., C. (2007) Source apportionment of fine particulate matter ($PM_{2.5}$) at a rural Ohio River Valley site. *Atmos. Environ.* **41**, 9231-9243 (doi: 10.1016/j.atmosenv.2007.07.061).

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. (1997) Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems* **37**, 23-35.

Paatero P. (1999) The multilinear engine - A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Graphical Statistics* **8**, 854-888.

Paatero P. (2000) User’s guide for positive matrix factorization programs PMF2 and PMF3, part 1: tutorial. Prepared by University of Helsinki, Finland, February.

Paatero P. (2000) User’s guide for positive matrix factorization programs PMF2 and PMF3, part 2: reference. Prepared by University of Helsinki, Finland, February.

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

Pekney N.J., Davidson C.I., Robinson A., Zhou L., Hopke P., Eatough D., and Rogge W.F. (2006) Major source categories for $PM_{2.5}$ in Pittsburgh using PMF and UNMIX. *Aerosol Sci. Technol.* **40**, 910-924.

Poirot R.L., Wishinski P.R., Hopke P.K., and Polissar A.V. (2001) Comparative application of multiple receptor methods to identify aerosol sources in northern Vermont. *Environ. Sci. Technol.* **35** (23), 4622-4636.

Politis D.N. and White H. (2003) Automatic block-length selection for the dependent bootstrap. Prepared by the University of California at San Diego, La Jolla, CA, February.

Polissar A.V., Hopke P.K., Paatero P., Malm W.C., and Sisler J.F. (1998) Atmospheric aerosol over Alaska 2. Elemental composition and sources. *J. Geophys. Res.* **103** (15), 19045-19057.

Ramadan Z., Eickhout B., Song X.-H., Buydens L.M.C., and Hopke P., K. (2003) Comparison of Positive Matrix Factorization and Multilinear Engine for the source apportionment of particulate pollutants. *Chemometrics and Intelligent Laboratory Systems* **66**, 15-28.

Reff A., Eberly S.I., and Bhave P.V. (2007) Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods. *J. Air & Waste Manag. Assoc.* **57**, 146-154.

Rizzo M.J. and Scheff P.A. (2007) Utilizing the chemical mass balance and positive matrix factorization models to determine influential species and examine possible rotations in receptor modeling results. *Atmos. Environ.* **41** (33), 6986-6998.

9.0 TRAINING EXERCISES

The following sections offer example PMF analyses of three common types of data sets: $PM_{2.5}$ data from the Speciation Trends Network (STN) and the Interagency Monitoring of Protected Visual Environments (IMPROVE) network, and speciated volatile organic compound (VOC) data from a Photochemical Assessment Monitoring Stations (PAMS) site. The data sets were installed in the EPA PMF v3.0 “Data” folder when EPA PMF v3.0 was installed and are provided as examples for analysis. Users can, on their own, follow the steps outlined in each example to better understand the PMF process and the interaction of the components described in the User’s Guide.

9.1 Baltimore, Maryland, STN $PM_{2.5}$ Data Set

The following sections detail a complete PMF analysis of a $PM_{2.5}$ data set from Baltimore, Maryland. The user should run EPA PMF with the data set provided in **balt_conc.xls and balt_unc.xls** and duplicate the analyses described below. The section headings correspond to the relevant tab in EPA PMF (italicized for reference). This exercise is intended to demonstrate the thought process and steps involved in reaching a solution using EPA PMF; it is not intended to be a complete source apportionment analysis.

9.1.1 Pre-PMF processing/Data set development

Concentration Input File
Prior to use in EPA PMF v3.0, the Baltimore data were downloaded from Air Quality System (AQS) and reformatted in MS Access; each row represents one sample and each column one species. Data below the detection limit (the maximum reported detection limit was used as a conservative limit for all samples) was substituted with one-half of the detection limit and missing data were substituted with the median value. Any samples missing either all of the metals or all of the ions/carbon species were excluded from analysis (30 samples). Missing data groups typically due to failure of one of the samplers so the distribution of $PM_{2.5}$ is unknown. Organic carbon (OC) was adjusted to organic matter (OM) by multiplying all values by 1.4 (White and Roberts, 1977; Turpin and Lim, 2001; Bae et al., 2006; Reff et al., 2007). Additionally, the percent below detection was calculated to guide the user in species categorization. Species with more than 95% of samples below the detection limit were not included in the data set for PMF (see Table 1). Three species were not included in the PMF data set to avoid double counting mass (sodium, potassium, and sulfur are represented by sodium ion, potassium ion, and sulfate, respectively).

Uncertainty Development
Prior to use in EPA PMF v3.0, uncertainties for the Baltimore data set were developed using collocated data (Wade et al., 2008 for more information). Data below detection were given an uncertainty of 5/6 of the detection limit and missing data were given an uncertainty of 4 times the median concentration (Polissar et al., 2001).

Table 9-1. *Percent below detection limit for all species included in STN $PM_{2.5}$ data set for the Baltimore, MD, Essex site. Species highlighted in yellow were not included in the PMF data set because more than 95% of samples were below detection; species highlighted in blue were not included because they were not sampled for the entire time period; species highlighted in green were not included because they are represented by other species.*

Parameter	Percent Below Detection
Aluminum	82%
Ammonium Ion	0%
Antimony	98%
Arsenic	95%
Barium	99%
Bromine	39%
Cadmium	98%
Calcium	2%
Cerium	100%
Cesium	100%
Chlorine	73%
Chromium	93%
Cobalt	100%
Copper	46%
Elemental Carbon	7%
Europium	99%
Gallium	100%
Gold	100%
Hafnium	100%
Indium	99%
Iridium	100%
Iron	
Lanthanum	100%
Lead	84%
Magnesium	97%
Manganese	73%
Mercury	97%
Molybdenum	99%
Nickel	66%
Niobium	100%
Ocx Carbon	
Ocx2 Carbon	
Organic Carbon	
Phosphorus	99%
Pk1_Oc Stn	3%
Pk2_Oc Stn	
Pk3_Oc Stn	1%
Pk4_Oc Stn	5%
PM2.5 mass	
Potassium Ion	42%
Potassium	0%
Pyrolc Stn	89%
Rubidium	100%
Samarium	98%
Scandium	100%
Selenium	90%
Silicon	15%
Silver	100%
Sodium Ion	11%
Sodium	92%
Strontium	98%
Sulfate	
Sulfur	0%
Tantalum	99%
Terbium	96%
Tin	98%
Titanium	56%
Total Nitrate	
Tungsten	100%
Vanadium	67%
Yttrium	100%
Zinc	3%
Zirconium	99%

9.1.2 Analyze Input Data

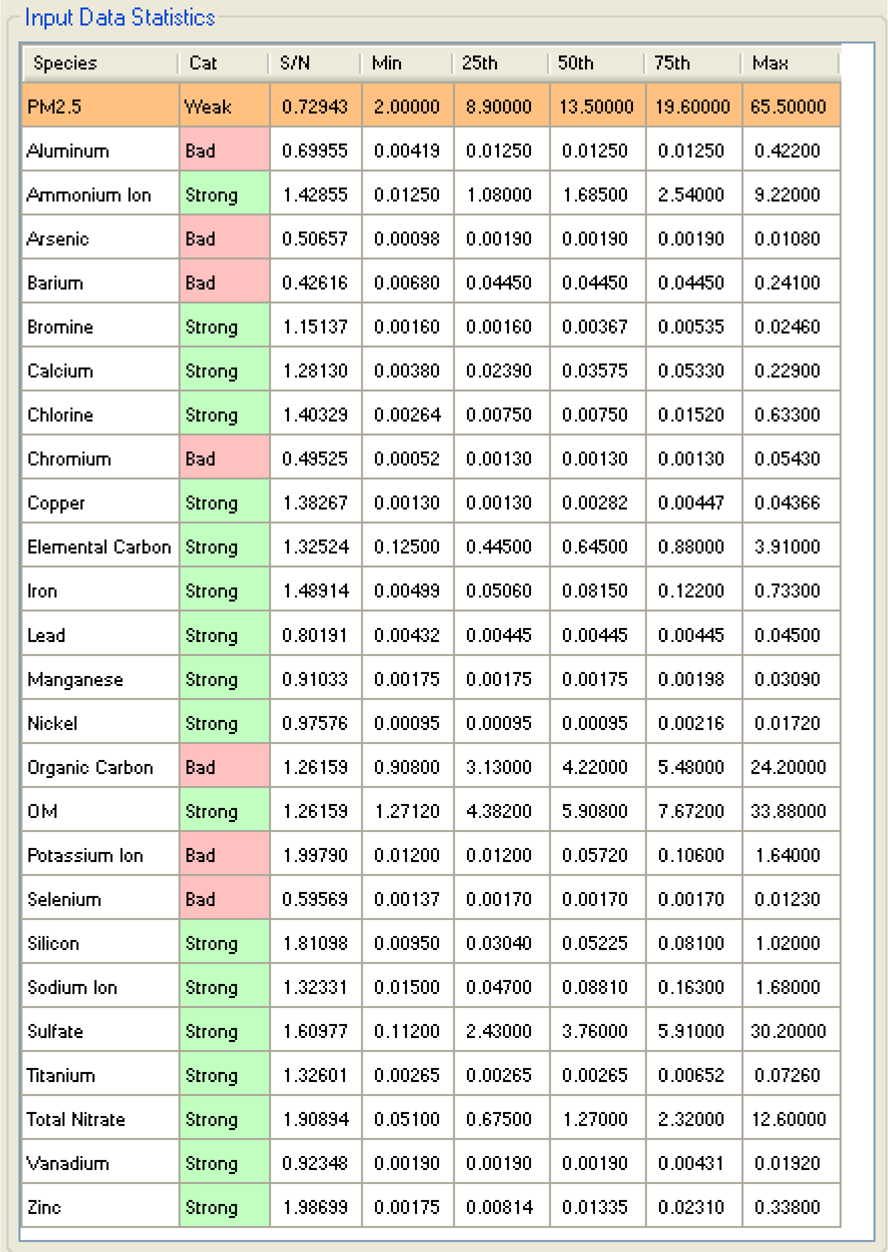
Characterizing Species (*Analyze Input Data: Concentration/Uncertainty and Concentration Time Series*)

The user should first examine the input data to determine if the uncertainties should be increased, by categorizing a species as “weak”, and if any species should be excluded, by categorizing a species as “bad.” There are several reasons to characterize a species as “bad”, for example, a high percentage of data below the detection limit and a low signal-to-noise ratio. Although “high” and “low” are relative terms, generally species with more than 75% of data below detection limit should be examined using a time series plot to determine if the species has a useful signal. Occasionally, *a priori* information is also used to determine if a species should be included in PMF. For example, a species might have a low signal-to-noise ratio but be useful as a tracer for a known local source; in this case, it might be beneficial to include that species in PMF.

Although no species had a signal-to-noise ratio less than 0.2, several species in the Baltimore data set were characterized as “bad” due to a combination of a high percentage of data below detection and a low signal-to-noise ratio: barium, arsenic, chromium, selenium, and aluminum (see the **Concentration/Uncertainty** screen**, Input Data Statistics** table, Figure 27). Time series plots, particularly for all species with a signal-to-noise ratio less than 1, were examined to support this decision (see **Concentration Time Series** screen, example in Figure 28). Because there were no species with a signal-to-noise ratio greater than 2 (indicating the uncertainty estimates are already conservative for this data set), the model was first run with all other species set to “strong”. Model results will be used to guide further categorization of species.

The user can also examine the percentiles provided on the **Concentration/Uncertainty** screen to verify that concentrations are within typical concentration ranges. Extreme high or low values could indicate errors in the data set or extreme events that would not be modeled well by PMF.

Initially, the “Extra Modeling Uncertainty” was left at 0%. Generally, at least 5% is appropriate and sensitivity tests using various values will be performed as part of the base model results.



Input Data Statistics

Species	Cat	S/N	Min	25th	50th	75th	Max
PM2.5	Weak	0.72943	2.00000	8.90000	13.50000	19.60000	65.50000
Aluminum	Bad	0.69955	0.00419	0.01250	0.01250	0.01250	0.42200
Ammonium Ion	Strong	1.42855	0.01250	1.08000	1.68500	2.54000	9.22000
Arsenic	Bad	0.50657	0.00098	0.00190	0.00190	0.00190	0.01080
Barium	Bad	0.42616	0.00680	0.04450	0.04450	0.04450	0.24100
Bromine	Strong	1.15137	0.00160	0.00160	0.00367	0.00535	0.02460
Calcium	Strong	1.28130	0.00380	0.02390	0.03575	0.05330	0.22900
Chlorine	Strong	1.40329	0.00264	0.00750	0.00750	0.01520	0.63300
Chromium	Bad	0.49525	0.00052	0.00130	0.00130	0.00130	0.05430
Copper	Strong	1.38267	0.00130	0.00130	0.00282	0.00447	0.04366
Elemental Carbon	Strong	1.32524	0.12500	0.44500	0.64500	0.88000	3.91000
Iron	Strong	1.48914	0.00499	0.05060	0.08150	0.12200	0.73300
Lead	Strong	0.80191	0.00432	0.00445	0.00445	0.00445	0.04500
Manganese	Strong	0.91033	0.00175	0.00175	0.00175	0.00198	0.03090
Nickel	Strong	0.97576	0.00095	0.00095	0.00095	0.00216	0.01720
Organic Carbon	Bad	1.26159	0.90800	3.13000	4.22000	5.48000	24.20000
OM	Strong	1.26159	1.27120	4.38200	5.90800	7.67200	33.88000
Potassium Ion	Bad	1.99790	0.01200	0.01200	0.05720	0.10600	1.64000
Selenium	Bad	0.59569	0.00137	0.00170	0.00170	0.00170	0.01230
Silicon	Strong	1.81098	0.00950	0.03040	0.05225	0.08100	1.02000
Sodium Ion	Strong	1.32331	0.01500	0.04700	0.08810	0.16300	1.68000
Sulfate	Strong	1.60977	0.11200	2.43000	3.76000	5.91000	30.20000
Titanium	Strong	1.32601	0.00265	0.00265	0.00265	0.00652	0.07260
Total Nitrate	Strong	1.90894	0.05100	0.67500	1.27000	2.32000	12.60000
Vanadium	Strong	0.92348	0.00190	0.00190	0.00190	0.00431	0.01920
Zinc	Strong	1.98699	0.00175	0.00814	0.01335	0.02310	0.33800

Figure 9-1.—Input Data Statistics table for the Baltimore data set within initial categorizations.

