EPA PMF 1.1 User's Guide

June 30, 2005

Shelly Eberly

U.S. Environmental Protection Agency
National Exposure Research Laboratory
Research Triangle Park, NC 27711

Acknowledgements

EPA PMF 1.1 would not be possible without the code base developed over the past ten years by Dr. Paatero for Positive Matrix Factorization (PMF2, PMF3, and ME-2). The EPA project officers for the series of EPA contracts that have supported and continue to support the development of EPA PMF include Janet Burke and Shelly Eberly.

Disclaimer

EPA through its Office of Research and Development funded and managed the research and development described here. It is being subjected to Agency review (internal and external), thus it has not been cleared for official distribution by the EPA. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Table of Contents

Section 1. Introduction

This user's manual is designed to guide a user through the use of the software EPA PMF 1.1 for a receptor modeling application. The manual covers details about what the user should do to operate the software, but it intentionally does not include guidance on how to interpret the output, other than to provide indicators about how well the model fits the user-provided data.

This manual is divided into 6 sections. The first section, this one, contains information about the development of the software and the model underlying the software. Section 2 describes the steps for installing, removing, or reinstalling the software from a computer system and offers important data management suggestions. Section 3 provides details about executing the software and Section 4 covers the output generated by the software. Section 5 covers some of the more advanced options, and the manual closes with a listing of numerous publications using PMF in Section 6.

The manual is a work in progress and users are encouraged to send suggestions to the software support point of contact listed below.

1.1 General Notes about User's Manual

Throughout the user's manual there are sample screen shots from the software and discussions about the contents of these screen shots. The screen shots have not been taken from one continuous application of the software to one data set, thus there is not necessarily continuity from one screen shot to the next. Where continuity is important for clarifying a point, the continuity of the example is maintained.

1.2 History of Development of Software

The richness of ambient air quality data sets has been increasing in recent years due to more species being measured, species being stratified by particle size, and sampling durations decreasing. To take advantage of these richer data sets, the receptor models have become more complex. One receptor model is Positive Matrix Factorization (PMF), developed by Dr. Paatero at the University of Helsinki in Finland in the mid 1990's. Since the release of the programs PMF2, PMF3 and ME2, there have been numerous applications in air quality to resolve source types or source regions. See the reference section at the end of this manual for a listing of several such applications.

To ensure that receptor modeling tools, both simple and complex, are available for use in the development and implementation of air quality standards, the United States Environmental Protection Agency's Office of Research and Development has and continues to develop a suite of receptor models that are freely distributed to the air quality management community. Where possible, a common modeling platform is used so that a user familiar with one of the models can easily transition and use another model. EPA PMF has been under development since 2003 and is now one of the models in this suite.

1.3 Status of Software

EPA PMF 1.1 is being subjected to internal and external peer review. Once the reviews are completed, a public workshop will be held late in 2005 to discuss the model, user's guide, associated guidance documents, and reviews. It is anticipated that a summary of this workshop and responses to comments will be available within two months after the workshop. If the reviews are favorable, then EPA PMF 1.1 will be officially released shortly after the workshop is completed.

Please send comments, questions, or suggestions regarding the software and user's guide to Shelly Eberly at eberly.shelly@epa.gov or (919) 541-4128. Feedback will be used to enhance future versions of the software.

1.4 Basics of Model Solved in EPA PMF

EPA PMF is basically a graphical user interface that has been "wrapped around" the Positive Matrix Factorization model, solved using the multilinear engine as implemented in the program ME-2. An extremely simplified version of the model is explained in this section. However, users of the software are strongly encouraged to read the papers listed in the references for this section as these papers contain the theoretical properties of the model and recommendations for applying the model. The description below is general and is not intended to contain all the details.

EPA PMF 1.1 solves the general receptor modeling problem using constrained, weighted, least-squares. The general model assumes there are *p* sources, source types or source regions (termed factors) impacting a receptor, and linear combinations of the impacts from the *p* factors give rise to the observed concentrations of the various species. Mathematically stated,

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

where x_{ij} is the concentration at a receptor for the j^{th} species on the i^{th} day, g_{ik} is the contribution of the k^{th} factor to the receptor on the i^{th} day, f_{kj} is the fraction of the k^{th} factor that is species *j*, and e_{ij} is the residual for the j^{th} species on the i^{th} day. In EPA PMF, it is assumed that only the x_{ij}'s are known and that the goal is to estimate the contributions (g_{ik}) and the fractions (or profiles) (f_{kj}). It is assumed that the contributions and mass fractions are all non-negative, hence the "constrained" part of the least-squares. Additionally, EPA PMF allows the user to say how much uncertainty there is in each x_{ij}. Species-days with lots of uncertainty are not allowed to influence the estimation of the contributions and profiles as much as those with small uncertainty, hence the "weighted" part of the least squares.

The task of EPA PMF is to minimize the sum of squares

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{s_{ij}} \right)^2$$

where s_{ij} is the uncertainty in the j^{th} species for day i. EPA PMF operates in a robust mode, meaning that "outliers" are not allowed to overly influence the fitting of the contributions and profiles.

If the model is appropriate for the data and if the uncertainties specified are truly reflective of the uncertainties in the data, then Q should be approximately equal to the number of data points in the concentration data set. Use this as a basic guide, only.

1.5 Software Support

To receive technical support for EPA PMF 1.1, please contact Shelly Eberly at eberly.shelly@epa.gov or (919) 541-4128. Please do not contact Dr. Paatero.

1.6 References

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 Computational and Graphical Statistics, Volume 1, Number 4, 854-888.

Section 2. Installing and Uninstalling the Software

2.1 System Requirements

The software is designed to run under Windows 2000 or Windows XP. If the monitor is smaller than 15 inches, some of the menus and pop-up informational boxes may not display in their entirety. Installation requires approximately 90 megabytes of hard disk space to be available. Model output from each run requires minimal disk space. The speed of execution depends on the memory, clock speed, bus speed, etc. of the user's computer. It is recommended that at least 512 megabytes of memory be available if operating under Windows XP. Less memory is needed under Windows 2000.

You must have system administration privileges on the machine to install or uninstall the software.

You must have write privileges in the directory in which EPA PMF is installed. For the default setting, this means that you must have write privileges in "c:\Program Files\EPA PMF 1."