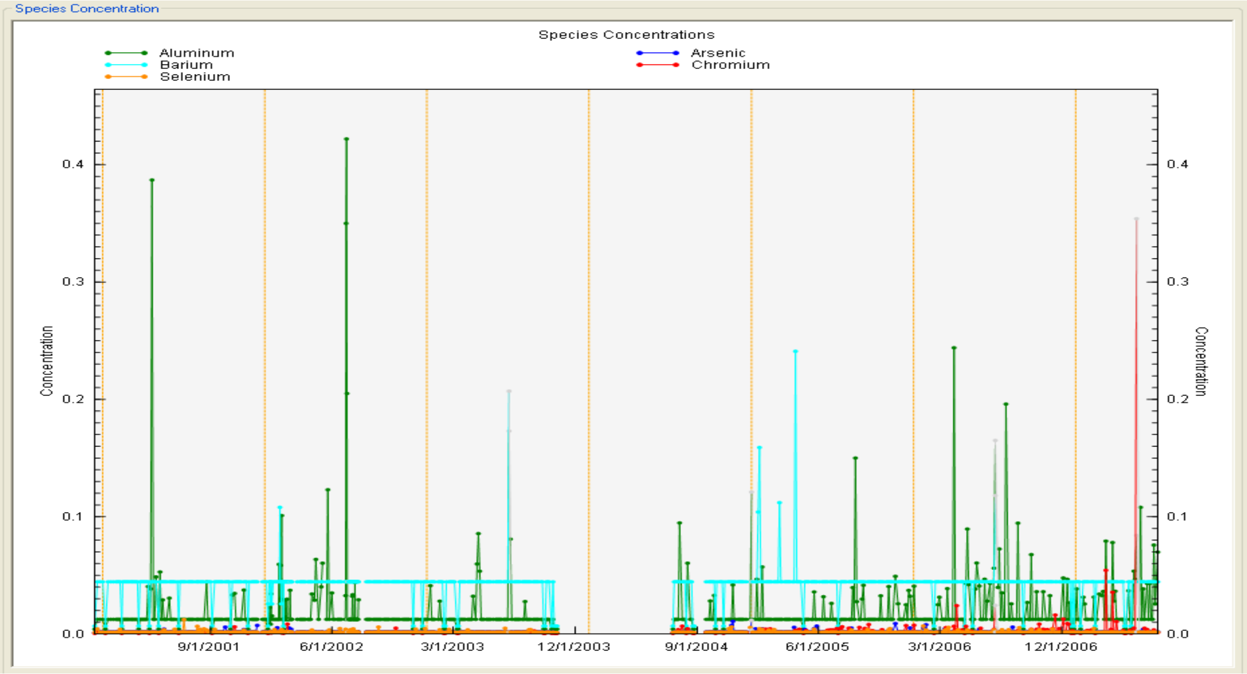


Figure 9-2.—Time series of species with low signal-to-noise ratios, Baltimore data set.

Relationships between Species (*Concentration Scatter Plot*)

Scatter plots between species should be examined for relationships that indicate a common source emitted both species (i.e. Si and Ti for crustal sources). In the Baltimore data set, silicon and calcium are loosely related, indicating a soil source and potentially a second calcium source (Figure 29, top). Iron and manganese are also related, indicating a potential steel source (Figure 29, bottom).

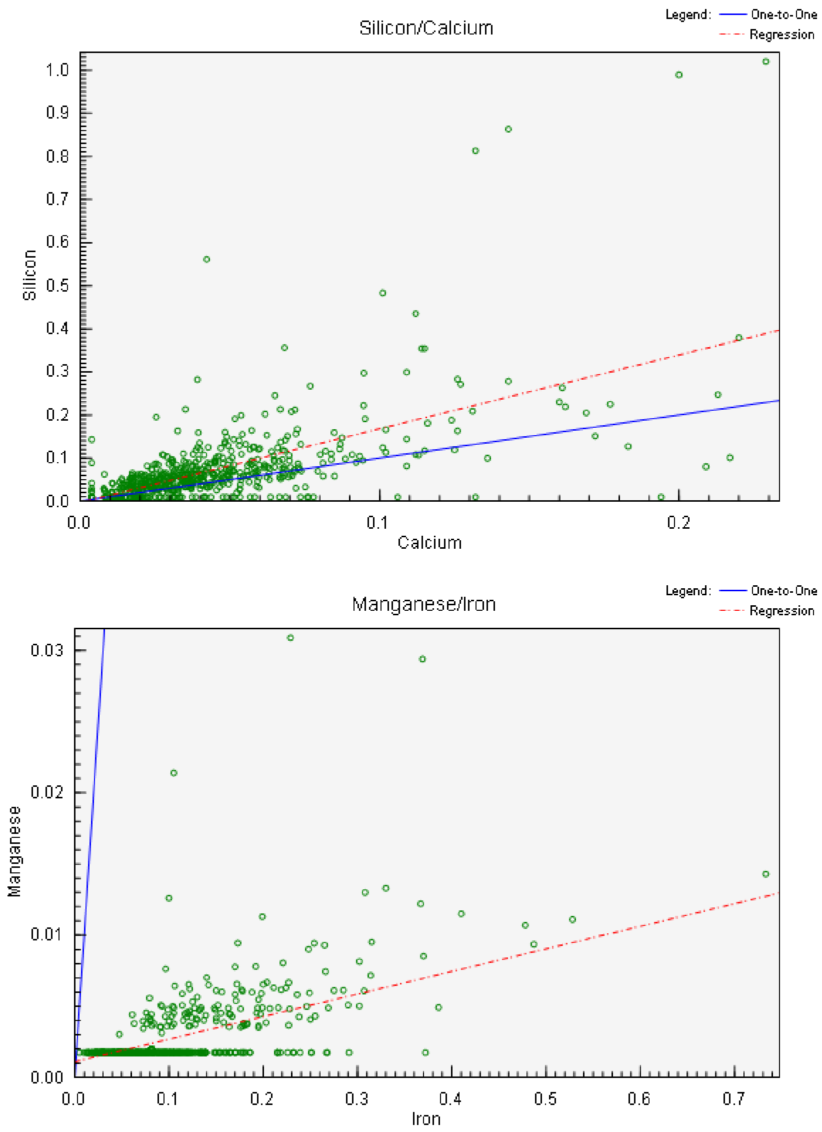


Figure 9-3.—Concentration scatter plots for soil elements (top) and steel elements (bottom).

Excluding Samples (*Concentration Time Series*)

The user should examine the **Concentration Time Series** plots to verify the species selected for PMF have seasonal patterns such as high sulfate during the summer as well as to identify unusual events. Often, these events are easily identified like fireworks on the Fourth of July which contribute to high levels of potassium, strontium, and other trace metals. These identified event samples should be excluded since the overall profiles may not capture the unique composition of the source or the profiles of non-event sources may be distorted. However, all data exclusions must be well justified and documented. The user should be aware that excluding a species will remove the sample from the analysis.

Initially, several samples were excluded due to extreme events, including 5/18/07 for high Fe concentrations (Figure 30, top); 2/28/02 for high chlorine concentrations; and 7/7/02, 7/8/02, 7/5/03, 1/1/05, and 7/4/06 for high potassium concentrations (Figure 30, bottom). The high metal and chlorine events are likely either one-time emissions from an unknown source or analytical errors. The potassium events are due to fireworks around the Fourth of July and New Year's Eve.

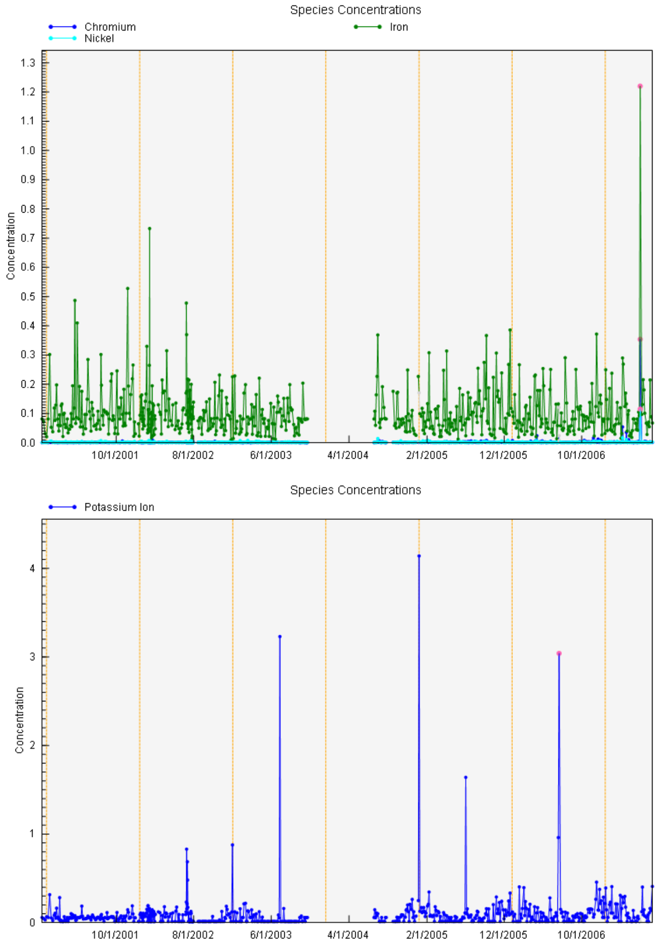


Figure 9-4.—Concentration time series illustrating extreme events for metals (top) and potassium ion (bottom).

9.1.3 Base Runs

Initial Model Parameters (*Model Execution*)
The model was run 20 times with 7 factors and a seed of 25. Enough runs should be performed to determine if the Q-values are stable. A constant seed was used to replicate results for training purposes; in practice, the user should generally use a random seed. All runs converged and the Q-values were very stable, with a range of only 1.1. The Q robust was within 1% of the Q true, indicating outliers are not heavily impacting the Q value. Both were within 50% of the Q theoretical (6778).

9.1.4 Base Run Results

Model Reconstruction (*O/P Scatter Plots, O/P Time Series*)
Examining the observed versus predicted scatter plots and time series, it is obvious that many species were not modeled well (Chlorine, Copper, Lead, Manganese, Nickel, Titanium, Vanadium). This could be due to incorrect uncertainties, improper categorization, too few factors being modeled, or simply that these species are not reliably quantified. Several species with signal-to-noise ratios less than 1 were poorly modeled, including lead, manganese, vanadium, and nickel. Examples of a well-modeled species, ammonium ion, and a poorly modeled species, lead, are presented in Figure 31. The poorly modeled species should be re-categorized as “weak” and the model re-run. Additionally, bromine, copper, and chlorine has too many scaled residuals above 3.0 and below -3.0, indicating that these species were not modeled as well. These species should also be characterized as “weak”. Most of the data for these poorly modeled species are at or below the detection limit; therefore, it is recommended to set them to the “bad” category.

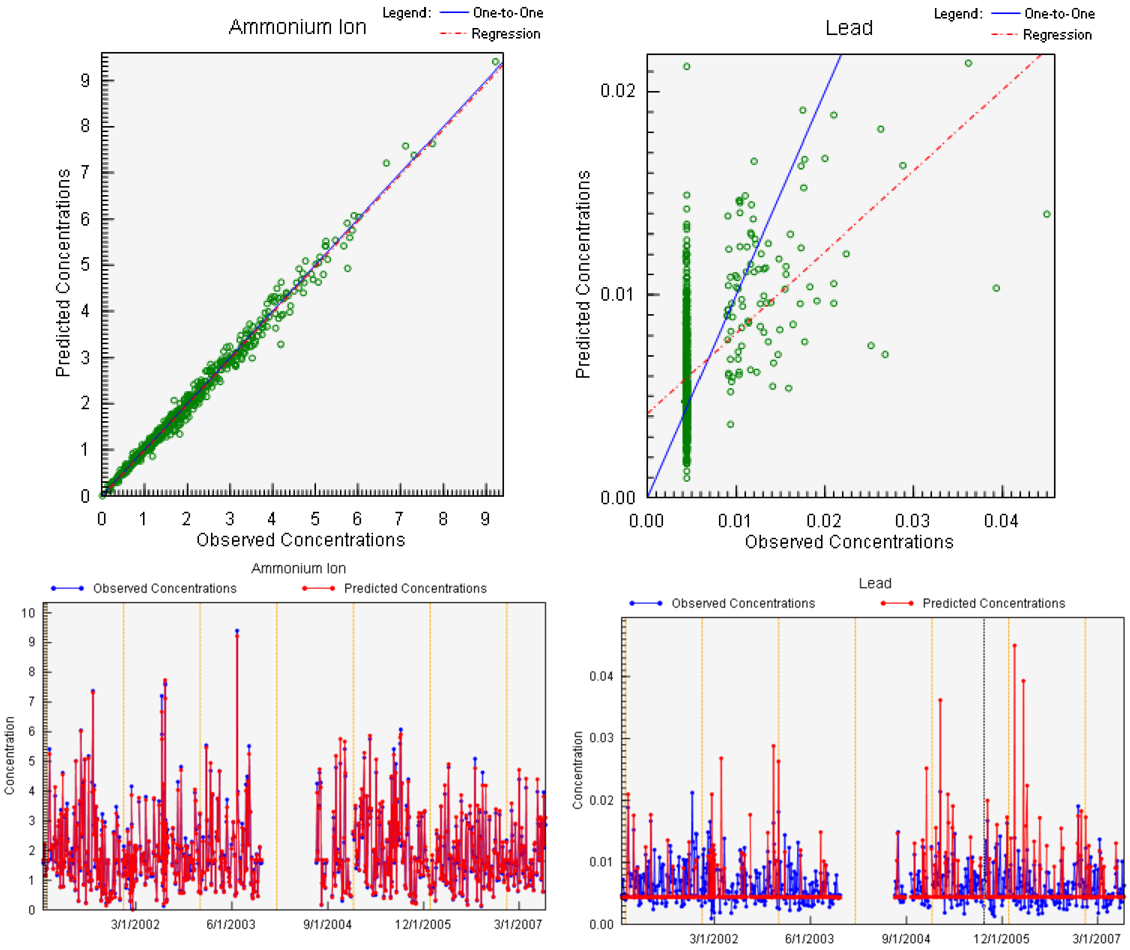


Figure 9-5.—Example output graphs for a well modeled species (ammonium ion, left) and a poorly modeled species (lead, right).

Factor Identification (*Profiles/Contribs, Aggregate Contribs*)
Factors may be identified using dominant species and temporal patterns. These are described in Table 2. Most factors make physical sense for the Baltimore area, except for the sodium ion factor. It is possible the uncertainty for sodium ion is too low. The model should be run with sodium ion characterized as "weak".

Table 9-2. *Identification of factors.*

Factor	Dominant Species	Temporal Pattern	Name
1	Zinc	None	Zinc smelter, Steel
2	Silicon, Calcium	Summer/Fall Peaks	Soil
3	Iron, Manganese	None	Steel
4	OM, EC	High in Winter	OM
5	Nitrate	High in Winter	Secondary Nitrate
6	Sulfate, Ammonium Ion	High in Summer	Secondary Sulfate
7	Sodium Ion	None	Sodium Ion

Rotations (*G-Space Plots*)

G-space plots of the solution should be examined to determine if there are edges. Figure 32 (left) shows a G-space plot with no edges. Figure 32 (right) has an edge that is indicated by a red line. The user should examine all G-space plots for edges. In this data set, several factors have edges that do not align with the axes, so Fpeak should be used to explore the rotational ambiguity of the data set.

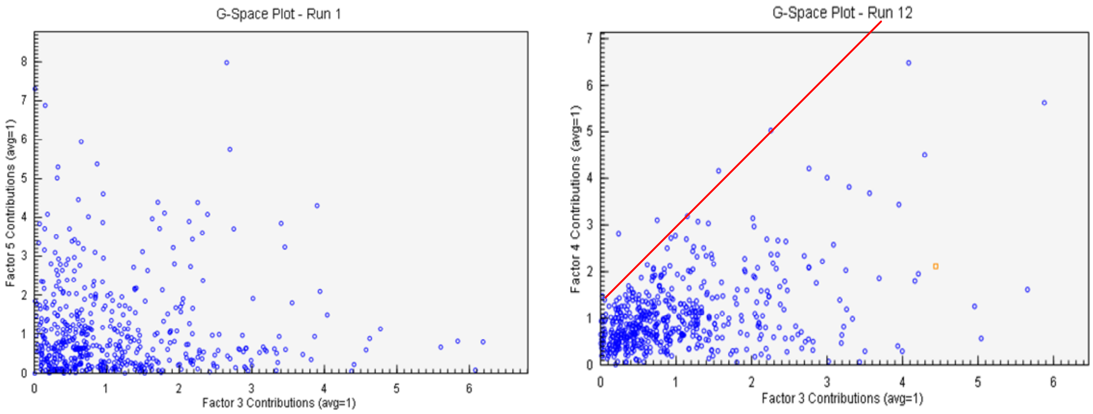


Figure 9-6.—Example of G-space plots for independent (left) and weakly dependent factors (right).

Mass Distribution (*Factor Pie Chart*)

Figure 33 shows the factor pie chart for the total mass variable ($PM_{2.5}$), which should be examined to ensure that the distribution of factors looks realistic. The major factors, secondary sulfate, OM, EC, and secondary nitrate, account for reasonable amounts of mass based on the distribution of ambient data.

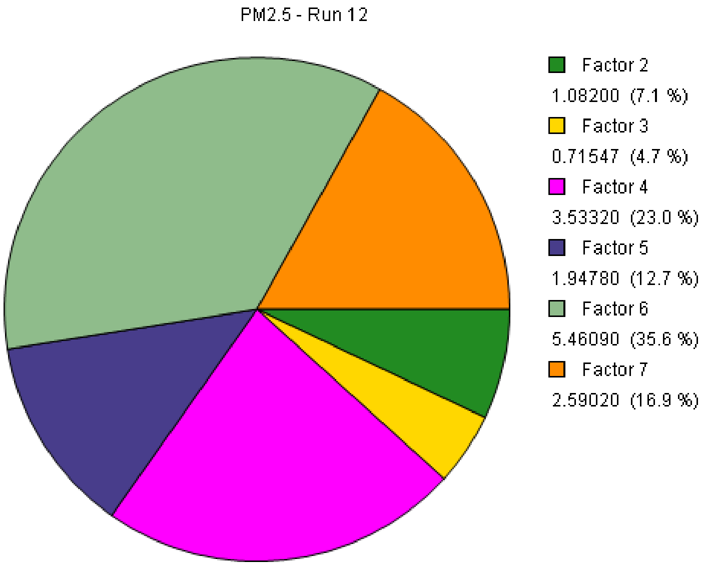


Figure 9-7.—Distribution of mass for total $PM_{2.5}$.

Base Model Runs with Updated Species Categorization

With each iteration, the user should examine all of the outputs/plots described above for effects of changes between runs. For brevity, only those plots that changed appreciably in this iteration will be described below.

In the second iteration, the Q-values are lower due to the increase in uncertainties for the weak species. They are still within an acceptable range and the Q robust is closer to the Q true. When the model is run with sodium ion “weak”, it is no longer its own factor. However, the observed/predicted values for several species are not well-matched and the OM factor has moderate loadings of many species, making it hard to identify. It is possible that too few factors are specified so multiple factors are combining into one factor that is difficult to interpret. An eight-factor solution should be explored next.

The initial eight-factor solution split the secondary sulfate factor into two separate factors. This is not a physically meaningful result. It is therefore likely that a seven-factor solution is a better fit to the data than an eight-factor solution.

As a sensitivity test, both nine-and six-factor solutions were also run. In the nine-factor solution, the soil elements were split into separate calcium and silicon factors. In the six-factor solution, the EC and OM combine into one factor. The steel factor also has excess EC in it.

Additional sensitivity tests were performed with various extra modeling uncertainty values. The Q values decrease slightly with increasing uncertainty, but are more stable. A value of 5% was chosen for the final solution.

9.1.5 Bootstrap Runs

Input Parameters (*Model Execution*)

After selecting a solution, the user should bootstrap that solution to determine if it is stable and will provide consistently similar results. The chosen Baltimore solution was bootstrapped 100 times with a seed of 25 (in order to replicate the results). The suggested block size of 10 and minimum r^2 of 0.6 were used.

Bootstrap Run Results

Output Diagnostics (*Summary*)

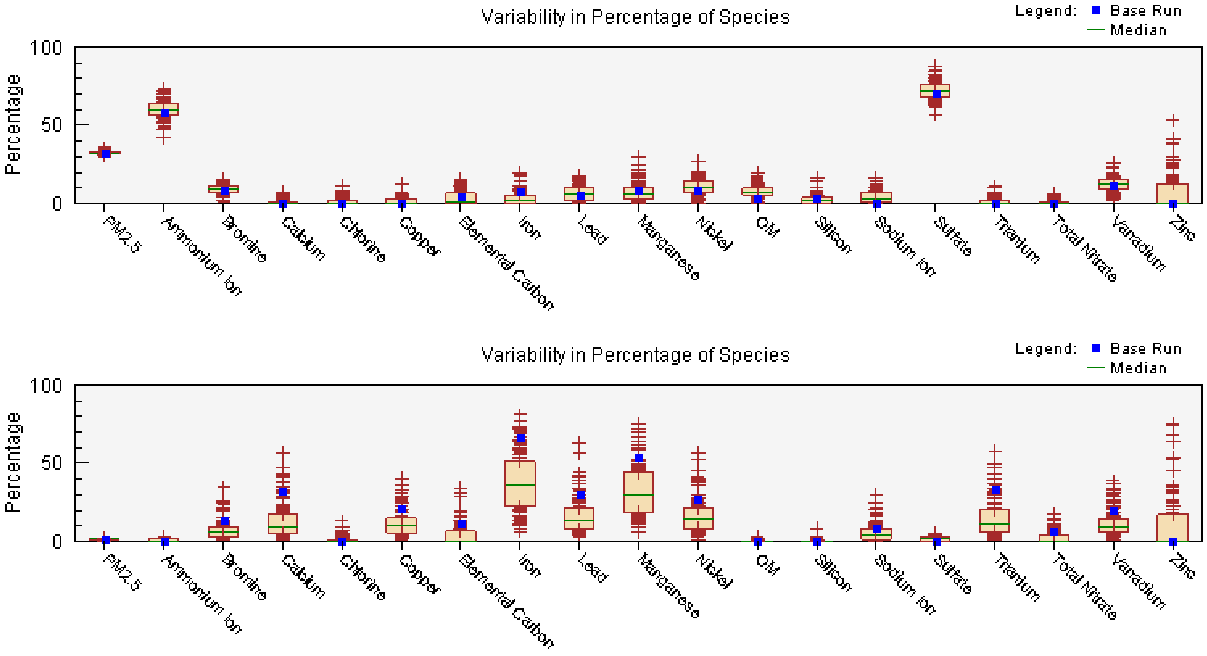
Of the 100 runs, all factors were mapped to a base factor in every run, and no factors were unmapped, indicating a stable result.

Factor Variability (*Box Plots*)

In examining bootstrapping graphics, the user should examine the interquartile range of species within each profile, both in terms of mass and percent. For this data set, the major factors (sulfate, nitrate) had very small interquartile ranges (Figure 34, top); more variability was seen in the industrial factors (Figure 34, bottom), which is not unusual.

The DDPs for this solution (found in the diagnostic file and onscreen on the Bootstrap Model Results/ Summary screen) agree with the bootstrapping box plots. The key species in the major factors have relatively small 90th and 95th percentiles (around 20%-40%). The key species in the industrial factors have larger percentiles, from 80-100%.

It should also be noted that some key species, such as the iron and manganese in the steel factor (Figure 34, bottom) have base run values that are not within in the interquartile range of the bootstrapping results. In this case, the medians of the bootstrapping result should be examined to determine if the factor was correctly identified. For the steel factor, the iron and manganese are still the key species, so the factor identification should not be changed.


Variability in Percentage of Species

Variability in Percentage of Species

Figure 9-8.— Example of bootstrapping profiles for secondary sulfate and steel factors.

9.1.6 Fpeak Runs

Input Parameters (*Model Execution*)

When the solution is rotated using Fpeak, the user should explore a variety of Fpeak values. Several attributes should be noted in determining which solutions are reasonable, including the change in Q values (from the original solution), changes in the profiles/contributions of the original solution, and the G-space plots between factors. A range of Fpeak values, positive and negative, should be used. Positive Fpeak values mainly impact the factor profiles and negative values impact the factor contributions. In this case, Fpeak values of -1.5 to +1.0 (by 0.1) were used with the first base run. Values less than -1.5 provided large increases in the Q value (hundreds of units) and values greater than +1.0 did not converge. A small increase in Q values (by a factor of 10) is acceptable, but larger increases may indicate over-rotation.

9.1.7 Fpeak Run Results

Maximal separation between factors (*G-Space Plots*)

Examining the G-space plots shows that negative values do not improve the edges in the G-space plots. In this example, a value of 0.3 was chosen as it improves the edges between factors somewhat, does not increase the Q value dramatically, and does not increase non-dominant species in the factor profiles. **Figure A-10** shows that the edges between Factors 4 and 7 increased some with this Fpeak value. However, it is important to examine all of the factors to ensure that independence between other factors was not lost.

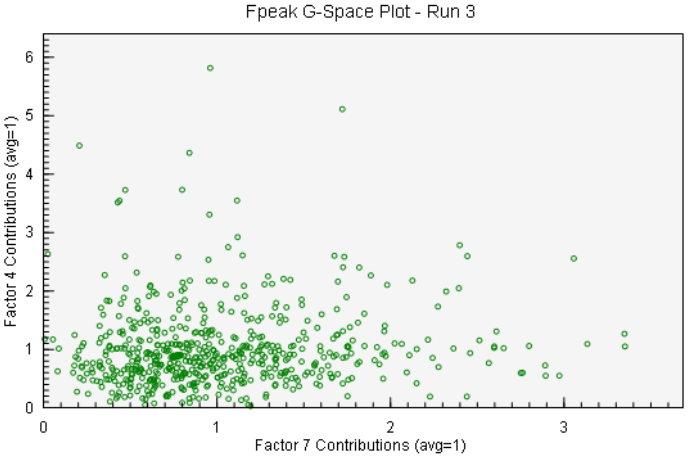


Figure 9-9.—G-space plot for Factor 4 versus Factor 7 with an Fpeak of 0.3.

Rotatated Factors (*Profiles/Contributions*)

Comparing the factor profiles for Fpeak = 0.3 to the base run profiles shows that the key species in the profiles are accentuated in the Fpeak solution. In the example in Figure 36, zinc, which is the key species, goes from about 80% to 100% and all other species decrease. Nitrate, ammonium ion, and the soil species, which were likely driving the overall mass of this factor, are gone or much lower in the Fpeak version. The contributions (Figure 37) did not change as noticeably, as expected with a positive Fpeak value.

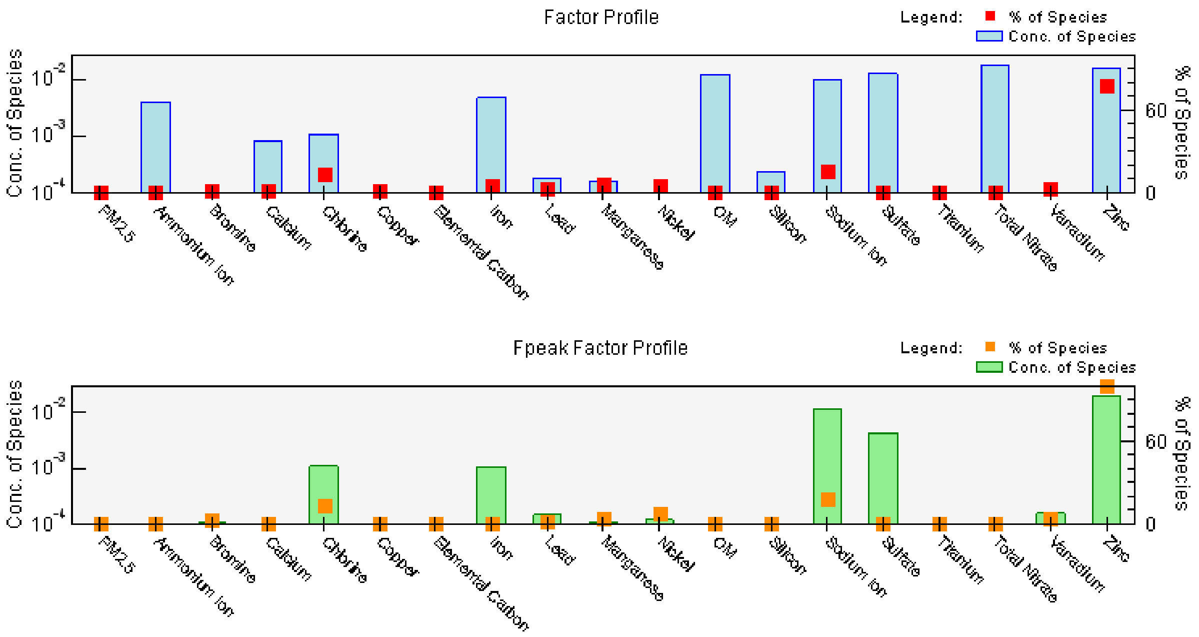


Figure 9-10.—Comparison of base run profile (top) and Fpeak run profile (bottom) for industrial zinc factor.