2.2 Installing the Software

To install the software,

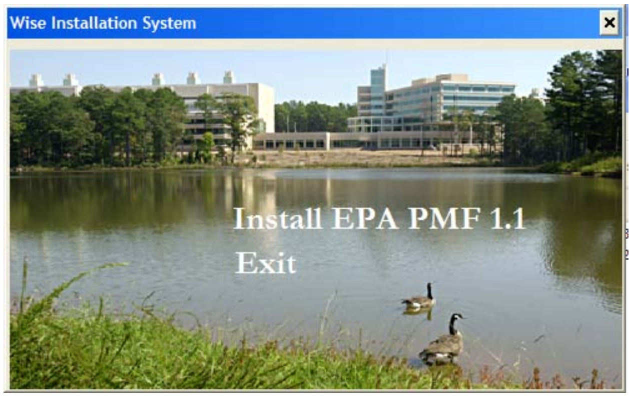


Figure 1. Opening Installation Screen

a. Insert the EPA PMF 1.1 CD into the CD-ROM drive. The following installation screen appears. If the opening installation screen does not appear automatically when the CD is inserted into the CD-ROM drive, go to Windows Explorer, double-click on the CD-ROM drive, and then double-click on the file "autoplay.exe" which is located on the CD.

b. Select "Install EPA PMF 1.1" and follow the on-screen instructions. Installation of EPA PMF includes installation of Matlab's Component Runtime Libraries so do not be concerned when queried about installing these. The default options in the installation are the correct options for most users. Installation may take several minutes.

c. Once the installation is complete, select "Exit" from the opening installation screen shown in Figure 1.

After installation, the user should see a new icon on the desktop. The icon should look like the following.

2.3 Uninstalling the Software

To uninstall the software, the user should execute these 3 steps:

a. Open the directory in which the software was installed. The default directory is c:\Program Files\EPA PMF 1.

b. Double-click on UNWISE.EXE. Follow the directions for performing a typical uninstallation.

c. Once UNWISE has uninstalled the software, use Windows Explorer to delete the directory in which the software was installed. For the default settings, this means delete the directory c:\Program Files\EPA PMF 1. Beware that if the user has stored data from his/her projects in this folder, the data will be deleted. As a result, it is strongly recommended that the user not store data and output files in the directory or subdirectories where EPA PMF 1.1 is installed. Instead, the user should store data and output files in directories that are unrelated to the software directory.

2.4 Reinstalling the Software

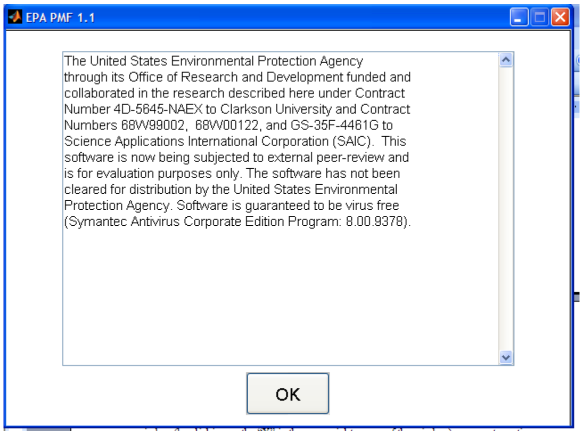
If the user has previously installed and uninstalled the software, when the user reinstalls the software, a screen about repairing or removing Matlab's Component Runtime Libraries may appear. If this dialog does appear, select "repair" and proceed. This will reinstall the libraries to your computer.

2.5 Data Management Suggestions

The input files that the user generates for EPA PMF and the output files generated by EPA PMF do not need to be stored in the same location as the software. In fact, it is strongly recommended that the input and output files not be stored in the directory where the software is installed. It is suggested that the user create a separate directory for each project and that the input files, output files, modeling details, and graphics be stored in these project-specific directories. These project-specific directories should not be subdirectories of the directory where EPA PMF was installed. For example, the project-specific directories should not be subdirectories of c:\Program Files\EPA PMF 1.

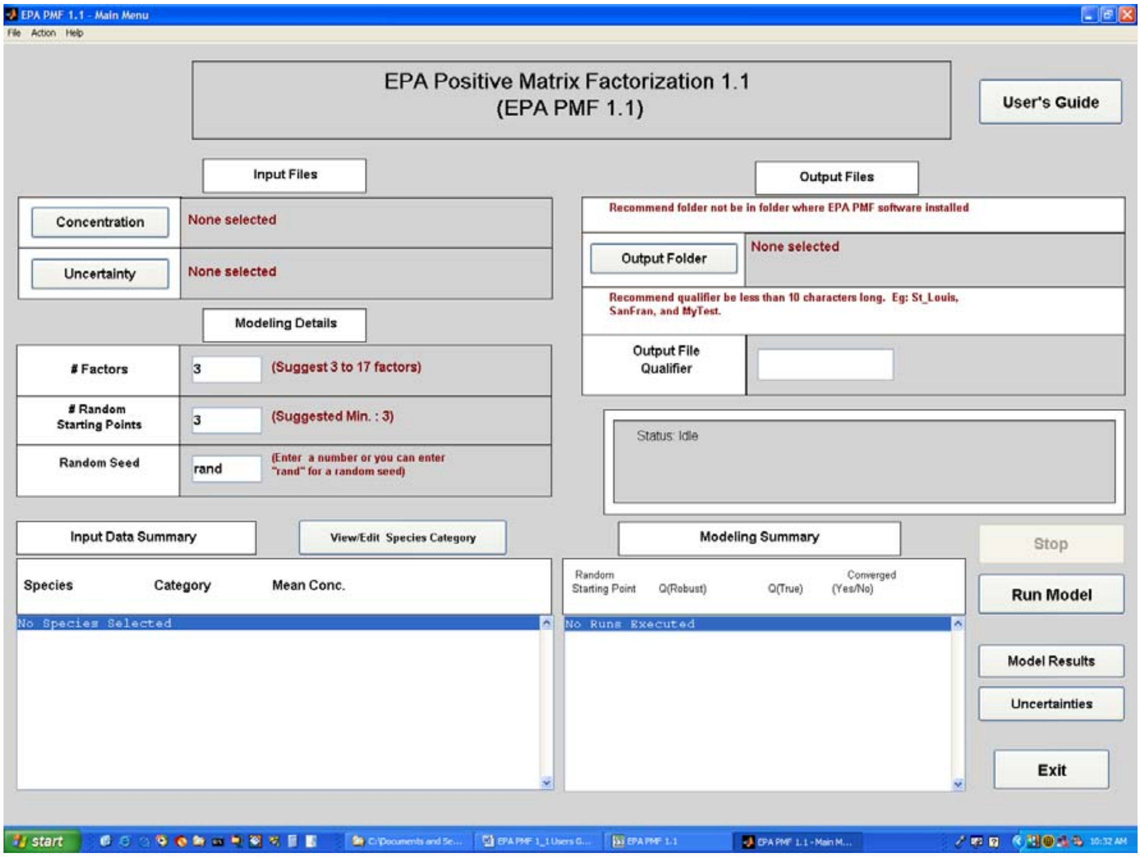
Section 3. Basic Running of EPA PMF

To execute the software, double-click on the EPA PMF 1.1 icon that has been installed on the desktop. A black, DOS window will open. The user may minimize the DOS window (not possible on some computer configurations). However, the user should not close the DOS window (by clicking on the "X" in the upper right corner of the window) now or at any time during the execution of the software as this will terminate EPA PMF 1.1.