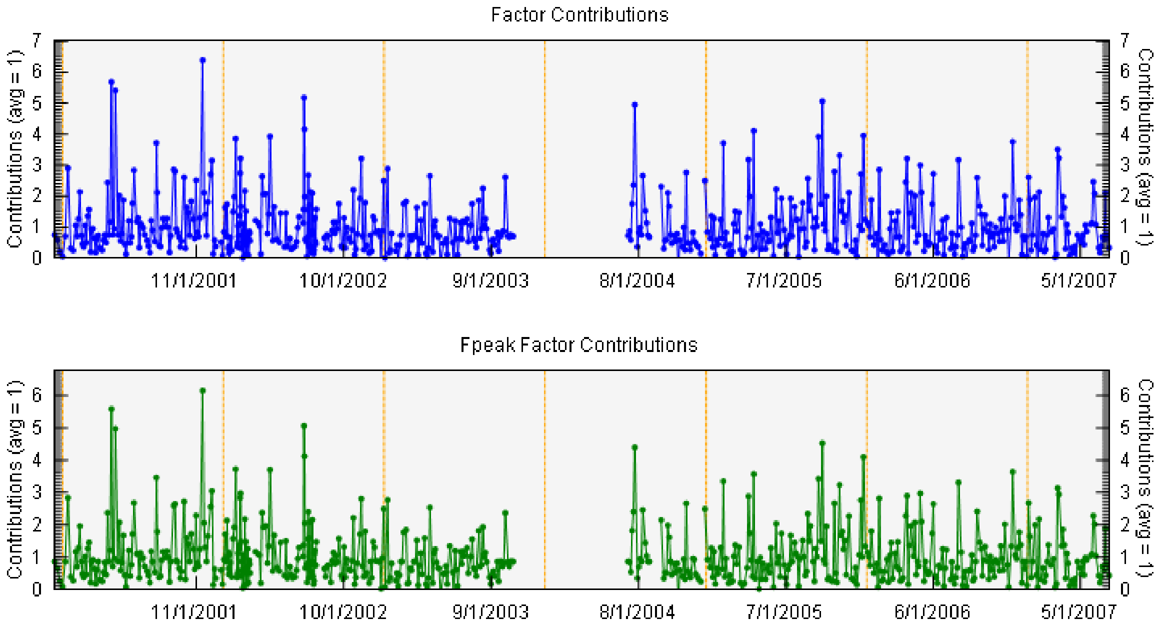


Figure 9-11.—Comparison of base run (top) and Fpeak run contribution (bottom) for industrial zinc factor.

Additional Analyses

The solution reached by PMF should be supported with additional analyses. For example, wind direction data and emissions inventories can be used to determine if local factors have high concentrations when winds are from the direction of known sources. The example in Figure 38 shows zinc emissions in the Baltimore area and a wind rose developed using wind data from the days with the highest zinc concentrations. The wind rose shows the highest zinc concentrations are almost always when wind is from the direction of the known zinc sources, confirming the identification of this factor as an “Industrial Zinc” factor.

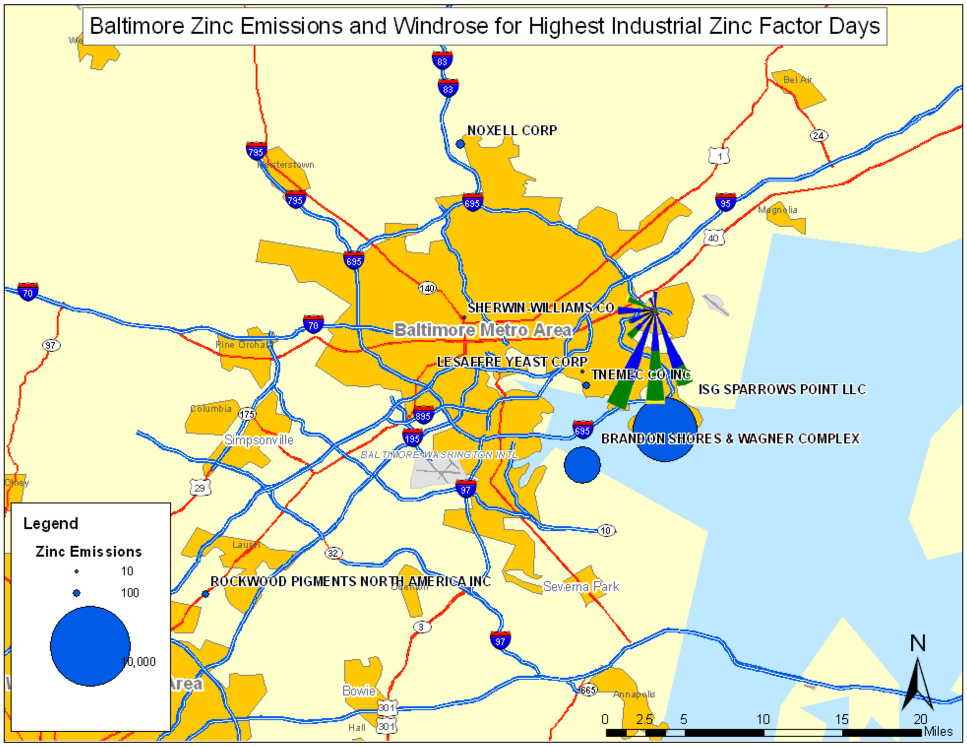


Figure 9-12.—Example wind rose for zinc factor.

9.1.8 References for this training exercise

Bae M.S., Schauer J.J., and Turner J.R. (2006) Estimation of the monthly average ratios of organic mass to organic carbon for fine particulate matter at an urban site. *Aerosol Sci. Technol.* **40**, 1123-1139, American Association for Aerosol Research.

Hyslop N.P. and White W.H. (2008) An evaluation of interagency monitoring of protected visual environments (IMPROVE) collocated precision and uncertainty estimates. *Atmos. Environ.* **42** 2691-2705.

Paatero P., Hopke P.K., Song X.H., and Ramadan Z. (2002) Understanding and controlling rotations in factor analytic models. *Chemometrics and Intelligent Laboratory Systems* **60**, 253-264.

Paatero P., Hopke P.K., and Philip K. (2003) Discarding or downweighting high-noise variables in factor analytic models. *Anal. Chim. Acta* **490**, 277-289.

Polissar A.V., Hopke P.K., and Poirot R.L. (2001) Atmospheric aerosol over Vermont: chemical composition and sources. *Environ. Sci. Technol.* **35** (23), 4604-4621.

Turpin B.J. and Lim H.-J. (2001) Species contribution to $PM_{2.5}$ mass concentrations: revisiting common assumptions for estimating organic mass. *Aerosol Sci. Technol.* **35** (10), 602-610.

White W.H. and Roberts P.T. (1977) On the nature and origin of visibility-reducing aerosols in the Los Angeles Air Basin. *Atmos. Environ.* **11**, 803-812.

Wade K.S., Brown S.G., Turner J.R., and Garlock J.L. (2008) Concentration value uncertainty estimates for source apportionment modeling of chemical speciation network data. Manuscript prepared for *A&WMA's 101st Annual Conference & Exhibition, Portland, OR, June 24-27* (STI-3271; Paper No. 632).

9.2 Sula Peak, Montana, Improve $PM_{2.5}$ Data Set

The following sections detail a complete PMF analysis of a $PM_{2.5}$ data set from Sula Peak, Montana (SULA). The user should run EPA PMF with the data set provided in **Sula.xls.** and duplicate the analyses described below. For all runs, a seed of 25 was used to ensure replicability. This exercise is intended to demonstrate the thought process and steps involved in reaching a solution using EPA PMF; it is not intended to be a complete source apportionment analysis.

9.2.1 Pre-PMF processing/Data set development

Concentration Input File
Data from the SULA1 site was downloaded from the Visibility Information Exchange Web System (VIEWS) web site. Data below the maximum reported detection limit for each species were substituted with one half of the detection limit and missing data were substituted with the median value. Additionally, the percent below detection was calculated to guide the user in species categorization. Species with more than 95% of samples below the detection limit were not included in the data set for PMF (see Table 3). The IMPROVE network reports missing data as -999; these values were left in the data set and the option within EPA PMF to replace missing values with the species median was chosen.

Uncertainty Input File
Uncertainties were provided for each species and sample except for organic carbon (OC) and elemental carbon (EC) and their fractions. For these species, 10% of the concentration was used as the uncertainty value.

Table 9-3. *Percent below detection limit for all species included in IMPROVE PM$_{2.5}$ data set for the SULA site. Species highlighted in yellow were not included in the PMF data set because more than 95% of samples were below detection; species highlighted in green were not included because they are represented by other species.*

Parameter	Percent Below Detection
ALf	58.4
ammNO3f	0
ammNO3f_bext	0
ammSO4f	0
ammSO4f_bext	0
ASf	97.4
BRf	1.0
CAf	26.7
CHLf	99.4
CLf	98.8
CRf	99.9
CUf	65.3
EC1f	39.7
EC2f	68.3
EC3f	91.9
ECf	0
ECf_bext	0
FEf	0.9
Hf	0.3
Kf	20.7
MF	7.5
MGf	98.8
MNf	93.2
N2f	99.7
NAf	93.2
NIf	91.3
NO3f	88.5
OC1f	92.5
OC2f	83.2
OC3f	58.7
OC4f	58.4
OCf	0
OMCf	0
OMCf_bext	0
OPf	70.4
PBf	24.2
Pf	99.7
RBf	64.6
RCFM	0
SEf	81.5
Sf	0.6
SIf	10.6
SO4f	28.3
SOILf	0
SOILf_bext	
SRf	49.9
TIf	79.1
Vf	99.7
ZNf	1.0
ZRf	97.9

9.2.2 Analyze input data

Characterizing Species (*Concentration/Uncertainty and Concentration Time Series*)
Species with more than 95% of data below detection and species represented by other species (i.e., duplication of mass) were categorized as “bad” (see highlighted species in Table 1-1). Additionally, aluminum (ALf) and silicon (SIf) were categorized as “bad” based on advisories from the IMPROVE program. PM$_{2.5}$ mass (MF) was chosen as the “Total Variable”. Total organic matter (OMf) and EC (ECf) were used in this analysis; therefore, OC fractions (OC1f, OC2f, OC3f, OC4f) and EC fractions (EC1f, EC2f, and EC3f) were excluded.

Relationships between Species (*Concentration Scatter Plot*)
Concentration scatter plots were examined for correlations indicating potential common influencing factors, such as meteorology or emissions. Soil species, in particular calcium (CAf) and iron (FEf), correlated well, indicating that these species are both predominantly crustal (Figure 39, top). EC and OM were also well correlated (Figure 39, bottom) which may indicate a common combustion source.

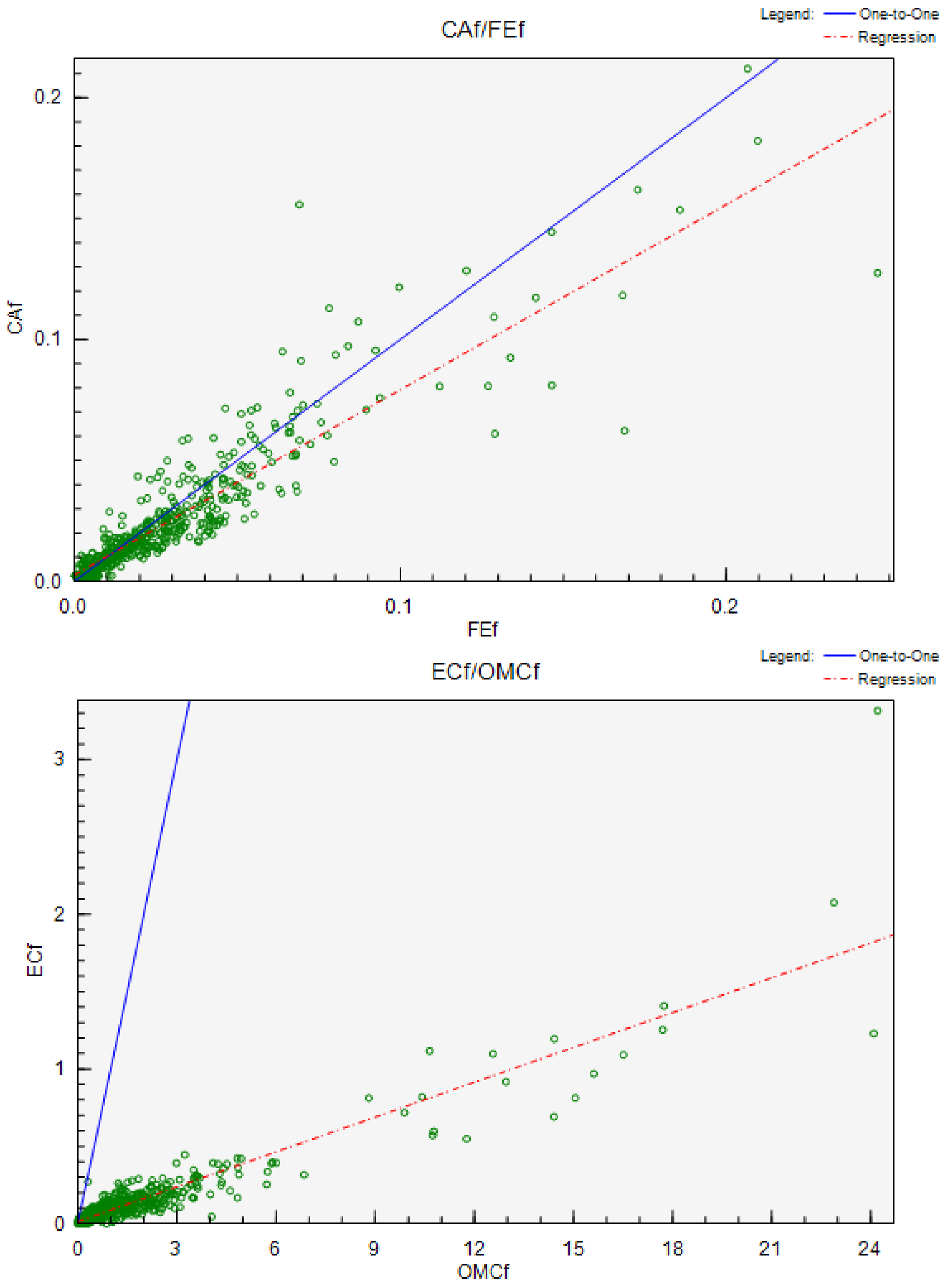


Figure 9-13.—Examples of well-correlated species.

Excluding Samples (*Concentration Time Series*)

Time series of each species were examined for outlier samples that should potentially be excluded from analysis. In this analysis, April 16, 2001 was excluded due to high concentrations of crustal elements (Figure 40, top) and April 5, 2006 was excluded due to high copper concentrations (Figure 40, bottom). Excluding days with high crustal elements is not always necessary; however, the ratios of the crustal species during this event are atypical and indicate an Asian soil event.

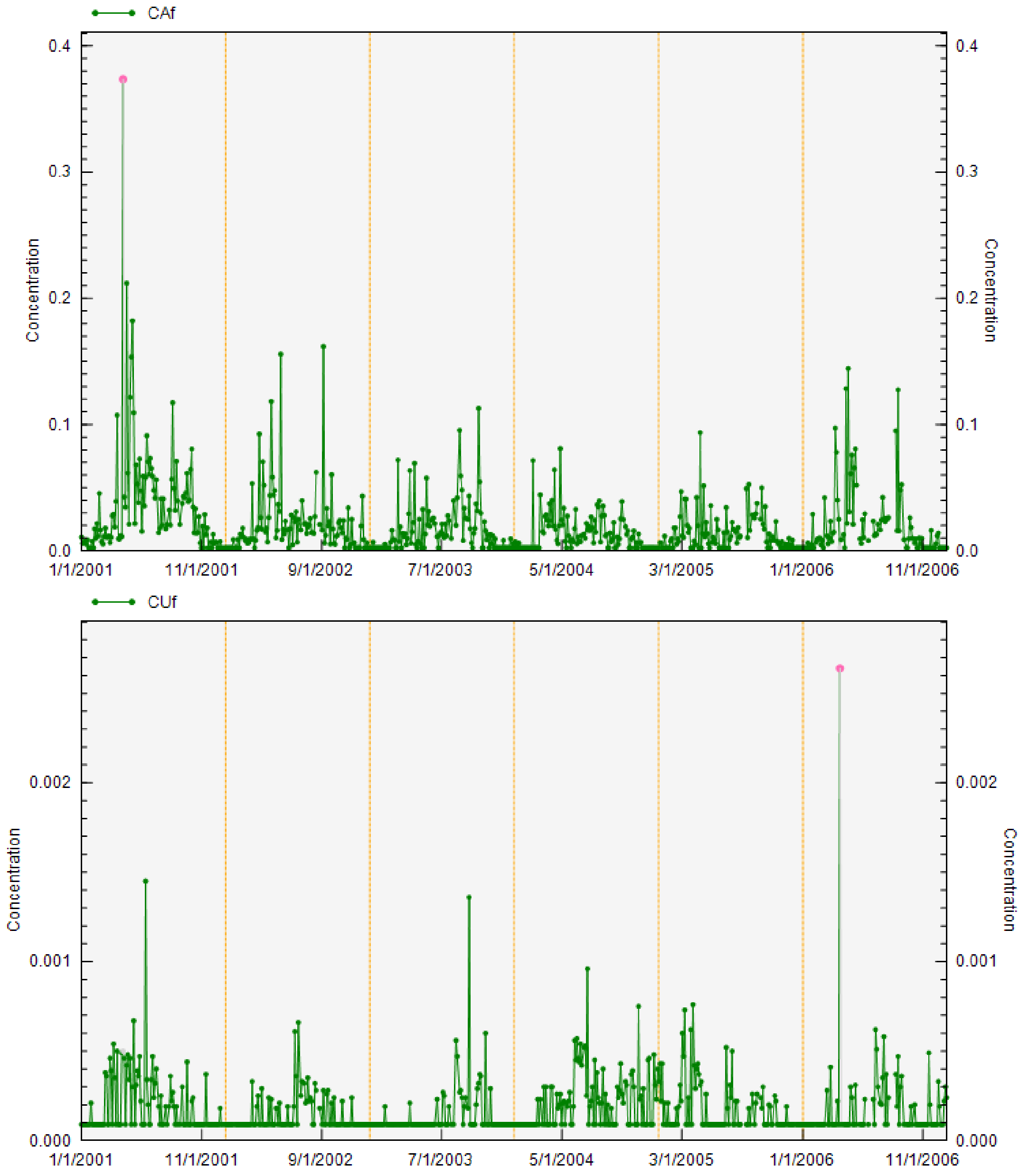


Figure 9-14.—Extreme events in calcium (top) and copper (bottom) concentrations.

9.2.3 Base Runs

Initial Runs (*Model Execution*)

The model was initially run with all included species categorized as “strong” and seven factors. Seven factors was selected based on the experience of analyzing similar data sets. A constant seed of 25 was used for reproducible results.

Model Evaluation (*Residual Analysis*, *O/P Scatter Plots*, *O/P Time Series*)

Several species had a large number of absolute scaled residuals greater than 3, including many metals. In particular, lead (PBf), rubidium (RBf), selenium (SEf), strontium (SRf). and copper (CUf) had large residuals; the graphs indicate poor observed-predicted correlations. Reported uncertainties for these species are often too low (Hyslop and White, 2008); these species were designated "weak" and the model was rerun.

In the second run, the Q-values were not stable. Examining the residual calculation in the diagnostic file shows that EC and bromine (BRf) contribute largely to the differences in solution space (i.e., variation in Q-values). Examining the "observed/predicted" graphs also shows that calcium (CAf) and titanium (TIf) were not modeled well. Uncertainties are often underestimated for these species. Additionally, no mass is apportioned to the zinc factor and the carbon fractions are separate factors, which is unexpected based on the correlation between EC and OM in the ambient data.

When the model is run with six factors and EC, BRf, CAf, and TIf are designated "weak", the Q values are still unstable. However, the zinc factor does have mass apportioned to it now. The zinc factor does have mass apportioned to it now, but the carbon fractions are still in separate factors. Examining the G-space plots of these factors shows they are very dependent. A five-factor solution should be explored next to see if EC and OM combine in one factor.

9.2.4 Final Base Run Results

Factor Identification (*Profiles/Contribs*, *Aggregate Contribs*)

The five factors were identified based on key species and temporal patterns. Factor identification is summarized in Table 4.

Table 9-4. *Identification of factors.*

Factor	Dominant Species	Temporal Pattern	Name
1	Zinc	None	Zinc smelter, Steel
2	Ammonium Sulfate, Copper, Lead	High in Summer	Secondary Sulfate/Transported Industry
3	Ammonium Nitrate	High in Winter	Secondary Nitrate
4	EC, OM, Potassium	High in Fall	Combined Carbon/Burning
5	Calcium, Iron, Titanium	High in Summer	Soil

Rotations (*G-Space Plots*)

Examination of G-space plots of the factors show evidence of an edge. In particular, the plot of Factors 1 and 4 and the plot of Factors 3 and 5 both exhibit an edge (Figure 41). The Fpeak feature should be used on the base solution to explore the rotational ambiguity.

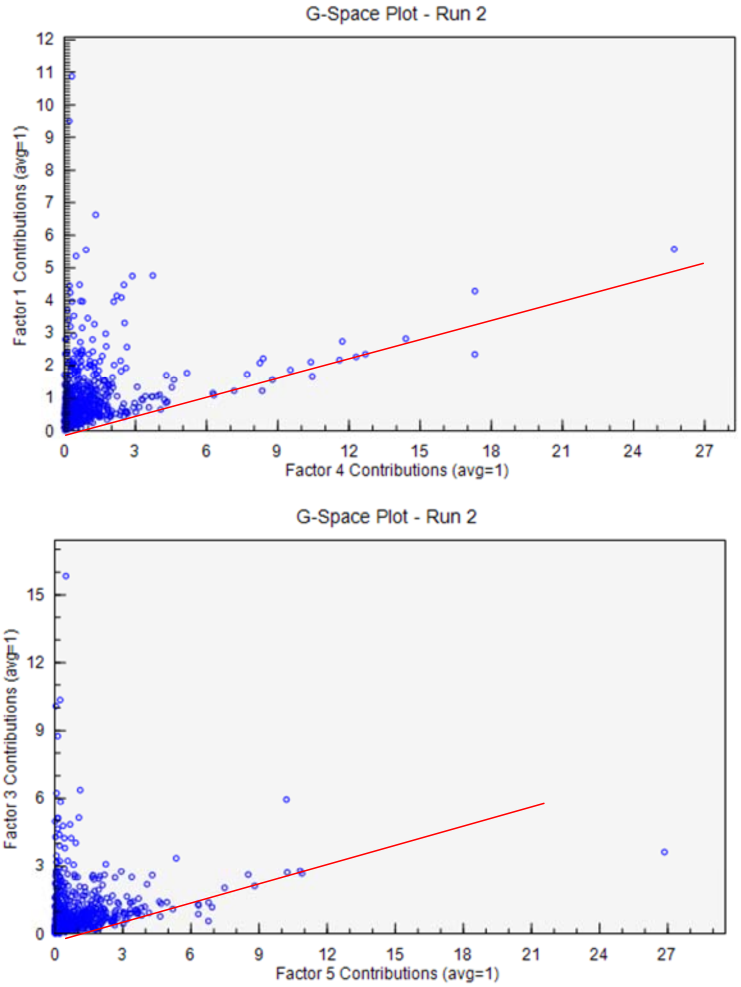


Figure 9-15.—G-Space plots indicating rotation of solution.

Mass Distribution (*Factor Pie Chart*)

Examination of the factor pie chart for the total mass (MF) shows that the combined carbon/burning factor is by far the largest contributor to total mass (Figure 42). The soil and secondary sulfate/transported industrial factors are also large contributors, while the secondary nitrate and industrial zinc factors account for only a few percent of the mass. This distribution seems reasonable for a remote site that is expected to be influenced mostly by burning and transport.

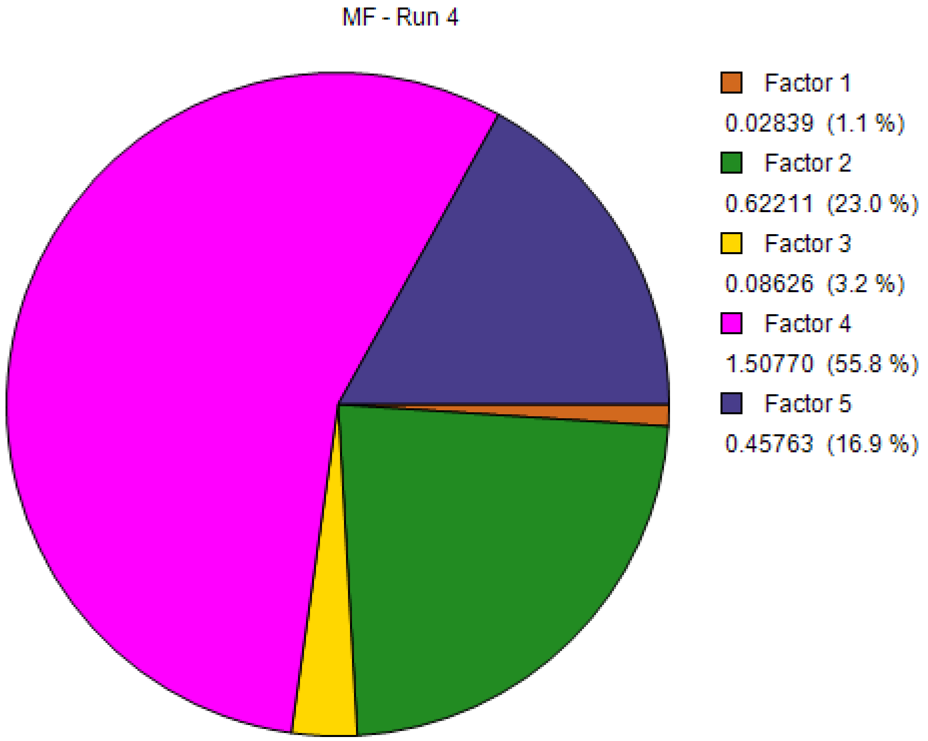


Figure 9-16.—Distribution of total mass among factors.

9.2.5 Bootstrap Runs

Input Parameters (*Model Execution*)

Bootstrapping was run on the final five-factor solution. Default bootstrapping parameters were used, including starting with base run 4, performing 100 runs, using an r-value of 0.6, and using the suggested block size of 20.

9.2.6 Bootstrap Run Results

Output Diagnostics (*Summary*)

All 100 runs were mapped to a factor and no factor had more than 100 runs mapped to it. This result indicates that the solution is stable.

Factor Variability (*Box Plots*)

For most factors, the interquartile ranges of bootstrapping results are very small (about 10%) (Figure 43, top). The exceptions are factors with trace metals zinc, lead, rubidium, calcium, and selenium, which are not unexpected as these concentrations are often near the detection limit. This agrees with the DDP results.

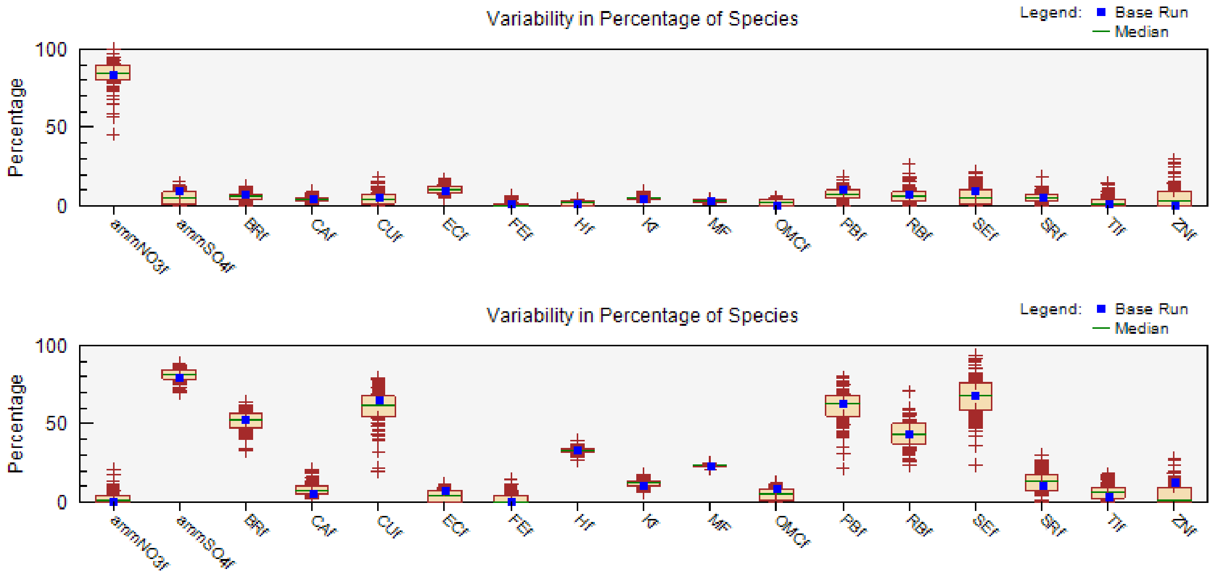


Figure 9-17.—Bootstrap model results for the secondary nitrate factor (top) and secondary sulfate/transported industry factor (bottom).