The United States Environmental Protection Agency through its Office of Research and Development funded and collaborated in the research described here under Contract Number 4D-5645-NAEX to Clarkson University and Contract Numbers 68W99002, 68W00122, and GS-35F-4461G to Science Applications International Corporation (SAIC). This software is now being subjected to external peer-review and is for evaluation purposes only. The software has not been cleared for distribution by the United States Environmental Protection Agency. Software is guaranteed to be virus free (Symantec Antivirus Corporate Edition Program: 8.00.9378).

After a few moments, a second window automatically opens. It contains the disclaimers for the software, as shown. Press "OK" to continue. The main EPA PMF menu appears, as shown in Figure 2. Each section of the menu is described in the following sections. Note that in the upper right corner of the Main Menu there is a "User's Guide" button. Press this button at any time to access this user's manual. The manual is also accessible from the Help drop-down menu.



EPA Positive Matrix Factorization 1.1
(EPA PMF 1.1)

Concentration: None selected

Uncertainty: None selected

Recommend folder not be in folder where EPA PMF software installed

Output Folder: None selected

Recommend qualifier be less than 10 characters long. Eg: St_Louis, SanFran, and MyTest.

Output File Qualifier:

Factors: 3 (Suggest 3 to 17 factors)

Random Starting Points: 3 (Suggested Min. : 3)

Random Seed: rand (Enter a number or you can enter "rand" for a random seed)

Status: Idle

Species	Category	Mean Conc.
No Species Selected		

Random Starting Point	Q(Robust)	Q(True)	Converged (Yes/No)
No Runs Executed			

Figure 2. EPA PMF Main Menu

Note that there are example data sets located in the Data subdirectory of the directory in which EPA PMF 1.1 was installed.

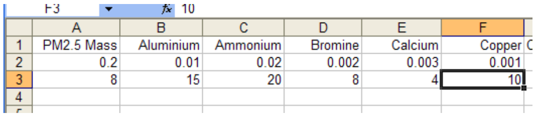
3.1 *Inputting Data*

The user must specify two input files, one file with the concentrations and one with the uncertainties associated with those concentrations. The format of the input files may be delimited files (comma, tab, space) or Excel spreadsheets.

The concentration file should be organized such that each row represents a sample (for examples, samples may be hourly, 3 hours, or 24-hours). The first row should list the species names and the subsequent rows should contain the concentrations, one row for each sample. Thus the number of rows equals the number of samples in the data set plus 1, the one extra row for the species names. The columns of the concentration file represent the various species analyzed in each sample. The concentration file may include dates, however, the dates must be in the first column. Also, files with dates should be sorted in ascending sequence by date (first row has data for oldest day, second row has data for second oldest day). See the stn_conc.csv file in the Data subdirectory for an example of a concentration file that contains dates and is in a comma-separated format.

The uncertainty file may be provided in one of two ways. An uncertainty can be specified for each species for each sample. That is, the uncertainty file can have the same number of rows and columns as the concentration file. Alternatively, the user can specify the method detection limit and a percentage for each of the species. In both cases, the first row of the uncertainty file should contain the species names and the order of the species must be the same as the order on the concentration file.

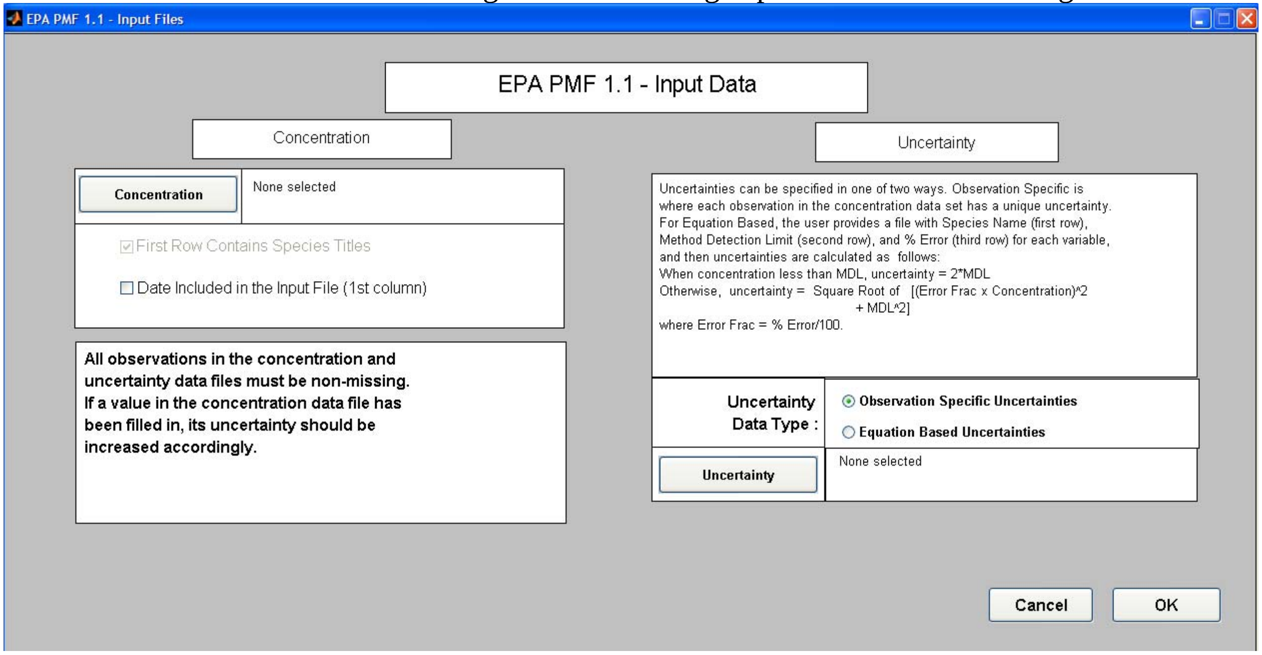
If the user chooses to specify the method detection limit and a percentage uncertainty for each species, then the user should construct a file (delimited or Excel spreadsheet) where the first row contains the species names, the second row contains the method detection limit (MDL) for each of the species, and the third row contains a percentage (e.g. 8 for 8 percent) uncertainty for each of the species. The number of columns in this file should equal the number of species in the concentration file. A partial example spreadsheet is given below showing that for PM2.5 Mass, the MDL is 0.2 ug/m3 and the uncertainty is 8% and for Copper, the MDL is 0.001 ug/m3 and the uncertainty is 10%. The software then uses the MDL and percentages to calculate uncertainties to be used by EPA PMF as follows. Be certain that the units for the MDLs are the same as those for the concentrations. For example, both MDL and concentration are in ug/m3 or both are in ng/m3.



	A	B	C	D	E	F
1	PM2.5 Mass	Aluminium	Ammonium	Bromine	Calcium	Copper
2	0.2	0.01	0.02	0.002	0.003	0.001
3	8	15	20	8	4	10
4						

$$\boxed{\begin{aligned}\text{Uncertainty} &= 2\times MDL,\ \text{if concentration} \leq MDL\\ &= \sqrt{(\text{percentage}\times\text{concentration})^2+(MDL)^2}\ ,\text{if concentration} > MDL\end{aligned}}$$

To input the data, the user clicks either on the Concentration button or the Uncertainty button on the main menu shown in Figure 2. This brings up the menu shown in Figure 3.