9.2.7 Fpeak Runs

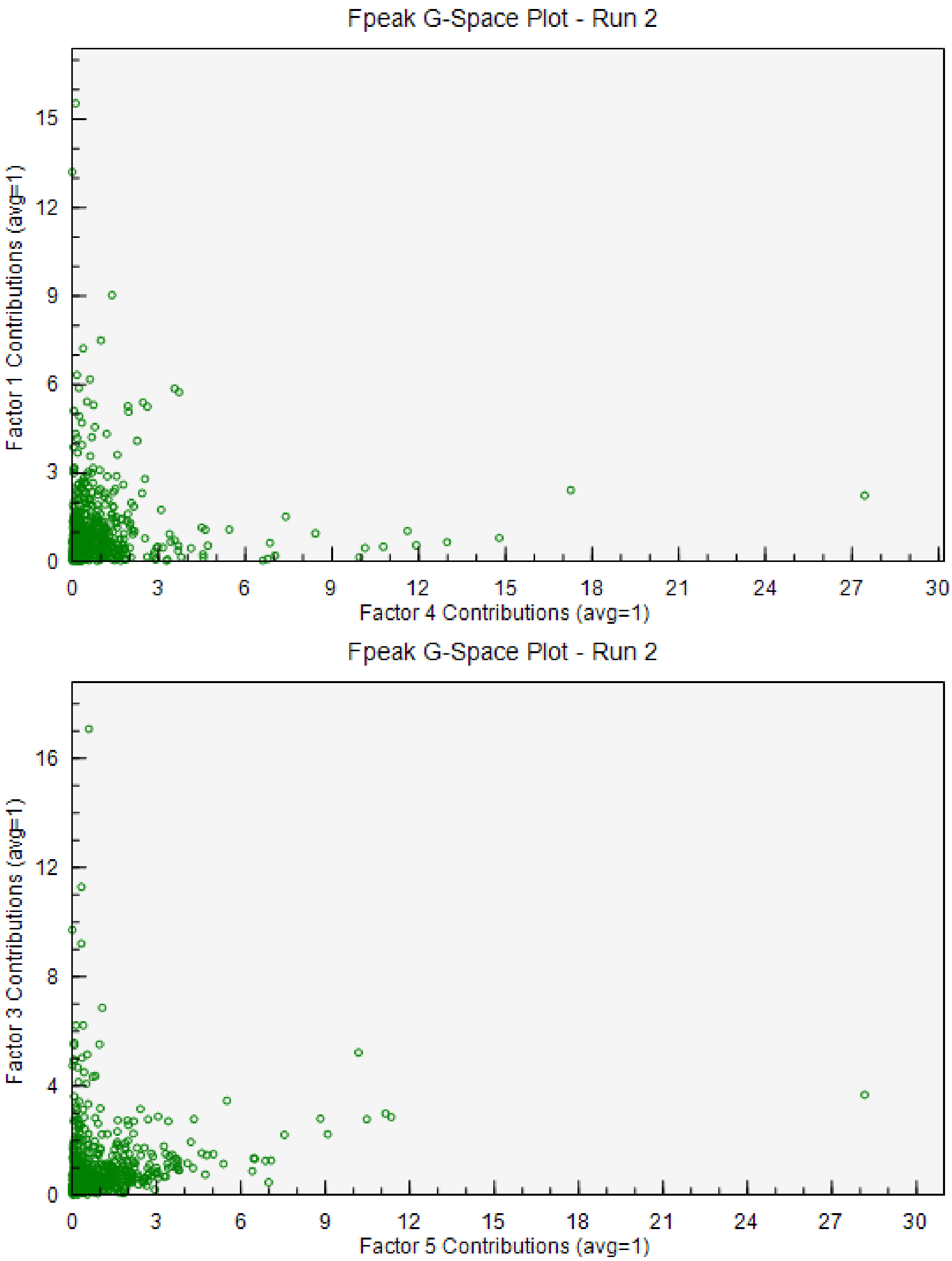
Input Parameters (*Model Execution*)

Examination of G-space plots between factors showed factor pairs with edges. A range of Fpeak values (from -2 to +2) was used to explore the solution. Positive values of Fpeak did not affect the G-space plots. Negative values beyond -1.5 increased the Q-value by more than 100 units. A value of -1.4 was shown to increase the independence of factors the most without increasing the Q value more than 100 units.

9.2.8 Fpeak Run Results

Contrast Between Factors (*G-Space Plots*)

Using an Fpeak value of -1.4 increased the contrast between Factors 1 and 4 (see Figure 44, top). However, the G-Space plot of Factors 3 and 5 has an edge (Figure 44, bottom). More advanced rotational tools are necessary to explore these results.

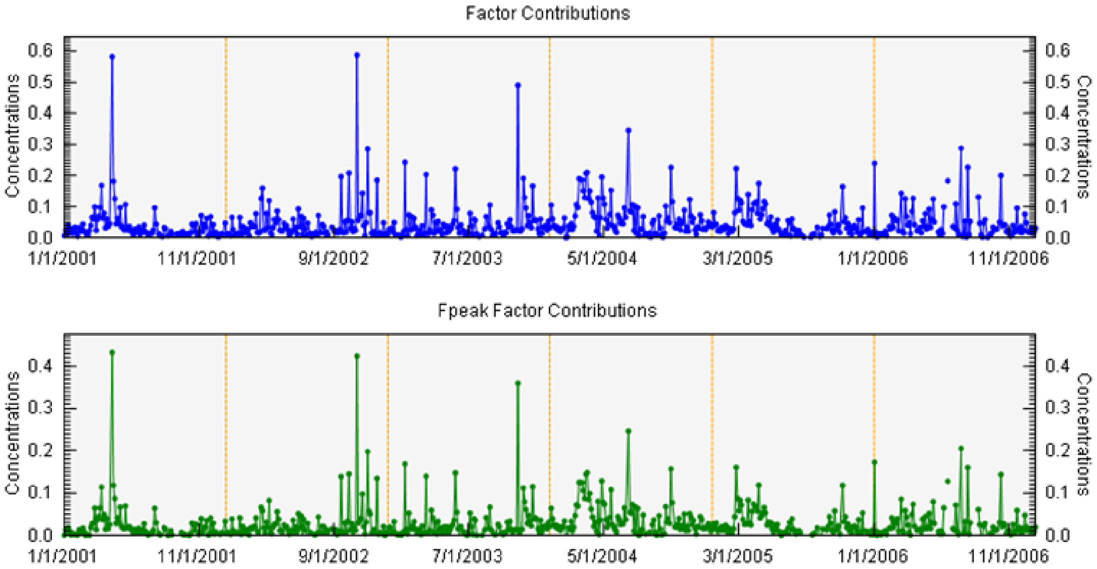

Fpeak G-Space Plot - Run 2

Fpeak G-Space Plot - Run 2

Figure 9-18.—G-Space plots after application of Fpeak at -1.4.

Rotation of Factors (Profiles/Contributions)

Using a negative Fpeak value may impact the contributions more than the profiles of the factors. Figure 45 shows that using an Fpeak of -1.4 pulls some contributions towards 0, as was shown in the G-space plots.


Factor Contributions

Fpeak Factor Contributions

Figure 9-19.—Comparison of contributions of the secondary nitrate factor in the base run (top) and Fpeak run (bottom).

Additional Analyses

Further analysis outside EPA PMF should be performed to verify these results. Satellite data and forest fire inventories may be examined to determine if large fires were present on the days when the combined carbon/burning factor was high. Wind direction and trajectory analyses can be used to determine the likelihood of fires impacting the site on these days. Trajectory analysis can also be used to examine the secondary sulfate/transported metals factor. Additionally, emissions inventories for zinc may be examined to determine if there is a zinc source in the area.

9.3 Baton Rouge, Louisiana, PAMS VOC Data Set

The following sections detail a PMF analysis of a PAMS VOC data set from Baton Rouge, Louisiana. The user should run EPA PMF with the data sets provided in **BatonRouge.xls** and duplicate the analyses described below. This exercise is intended to demonstrate the thought process and steps involved in reaching a solution using EPA PMF; it is not intended to be a complete source apportionment analysis.

9.3.1 Pre-PMF processing/Data set development

Concentration Input File

Data for this analysis were downloaded from AQS. All hourly PAMS VOC data for June–August 2005 (682 samples) at the Baton Rouge site were downloaded for potential inclusion in PMF. Table 5 lists the species available and the percent below detection.

Table 9-5. *Percent below detection limit for all species included in the PAMS VOC data set for the Baton Rouge site. Species highlighted in yellow were not included in the PMF data set because more than 50% of samples were below detection; species highlighted in green were not included because they had noticeable step changes in concentrations, indicating a change in collection or analysis methods. Species boldfaced were used in PMF.*

Parameter	Percent Below Detection	Parameter	Percent Below Detection	Parameter	Percent Below Detection
1,2,3-Trimethylbenzene	48%	**Benzene**	0%	**N-Decane**	30%
1,2,4-Trimethylbenzene	9%	Cis-2-Butene	52%	N-Heptane	4%
1,3,5-Trimethylbenzene	36%	Cis-2-Pentene	24%	**N-Hexane**	0%
1-Butene	13%	Cyclohexane	15%	N-Nonane	21%
1-Pentene	14%	Cyclopentane	25%	N-Octane	13%
2,2,4-Trimethylpentane	1%	**Ethane**	0%	**N-Pentane**	0%
2,2-Dimethylbutane	29%	**Ethylbenzene**	4%	N-Propylbenzene	58%
2,3,4-Trimethylpentane	12%	**Ethylene**	0%	**N-Undecane**	30%
2,3-Dimethylbutane	12%	**Isobutane**	0%	**O-Ethyltoluene**	38%
2,3-Dimethylpentane	22%	**Isopentane**	0%	**O-Xylene**	5%
2,4-Dimethylpentane	29%	**Isoprene**	6%	P-Diethylbenzene	68%
2-Methylheptane	34%	Isopropylbenzene	75%	P-Ethyltoluene	42%
2-Methylhexane	6%	M_P Xylene	1%	**Propane**	0%
2-Methylpentane	1%	M-Diethylbenzene	71%	**Propylene**	0%
3-Methylheptane	35%	Methylcyclohexane	15%	**Styrene**	20%
3-Methylhexane	3%	Methylcyclopentane	1%	**Toluene**	0%
3-Methylpentane	0%	M-Ethyltoluene	18%	Trans-2-Butene	54%
Acetylene	0%	**N-Butane**	0%	Trans-2-Pentene	16%

Uncertainty Data Set

Uncertainties are not regularly reported for PAMS VOC data. For this analysis, 20% of the concentration was used as the initial uncertainty for each species.

9.3.2 Analyze input data

Characterizing Species (*Concentration/Uncertainty and Concentration Time Series*)

For the initial run, all included species were left as strong. Signal-to-noise ratios are not as useful in this analysis because all species were given a 20% uncertainty; therefore species categorizations will be evaluated based on residuals and observed predicted statistics after the initial base run. No species was included as a total variable in this data set.

Relationships between Species (*Concentration Scatter Plot*)

Scatter plots between species are examined to evaluate relationships between species which may indicate a common source. In the Baton Rouge data set, expected relationships between gasoline mobile species (such as toluene and o-xylene) and heavy duty mobile species (such as decane and undecane) were seen (Figure 46). Ethane and propane show some evidence of bifurcation, potentially indicating a mix of fresh sources from petrochemical processing/natural gas use and aged carryover from other areas. Benzene and styrene, often mobile-dominated species, were not well correlated with mobile species, likely due to additional petrochemical sources in the area. Several large refineries in the area could be contributing to these concentrations.

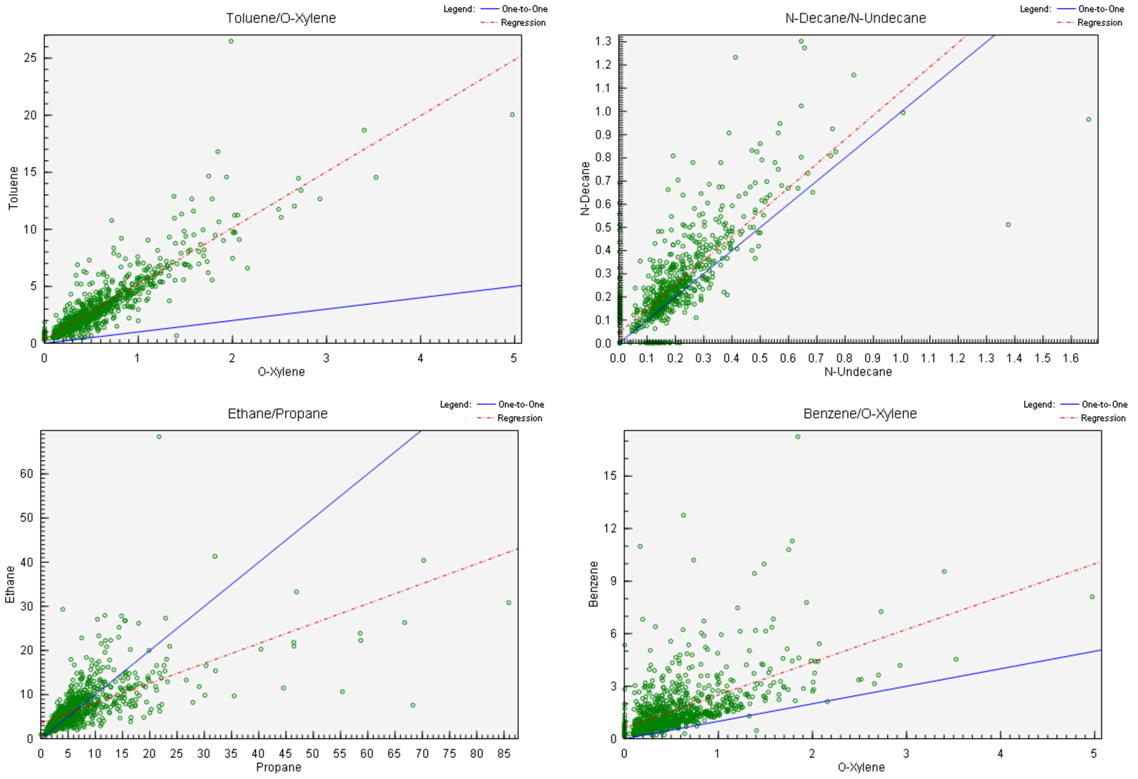


Figure 9-20.—Relationships between ambient concentrations of various species.