EPA PMF 1.1 - Input Files
EPA PMF 1.1 - Input Data
Concentration
Concentration
None selected
☑ First Row Contains Species Titles
□ Date Included in the Input File (1st column)
Uncertainty
Uncertainty
Data Type :
● Observation Specific Uncertainties
○ Equation Based Uncertainties
Uncertainty
None selected
Cancel
OK

All observations in the concentration and uncertainty data files must be non-missing. If a value in the concentration data file has been filled in, its uncertainty should be increased accordingly.

Uncertainties can be specified in one of two ways. Observation Specific is where each observation in the concentration data set has a unique uncertainty. For Equation Based, the user provides a file with Species Name (first row), Method Detection Limit (second row), and % Error (third row) for each variable, and then uncertainties are calculated as follows:
When concentration less than MDL, uncertainty = 2*MDL
Otherwise, uncertainty = Square Root of [(Error Frac x Concentration)^2 + MDL^2]
where Error Frac = % Error/100.

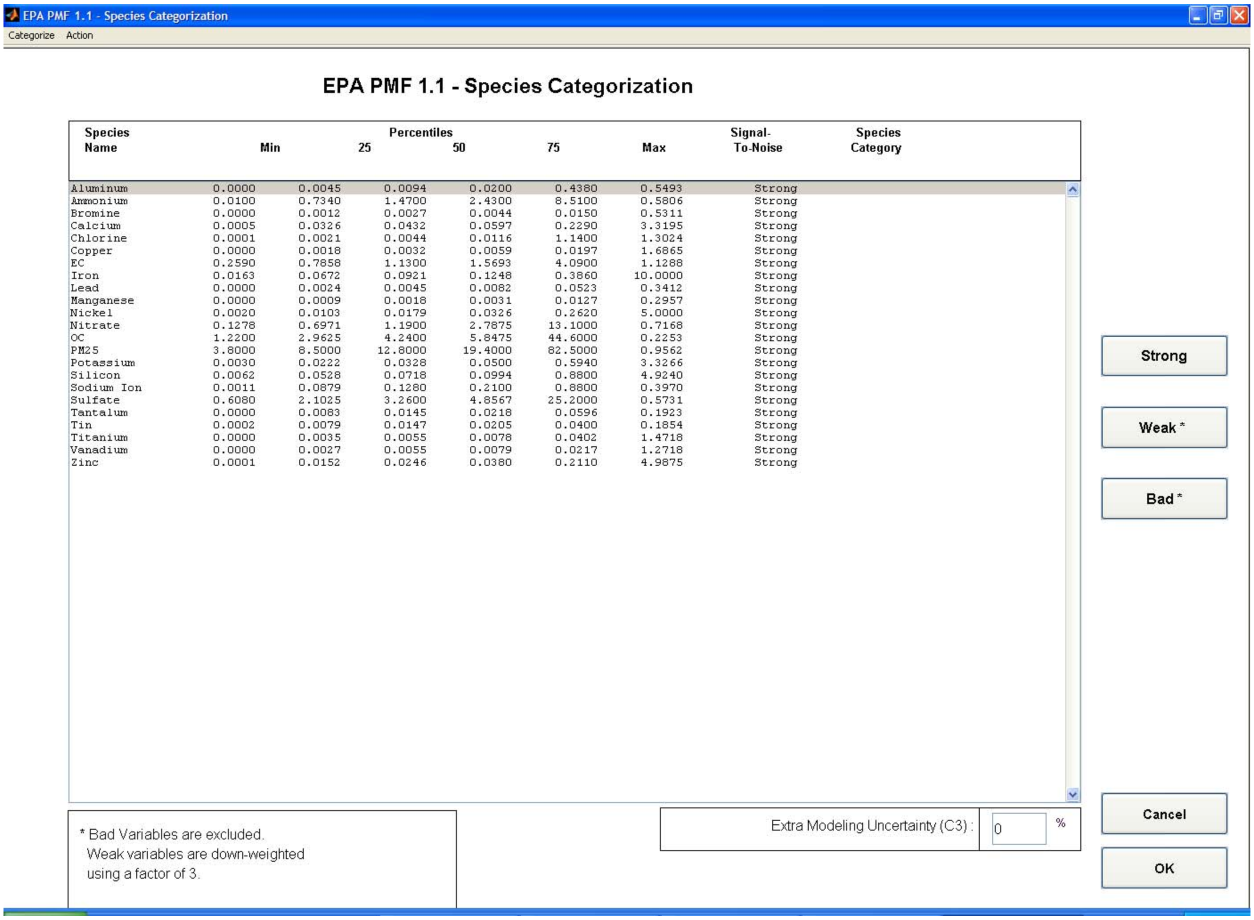
Figure 3. EPA PMF Input Data Menu

Click on the "Concentration" button on the left half of the screen to specify the file containing the concentration data. A browser window is opened from which the user can navigate to the desired concentration file. After selecting a file, the user needs to indicate whether the first column of the concentration file represents dates by selecting/deselecting the option on the Input Data menu (Figure 3).

NOTE: As stated on the Input Data menu, all data in the input files must be non-missing! This is true for the concentration and the uncertainty files.

To input the uncertainty data, from the Input Data menu (Figure 3), specify whether the uncertainties are sample-species specific (the first radio button) or are to be calculated (the second radio button). Then, as with the concentrations, click on the "Uncertainty" button and navigate to the desired file. Press "OK" on the Input Data menu once the concentration and uncertainty files have been specified.

The software brings up a page summarizing information about the concentration data. An example is provided in Figure 4. The page shows the names for each of the species, as specified by the labels provided by the user on the concentration file. For each species, a distribution of the concentrations is provided. The distribution shows the minimum, 25th percentile, median (50th percentile), 75th percentile and maximum concentration, and a "signal to noise" ratio, which is described in the next paragraph. The last column on this menu shows whether each species is Strong, Weak, or Bad, which is also described in the next paragraph. (Note that on some monitors, as shown in Figure 4, the column labels and the column data do not align. For example, the first column of numbers being reported contains the minimum value for each species; however, the word “Min” is not directly over the first column. The order of the columns is as indicated by the labels.)



EPA PMF 1.1 - Species Categorization

Species		Percentiles				Signal-	Species
Name	Min	25	50	75	Max	To-Noise	Category
Aluminum	0.0000	0.0045	0.0094	0.0200	0.4380	0.5493	Strong
Ammonium	0.0100	0.7340	1.4700	2.4300	8.5100	0.5806	Strong
Bromine	0.0000	0.0012	0.0027	0.0044	0.0150	0.5311	Strong
Calcium	0.0005	0.0326	0.0432	0.0597	0.2290	3.3195	Strong
Chlorine	0.0001	0.0021	0.0044	0.0116	1.1400	1.3024	Strong
Copper	0.0000	0.0018	0.0032	0.0059	0.0197	1.6865	Strong
EC	0.2590	0.7858	1.1300	1.5693	4.0900	1.1288	Strong
Iron	0.0163	0.0672	0.0921	0.1248	0.3860	10.0000	Strong
Lead	0.0000	0.0024	0.0045	0.0082	0.0523	0.3412	Strong
Manganese	0.0000	0.0009	0.0018	0.0031	0.0127	0.2957	Strong
Nickel	0.0020	0.0103	0.0179	0.0326	0.2620	5.0000	Strong
Nitrate	0.1278	0.6971	1.1900	2.7875	13.1000	0.7168	Strong
OC	1.2200	2.9625	4.2400	5.8475	44.6000	0.2253	Strong
PM25	3.8000	8.5000	12.8000	19.4000	82.5000	0.9562	Strong
Potassium	0.0030	0.0222	0.0328	0.0500	0.5940	3.3266	Strong
Silicon	0.0062	0.0528	0.0718	0.0994	0.8800	4.9240	Strong
Sodium Ion	0.0011	0.0879	0.1280	0.2100	0.8800	0.3970	Strong
Sulfate	0.6080	2.1025	3.2600	4.8567	25.2000	0.5731	Strong
Tantalum	0.0000	0.0083	0.0145	0.0218	0.0596	0.1923	Strong
Tin	0.0002	0.0079	0.0147	0.0205	0.0400	0.1854	Strong
Titanium	0.0000	0.0035	0.0055	0.0078	0.0402	1.4718	Strong
Vanadium	0.0000	0.0027	0.0055	0.0079	0.0217	1.2718	Strong
Zinc	0.0001	0.0152	0.0246	0.0380	0.2110	4.9875	Strong

* Bad Variables are excluded.
Weak variables are down-weighted
using a factor of 3.

Extra Modeling Uncertainty (C3) : 0 %

Figure 4. EPA PMF Species Categorization Menu

In 2003, Paatero and Hopke noted that including all species in factor analytic models may degrade the solution. Species that are always below their detection limit or species that have a lot of error in their measurements relative to the magnitude of their concentrations should not be included in such models. Paatero and Hopke suggested looking at a signal to noise ratio for each species to decide if the species was strong enough to include in the model, was bad and therefore should be severely downweighted if not removed from the modeling, or was between strong and bad (termed weak) and should be retained in the model but slightly downweighted. In EPA PMF 1.1, the user can specify whether a species is Strong, Weak, or Bad. This is done by highlighting the species (or multiple species by holding down the CTRL key) and then clicking on the Strong, Weak, or Bad buttons on the right of the Species Categorization menu (Figure 4). As stated at the bottom of this menu, species that are labeled “Bad” are removed from the analysis and species that are labeled “Weak” have their uncertainties (as provided by the user) increased by a factor of 3 prior to modeling.

The "signal-to-noise" ratio for species *j* is estimated as $\left(\frac{1}{2}\right)\sqrt{\sum_{i=1}^{n} x_{ij}^2 \Big/ \sum_{i=1}^{n} s_{ij}^2}$, where x_{ij} is the user-provided concentration for species *j* on day *i* and s_{ij} is the user-provided uncertainty for species *j* on day *i*. Smaller signal-to-noise ratios indicate that the species is more noisy. Larger signal-to-noise ratios indicate species that have more

The "Species Category," the last column in the Species Categorization menu, is set to "Strong" initially for all species. This does not mean that all species are necessarily good species to use in the modeling. The user needs to analyze the signal to noise ratios, to look at the percentage of time each species is greater than its MDL, and to use information from the lab about the quality of various species to determine which category is most appropriate for each species.

Regarding the field "Extra Modeling Uncertainty" in the lower right corner of the Species Categorization menu (Figure 4), there are several reasons that additional uncertainty could be added to the model. One reason is that the user believes that the uncertainties specified in the uncertainty file are low. A second reason is that data do not exactly meet the modeling assumptions, namely the assumption that the ratios of species in each factor (the profiles) do not vary through time. If the user has reason to believe that extra modeling uncertainty is warranted, the user can specify that additional uncertainty in the box in the lower right corner of the Species Categorization menu (Figure 4). Alternatively, the user can generate a new uncertainty matrix with the additional uncertainty (outside of EPA PMF) and input that new uncertainty matrix into EPA PMF. The amount of extra modeling uncertainty that the user can specify in the Species Categorization menu is between 0 and 25 percent. If the user wants to add 5 percent, then the user should specify "5" in the box next to "Extra Modeling Uncertainty (c3)."

Press "OK" to return to the main menu (Figure 2). The lower left area of the Main Menu, titled "Input Data Summary," will now reflect some of the information about the species, as a reminder to the user what was specified. It is easy to return to the Species Categorization menu should the user wish to review the concentration distributions or change the categorization of one or more of the species. To return to the Species Categorization menu, simply press the "View/Edit Species Category" button located above the Input Data Summary window. Alternatively, select "View/Edit Species Categorization" option from the "File" drop down menu located in the upper left corner of the Main Menu.

3.2 Selecting Modeling Details

On the Main Menu (Figure 2), there is a section called Modeling Details. In this section, the user specifies the number of factors that EPA PMF is to resolve, the number of random starting points to try, and seeds for the random number generator.

The number of factors that EPA PMF is to resolve should be greater than 2 (since 2 or fewer factors is generally unrealistic for air quality environmental data) and less than 18 (since air quality environmental data generally does not have sufficiently unique and/or variable species to infer more factors). The user specifies the number of factors by highlighting the box next to "# Factors" and typing in the desired number of factors. Currently, EPA PMF provides no suggestions about how many factors to model. As more factors are selected, the time required for finding a solution increases.

The next box in the Modeling Details is for specifying the number of random starting points that EPA PMF should try. EPA PMF uses numerical algorithms for finding a solution that minimizes Q (see Section 2), and as such, it is possible for the software to get "stuck" in a local minimum. Starting from several random starting points can show the user whether there are some solutions that appear not to be at a global minimum. Try running at least 5 random starting points, initially. If the resulting Q's are very similar, it likely is the case that all 5 solutions are near the global minimum. For a final publication or model supporting a critical decision, the user should try 20 or more random starting points to be sure that the model is finding the global minimum. See the model output sections regarding Q for more description on the interpretation of the multiple random starting points. As more random starting points are selected, the time required for completing the modeling increases. Also, if the user changes anything about the modeling (input concentrations, uncertainties, number of factors, or species categorization), the user must retest to see if there are local minima. For example, if a 5-factor solution does not appear to have local minima, this does not mean that a 6-factor solution will not have local minima.