Excluding Samples and Species (*Concentration Time Series*)

Time series of each pollutant were examined to look for extreme events that should be removed from the analysis. Five samples were removed due to events in various species (Figure 47): 8/5/05 9:00:00 PM (2-methylheptane), 8/7/05 09:00:00 AM and 12:00:00 PM (n-undecane), 8/6/05 6:00:00 AM (o-ethyltoluene), 7/21/05 9:00:00 AM (propylene). The 8/5/05–8/7/05 samples were possibly part of the same event, further data analysis outside of EPA PMF could be used to confirm if the data are real and informative.

Several species had noticeable step changes in concentrations (see example in Figure 48), indicating a change in sampling or analytical method. These types of changes may be identified as separate sources, therefore these species (3-methylheptane, m/p-xylene, and m-ethyltoluene) were excluded from analysis.

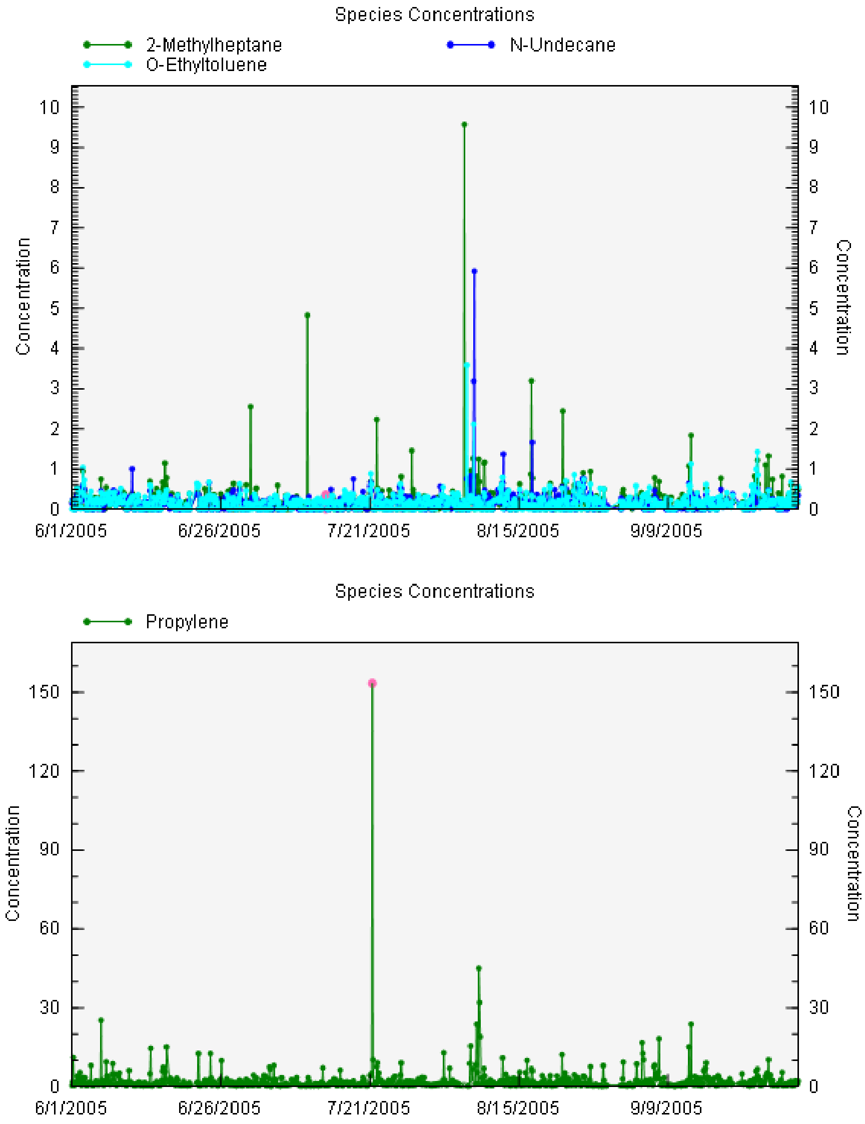


Figure 9-21.—Extreme values excluded from analysis.

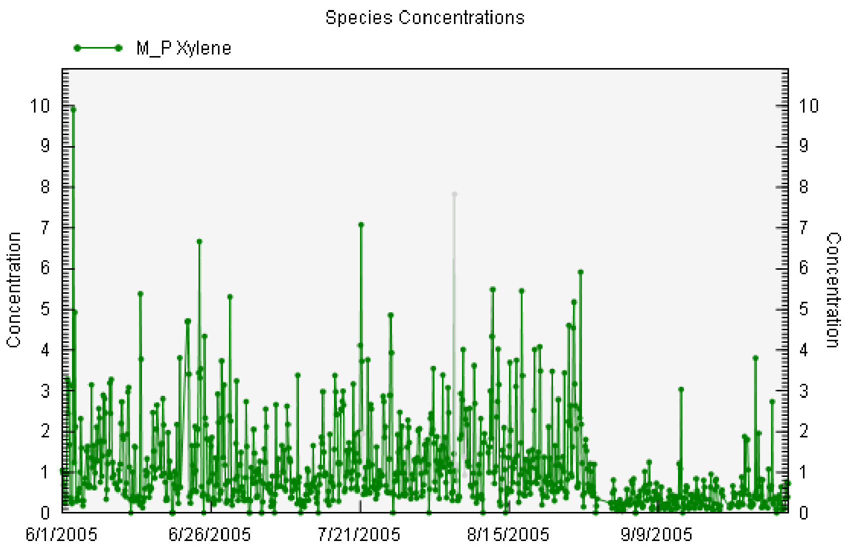
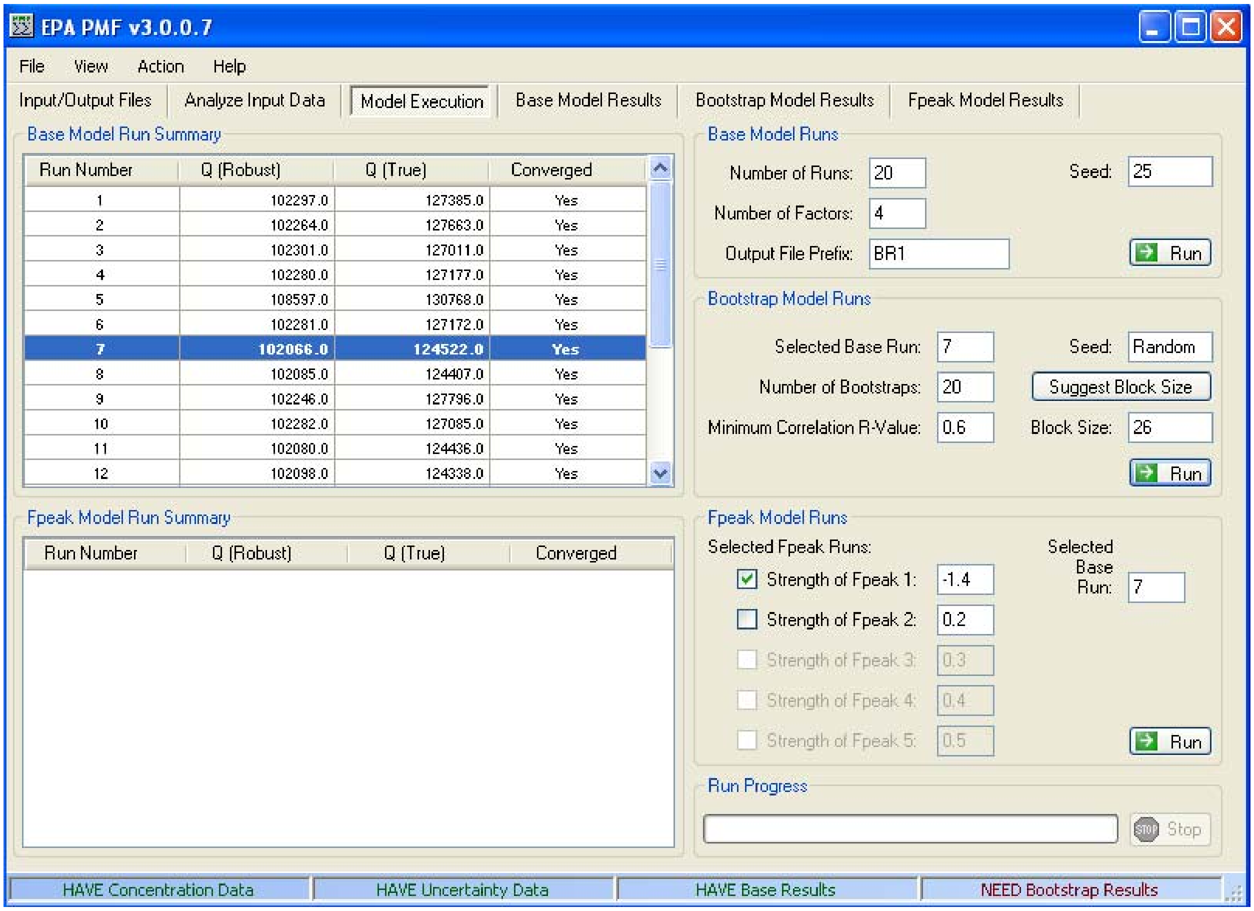


Figure 9-22.—Example of step change in concentrations.

9.3.3 Base Runs

Initial Model Parameters (*Model Execution*)

Initially, 20 base runs with 4 factors and a seed of 25 were run. In this iteration, the Q values varied by several hundred units, indicating the solution may not be stable (Figure 49).



Base Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged
1	102297.0	127385.0	Yes
2	102264.0	127663.0	Yes
3	102301.0	127011.0	Yes
4	102280.0	127177.0	Yes
5	108597.0	130768.0	Yes
6	102281.0	127172.0	Yes
7	**102066.0**	**124522.0**	**Yes**
8	102085.0	124407.0	Yes
9	102246.0	127796.0	Yes
10	102282.0	127085.0	Yes
11	102080.0	124436.0	Yes
12	102098.0	124338.0	Yes

Fpeak Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged

Base Model Runs

Number of Runs: 20
Seed: 25
Number of Factors: 4
Output File Prefix: BR1

Bootstrap Model Runs

Selected Base Run: 7
Seed: Random
Number of Bootstraps: 20
Minimum Correlation R-Value: 0.6
Block Size: 26

Fpeak Model Runs

Selected Fpeak Runs:
Selected Base Run: 7

☑ Strength of Fpeak 1: -1.4
□ Strength of Fpeak 2: 0.2
□ Strength of Fpeak 3: 0.3
□ Strength of Fpeak 4: 0.4
□ Strength of Fpeak 5: 0.5

HAVE Concentration Data | HAVE Uncertainty Data | HAVE Base Results | NEED Bootstrap Results

Figure 9-23.—Model Execution tab after completion of first round of base runs.

9.3.4 Base Run Results

Model Reconstruction (*O/P Scatter Plots*, *O/P Time Series*)

Residuals of many species were skewed high or bimodal (Figure 50). These species, n-decane, n-undecane, o-ethyltoluene, and styrene, also had poor observed-predicted plots, which illustrate that peak concentrations are not modeled well (Figure 51, left) and, for n-decane and n-undecane, low concentrations (below the detection limit) are not well modeled (Figure 51, right). All these species will be recategorized as weak as they are typically not as well measured as other PAMS species.

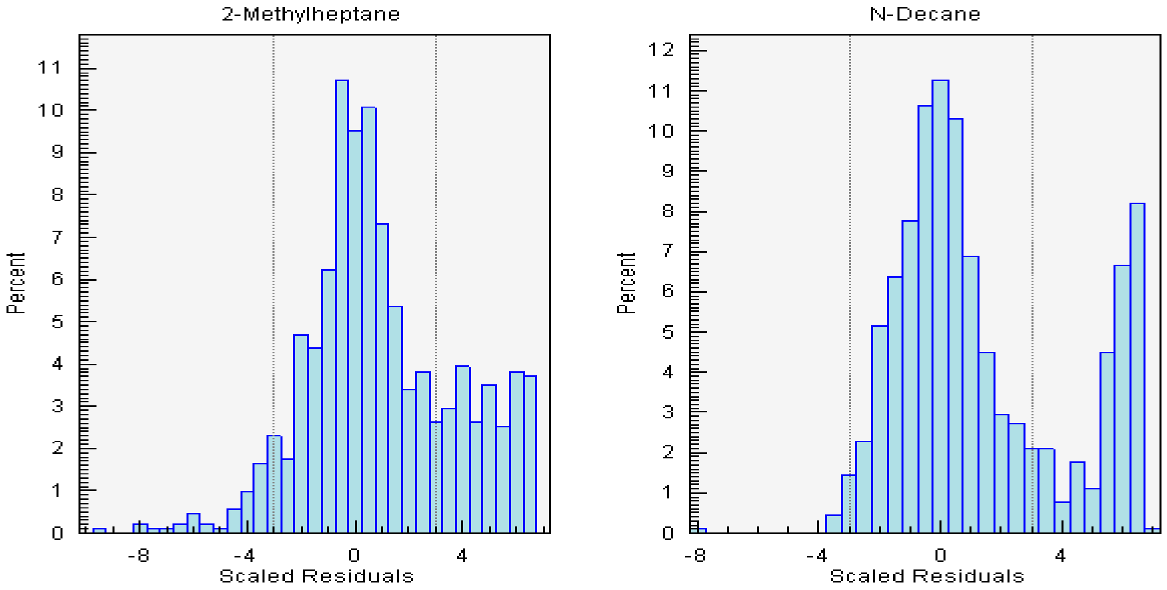


Figure 9-24.—Non-normal scaled residuals.