The last box in the Modeling Details lets the user control the random number generation, if so desired. Controlling the random number generator is useful if the user wants to exactly reproduce previous results, which can be particularly helpful if two people in different locations want to perform some modeling of a common data set and discuss the results. To specify the seed used in the random number generation, highlight the box next to "Random Seed" and type a number. If the user prefers to randomly generate seeds, specify "rand" in the box next to "Random Seed."

3.3 Specifying Location and Qualifying Name for Modeling Outputs

EPA PMF generates several types of output files. One file contains the resulting profiles. Another files contains the sample-specific contributions of each factor. Yet another file contains diagnostics about the model. And several files are generated if the user asks for graphics to be saved. To minimize the number of file names specified, the user need only provide the directory into which all output files are to be stored and a qualifying name for all the output files. Then EPA PMF will generate files names based on the qualifying name and store all the results in the user-specified directory. The output files generated are as follows. Given how long the file names can become, the user-specified qualifier can not exceed 15 characters.

File name generated by EPA PMF (where "qual" is specified by user)	**Contents of File (explained in detail in Section 4)**
*qual*_profile.txt	Profile information (in units same as concentrations)
*qual*_contrib.txt	Sample-specific contributions for each factor (where average for each factor is 1)
*qual*_resid.txt	Residuals and standardized residuals from modeling

File name generated by EPA PMF (where “qual” is specified by user)	Contents of File (explained in detail in Section 4)
*qual*_diag.txt	Diagnostics from modeling, including details about modeling (number of factors, random number generator, species), Q values for each random starting point, regression results, large residuals, information about uncertainties in the modeling results
*qual*_strength.txt	Strength of each factor. All strengths equal 1, unless using ACB model.
*qual*_rrR_factF_prof.jpg *qual*_rrR_factF_contrib.jpg	Graphical representations of the F’th factor for the R’th random run. Graphs of profiles and time series of contributions.
*qual*_profiles_BOOT.txt	Detailed listing of all bootstrapping results.
*qual*_uncert_rrR_factF.jpg	Graphical representations of the uncertainty in the profile for the F’th factor in the R’th random run.

3.4 Executing EPA PMF

After specifying the input files, modeling details, and the directory and qualifier for the output files, press “Run Model” from the Main Menu (Figure 2). The Main Menu will fade, a status bar will show how many of the multiple random starting points have completed execution, and resulting robust and true Q values are summarized in the Modeling Summary window on the Main Menu. The Modeling Summary window also indicates whether each of the random runs converged. If you realize that something was incorrectly specified for the modeling runs, press “Stop” and EPA PMF will prompt to user about whether to terminate the current random start only or all random starts. It is recommended that you terminate all random starts. Unstable behavior in the software may be observed if stopping only the current random start. After pressing “Stop,” be patient as it may take a few moments before the modeling run stops.

After execution, EPA PMF must format the output, which takes a few moments. Watch the Status Bar window to see which files are being formatted. When the window displays “Status: Run Completed,” then the modeling is complete.

If any of the random runs did not converge, the user should investigate the cause of the non-convergence. It may be due to EPA PMF needing more “time” to find a solution, which can easily be arranged, or may be due to a serious problem with the user-specified input files. If a model run does not converge, no further analysis of that run is allowed.

EPA PMF finds a solution that minimizes Q using a numerical algorithm. To prevent the software from searching indefinitely for a solution, a maximum number of steps is specified. If this maximum number is exceeded, the software stops that model run and reports that the random run did not converge. The final column of the Modeling Summary window tells the user which runs converged to a minimum Q before reaching the maximum number of steps and which did not converge, either due to exceeding the number of allowable steps or for something more serious. To see if the maximum number of steps was exceeded requires reviewing two files. First, open the file containing the modeling diagnostics (*qual*_diag.txt) and locate the section showing the results for the random runs. The number of steps taken for each run is reported in the last column, as shown in the box to the right. The number of allowable steps is in the file PMF_bs_1d.ini in the directory in which EPA PMF was installed. The section of the INI file containing the maximum allowable steps is shown below and the maximum number of steps is highlighted by the red circle. If the number of steps in *qual*_diag.txt exceeds the maximum number specified in the INI file, then increase the number allowed in the INI file and rerun EPA PMF. (Additional details about modifying the INI file are in Section 5.) If the number of steps taken does not exceed the number of steps allowed and yet the run did not converge, this most likely means that there is a serious problem with the data. Please contact the software support contact listed in Section 1 should you encounter such an occurrence.

```
(Partial listing from qual_diag.txt)
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Q Values for random-start runs
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Random Run  Q(Robust) Q(True) Converged(Y/N)  # Steps
    1      28331.00     36658.20   Yes           304
    2      28364.00     36578.20   Yes           384
    3      28361.10     36569.00   Yes           287
    4      28363.60     36584.80   Yes           401
    5      28362.70     36575.50   Yes           384
    6      28365.10     36570.70   Yes           485
```

```
(Partial listing from PMF_bs_1d.ini)
% Convergence tests and other parameters for the three
% iteration levels.                  %---Reserved --
convtests                            %--for future--
0.1000,       20,        800,      0,       0,     0.001,    %level 1
0.0050,       50,       2000,      0,       0,     0.00005,   %level 2
0.0003,      100,       5000,      0,       0,     0.000005;  %level 3
% deltaQ   consecut.    max cumul.   not     not    gg2 norm
%   test      steps     step count   used    used     test
```

Note that if the user changes anything about the modeling (input concentrations, uncertainties, number of factors, or species categorization), the user must retest to see that the maximum number of steps is sufficiently large.

3.5 Saving Model Specifications

Specifying the input and output files, modeling parameters, and species categories takes time. Once the user has input this information once, the user can save these settings to a file. This file can then be loaded in the future and the user will have all the same settings previously specified. To save the model specifications, select "File" from the upper left corner of the Main Menu (Figure 2). A drop-down menu appears. Select "Save Run Profile." The user will be prompted, via a Windows Explorer window, to specify the folder and file in which to save the model specifications. To load the model specifications, all the user needs to do is select "Load Prior Run Profile" from the "File" drop down menu and specify the folder and file in which the specifications were stored. The user may view the file containing the model specifications by opening the specifications file with Excel.

Note: it generally is not possible to move the model specification file directly from one computer to another and have it work correctly. This is because the model specification file stores locations of files and rarely is the folder structure the same from one machine to the next. It is possible to open the model specifications file in Excel, modify the paths of the folders, save the file, and then load the modified version into EPA PMF.