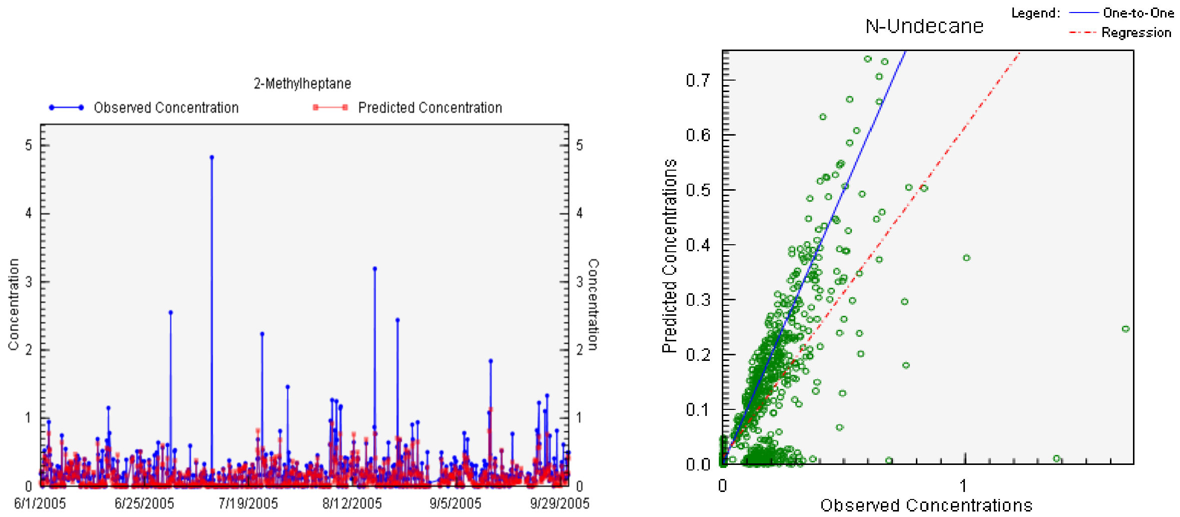


Figure 9-25.—Observed/Predicted plots of poorly modeled species.

Factor Identification (*Profiles/Contribs*, *Aggregate Contribs*)

Profiles and contributions were examined to identify factors. In the initial run, the first factor had a large contribution of both n-decane and n-undecane, but also had contributions of 2-methylheptane, o-ethyltoluene, and styrene. The second factor was isoprene, which is a biogenic marker. The third factor had ethane, ethylene, propane, and propylene, representing natural gas/petrochemical industry, along with acetylene and benzene, which are traditionally mobile source markers. The final factor contained the butanes and pentanes, indicative of a solvent source. Because of the non-normal residuals and the appearance of unexpected species in some of the factors, and the lack of a clear mobile factor, a higher number of factors should be explored.

Because this data set consists of summer only data for one year, the seasonal and annual aggregate contributions are not useful. The n-decane/n-undecane factor may have a day of week pattern if it represents heavy-duty traffic, but in this iteration no trend was evident.

Rotations (*G-Space Plots*)

Examination of G-space plots showed some rotation may be present. In particular, the Factor 1 versus Factor 3 plot shows a clear edge (**Figure 52**). Using a different number of factors may eliminate this, but it should be revisited with each iteration.

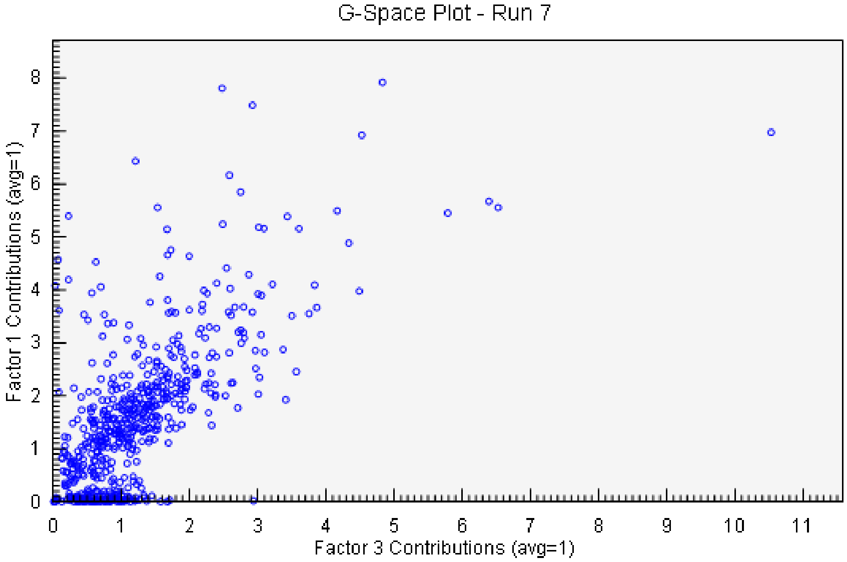


Figure 9-26.—Illustration of rotational ambiguity in the initial base run solution.

Species Distribution (*Factor Pie Chart*)

Because no total variable was used for this data set, the factor pie charts are best used to see the distribution of individual species among the factors. Key species, such as mobile tracers or toxic species are of particular interest. For example, Figure 53 shows that while most of the benzene is in Factor 3 (the ethane/ethylene, propane/propylene factor), a large fraction of it is also in Factor 4 with the butanes and pentanes.

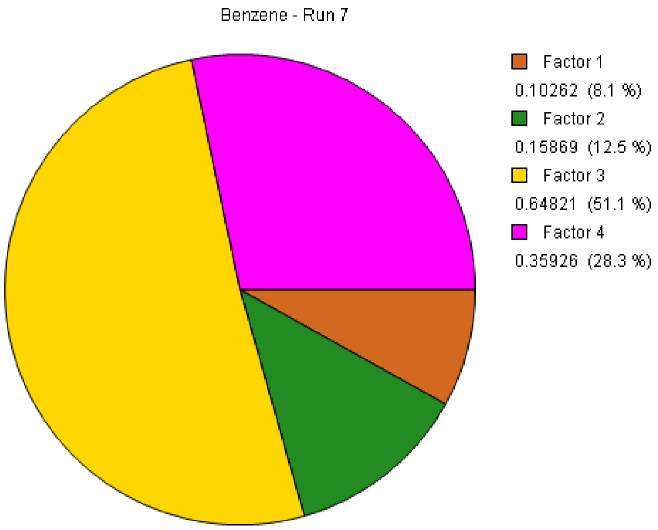


Figure 9-27.—Apportionment of benzene to factors resolved in initial base run.

Base Model Runs with Updated Species Categorization

The model was next run with five factors and with n-decane, n-undecane, o-ethyltoluene, and styrene categorized as weak. The Q values were still not stable, but the scaled residuals were more reasonable. The additional factor in this iteration is characterized by 3-methylhexane, ethylbenzene, o-xylene, and styrene, which, with the exception of styrene, are largely mobile markers. To try to stabilize the Q-values, this iteration was re-run with the extra-modeling uncertainty set to 15%. There are still two distinct minima. Examining the sum of the squares of the differences in residuals shows that 2-methylheptane and 3-methylhexane are varying the most between runs. These species were re-categorized as weak.

When the model is run with the additional weak species, Q-values are stable and the residuals for all species are reasonable. The five factors in this solution are evaporation (pentanes/butanes), heavy duty (n-decane/n-undecane), biogenic (isoprene), natural gas/industry (ethane/ethylene/propane/propylene), and mobile (acetylene, ethylbenzene, o-xylene, and toluene). A six-factor solution was explored to see if any additional factors could be identified. An additional factor with propylene and ethylene was resolved as a factor independent of propane and ethane. The propylene/ethylene factor likely represents the petrochemical industry whereas the ethane/propane represents natural gas and accumulation of aged air. The edges observed in the initial solution are still present in this solution (Figure 54). A seven-factor solution resolved an independent n-hexane factor. Additional exploration of sources in the area is needed to confirm if this is a physically realistic factor. For this analysis, the six-factor solution will be considered the final solution.

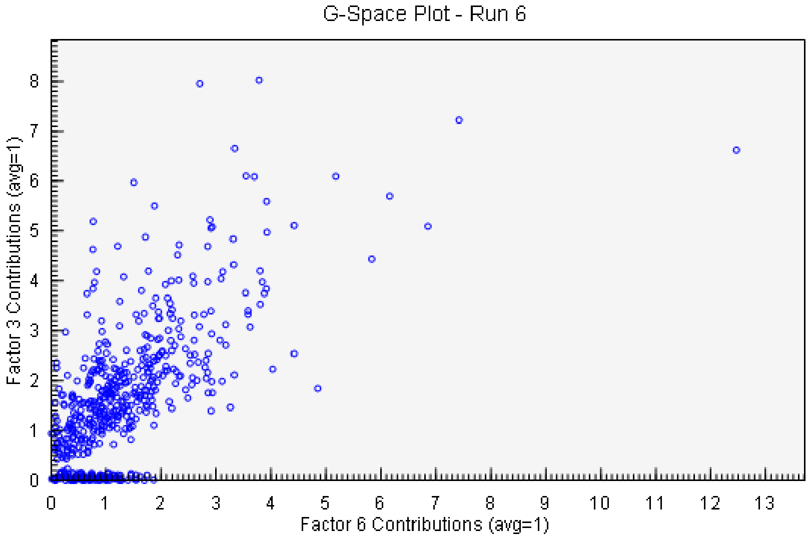


Figure 9-28.—Illustration of rotational ambiguity in final base run solution.

9.3.5 Bootstrap Runs

Input Parameters (*Model Execution*)

Bootstrapping was run with all of the default input parameters: base run 6, 100 bootstraps, minimum r value of 0.6, and suggested block size of 26. A seed of 25 was used to ensure replicability.

9.3.6 Bootstrap Run Results

Output Diagnostics (*Summary*)

Out of the 100 runs, at least 97 bootstrap factors were mapped to each base factor. Only 2 factors were unmapped. This indicates a relatively stable result. The unmapped factors should be examined to determine if any patterns are evident.

Factor Variability (*Box Plots*)

Most species had small interquartile ranges of around 15%, indicating little variability in the factors. The exception was 3-methylhexane, which had large interquartile ranges in Factors 1 (natural gas) (Figure 55, top), 4 (evaporation), and 6 (mobile). Because this species is already weak, additional runs excluding it should be explored.

The unmapped factor shows no obvious pattern (Figure 56, bottom), which is expected as only two factors were unmapped.

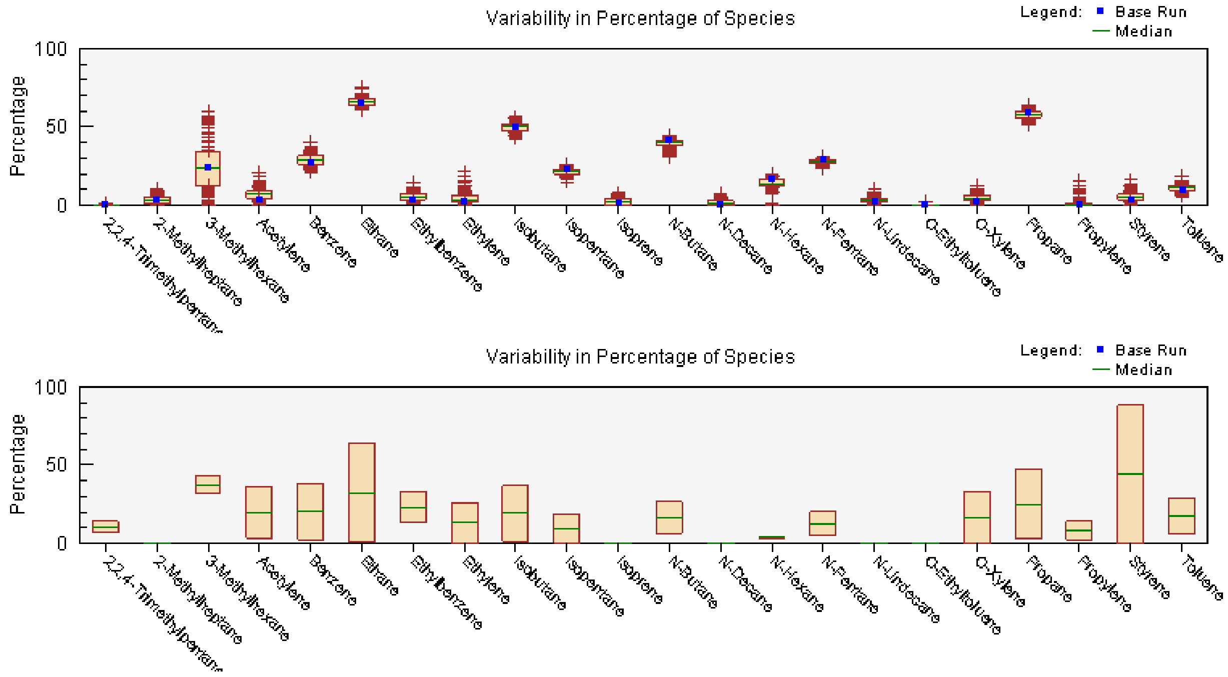


Figure 9-29.—Example of interquartile ranges for bootstrap results illustrating the relatively large interval for 3-methylhexane and the unmapped factor.

9.3.7 Fpeak Runs

Input Parameters (*Model Execution*)

As noted in the base model results section, some rotational ambiguity was observed in this solution. Specifically, G-space plots of Factors 3 and 6 have an edge that does not align to the axis and will be the focus of this section. Base run 6 was chosen as the starting point for Fpeak. Fpeak values between -1.5 and 1.5 were explored. Values beyond this range increased the Q values by more than 150 units.

9.3.8 Fpeak Run Results (G-Space Plots, Profiles/Contributions)

None of the Fpeak values tested produced a noticeable change in the G-space plots. Noticeable changes in the profiles and contributions were also not seen with the range of Fpeak values used. Additional rotational tools (available outside of EPA PMF using ME-2) should be used to further evaluate the rotations.

9.3.9 Additional Analyses

To support the source apportionment results, sources of VOCs in the area should be examined. If local emission inventories are available, these should be examined to determine if they agree with source apportionment results. If no speciated inventory is available, individual sources in the area should be evaluated. Wind direction analysis, using the point sources identified as well as information about local roads, would support the factor identification. Other years of data could also be modeled and compared to the 2005 data set used here.

10.0 ACRONYMS

Acronym	Definition
AQS	Air Quality System
DDP	Discrete Difference Percentiles
EC	Elemental Carbon
GUI	Graphical user interface
IMPROVE	Interagency Monitoring of Protected Visual Environments
MF	Total Mass
O/P	Observed/Predicted
OC	Organic Carbon
OM	Organic Matter
PAMS	Photochemical Assessment Monitoring Stations
PMF	Positive Matrix Factorization
S/N	Signal-to-noise ratio
STN	Speciation Trends Network
SULA	Sula Peak, Montana
UAC	User Account Control
VIEWS	Visibility Information Exchange Web System
VOC	Volatile Organic Compound

United States
Environmental Protection
Agency

(8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
$300

EPA 600/R-08/108
Month date
www.epa.gov