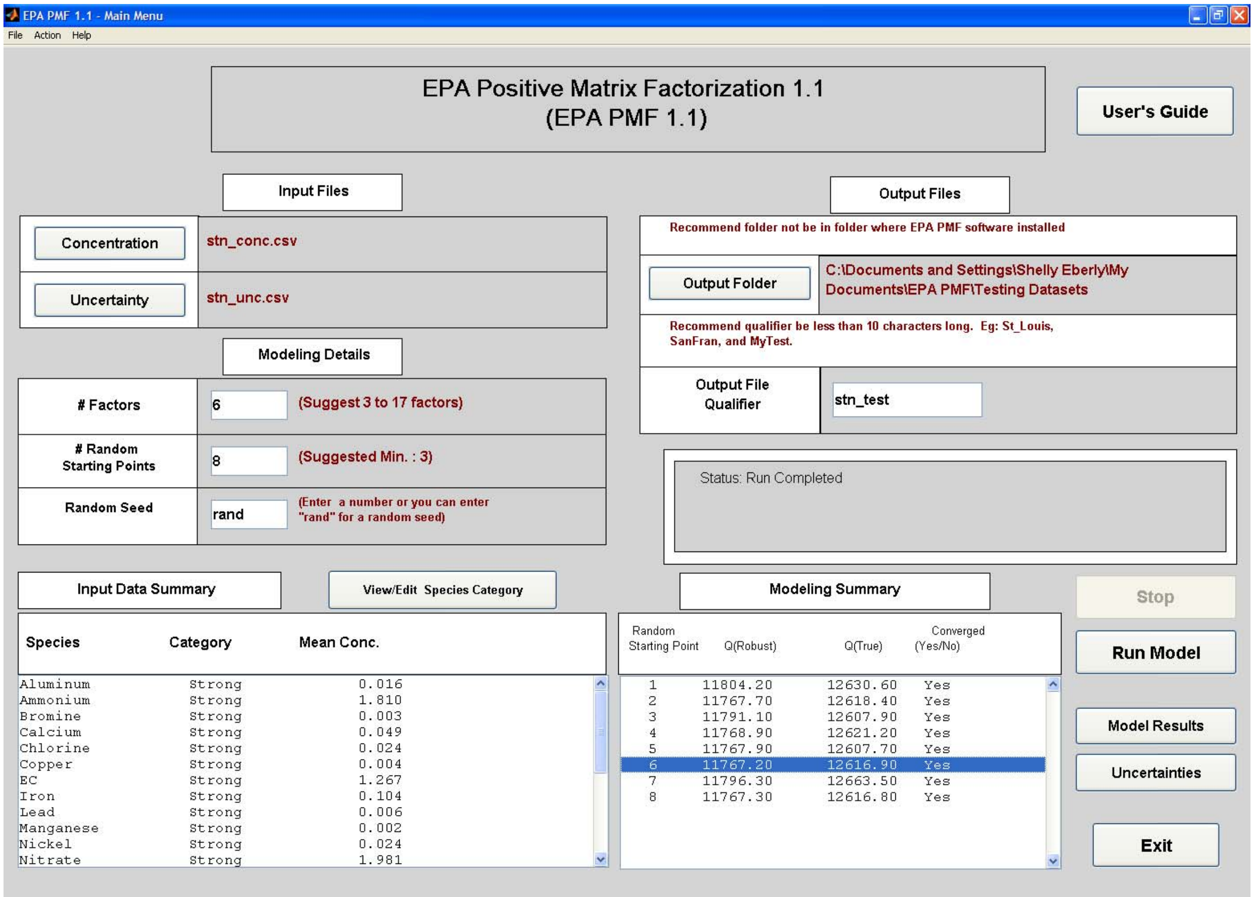
3.6 References

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

Section 4. Basic Output Produced by EPA PMF

This section describes the output produced by EPA PMF. It assumes that the software has been run at least once, if not multiple times for the random starting points. If the software has not been run, follow the instructions in Section 3 before returning to this section.

The following figure shows an example run of the software using the data stn_conc.csv and stn_unc.csv, both of which are provided in the Data subdirectory of the directory in which EPA PMF was installed. A 6-factor solution is being sought starting from 8 different random starting points.



Concentration	stn_conc.csv
Uncertainty	stn_unc.csv

Recommend folder not be in folder where EPA PMF software installed

Output Folder	C:\Documents and Settings\Shelly Eberly\My Documents\EPA PMF\Testing Datasets

Recommend qualifier be less than 10 characters long. Eg: St_Louis, SanFran, and MyTest.

Output File Qualifier	stn_test

# Factors	6	(Suggest 3 to 17 factors)
# Random Starting Points	8	(Suggested Min. : 3)
Random Seed	rand	(Enter a number or you can enter "rand" for a random seed)

Status: Run Completed

Species	Category	Mean Conc.
Aluminum	Strong	0.016
Ammonium	Strong	1.810
Bromine	Strong	0.003
Calcium	Strong	0.049
Chlorine	Strong	0.024
Copper	Strong	0.004
EC	Strong	1.267
Iron	Strong	0.104
Lead	Strong	0.006
Manganese	Strong	0.002
Nickel	Strong	0.024
Nitrate	Strong	1.981

Random Starting Point	Q(Robust)	Q(True)	Converged (Yes/No)
1	11804.20	12630.60	Yes
2	11767.70	12618.40	Yes
3	11791.10	12607.90	Yes
4	11768.90	12621.20	Yes
5	11767.90	12607.70	Yes
6	11767.20	12616.90	Yes
7	11796.30	12663.50	Yes
8	11767.30	12616.80	Yes

Figure 5. EPA PMF Main Menu after Execution

4.1 Results for Random Runs

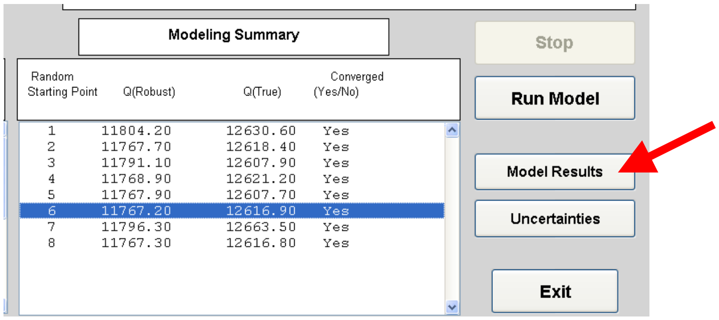
The Q values for the random starting points are displayed in the Modeling Summary section of the Main Menu (Figure 5) and they are also written to the diagnostics files. Two types of Q values are reported, one called robust and one called true. The robust Q value is the one for which the impact of outliers has been reduced. That is, no observation is allowed to have extreme influence in the fitting of the model. This should prevent over-fitting of these extreme values. The true Q value does not have the impact of the outliers reduced. Generally, using the robust Q is preferable for understanding how well the model fit the data. Only those random runs that converged should (and can) be analyzed further.

For this example, all of the random runs converged. Five of the robust Q values cluster around a minimum of 11,768 and three are scattered but larger. Based on these 8 random starts, it appears that the global minimum solution is around 11,768. The true Q's for the five clustering runs vary more, but the range is still small, indicating that the extreme values as a whole are being fitted approximately equally well for each of the five clustering solutions. Given the similarity in the robust and true Q values, the following analyses can be based on any of the five random runs. In general, the user should select the convergent solution with the smallest robust Q to continue. If there are multiple convergent solutions with the same, small robust Q, then select from these the one with the smallest true Q.

Note that the robust Q value should be approximately equal to the number of observations in the concentration file for a model where the uncertainties are correctly specified, the number of factors is right, and the data abide by the assumptions of the model. Lots of "if's!" In general, make sure that the robust Q is in the neighborhood of where it should be, called the theoretical Q. For the example, there are 375 days (rows) and 23 species (columns) meaning there are 8625 observations in the concentration file. The robust Q's of 11,768 are about 35% larger. This may be due to too few factors being specified for this dataset and/or the data not entirely abiding by the assumptions of the model. However, the robust Q is in a reasonable range of what is expected. Had the robust Q been 200 or 100,000, the user would have known there was a serious problem, most likely with the specification of the uncertainties.

4.2 *Plots of Factor Profiles and Contributions*

After running several random starting points and determining which solutions appear to be near the global minimum, the next step is to analyze one of the solutions further. Select one of the random runs associated with the global minimum for further analysis. For this example, we select the 6th random run. To select the run, move the mouse until the cursor is in the Modeling Summary of the Main Menu and highlight the 6th line of results, as shown to the right. Once the line is highlighted, press the "Model Results" button on the right of the Main Menu. Alternatively, "Model Results" can be reached by pressing "Action" located in the upper left corner of the Main Menu (Figure 5).



Random Starting Point	Q(Robust)	Q(True)	Converged (Yes/No)
1	11804.20	12630.60	Yes
2	11767.70	12618.40	Yes
3	11791.10	12607.90	Yes
4	11768.90	12621.20	Yes
5	11767.90	12607.70	Yes
6	11767.20	12616.90	Yes
7	11796.30	12663.50	Yes
8	11767.30	12616.80	Yes

Once "Model Results" has been pressed, EPA PMF produces up to 3 types of displays. One page, titled Modeling Diagnostics, shows how well EPA PMF is reproducing the original data. The results on this page are explained in detail in the following subsection. Other pages, titled Graphical Results, include graphical displays for each of the modeled profiles and their associated time series as well as annual, seasonal, and weekend/weekday summaries of the time series, assuming dates were provided on the input data sets.

Figure 6 shows the graphical display for one of the modeled factors, the top part showing the profile and the bottom showing the time series. Note that the title shows that the display is for the 4th factor of the 6th random run. The profile is displayed using two scales. The solid bars show the amount of each species apportioned to the factor where the units are the same as those on the concentration file. For the example above, the units on the concentration file are micrograms per cubic meter. Thus this factor is comprised of approximately 1.7 ug/m3 of nitrate, 1.9 ug/m3 of PM2.5, and 0.3 ug/m3 of ammonium. Use the scale on the left vertical axis of the graph to read the height of the solid bars. The compositions are also available in the profile output file.

The plot of the profiles also has stars for each of the species. The stars show what percentage of that species is apportioned to this factor. For example, the nitrate star is located close to 80% (using the scale on the right vertical axis to read the relative amount) meaning nearly 80% of all the nitrate is associated with this factor.

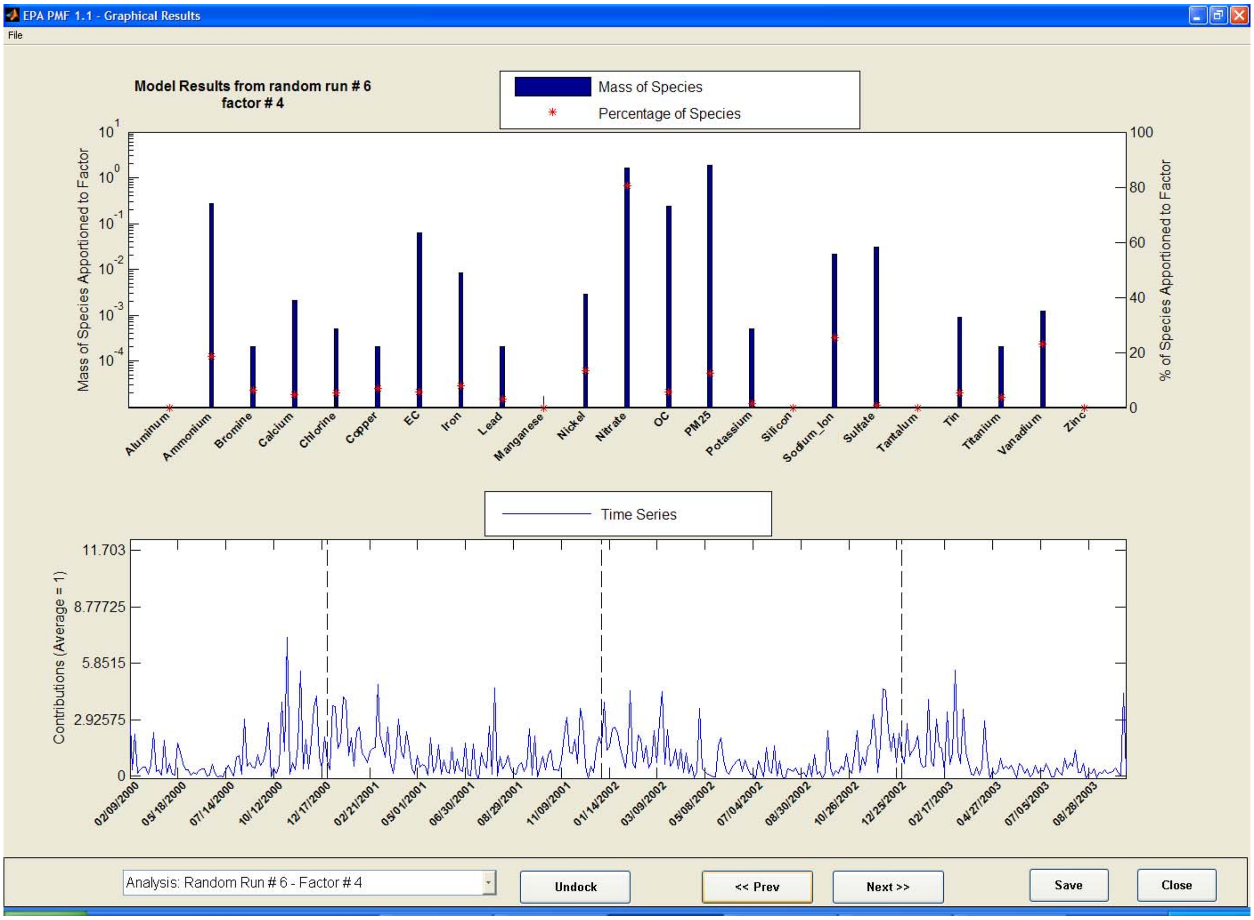


Figure 6. Example Plot of Factor Profile and Its Associated Contributions

In Figure 6, the bottom panel shows the time series of contributions associated with the factor, where the contributions average to 1. Thus a value of 3 in this plot means that the contribution for that factor to the receptor on that day is 3 times the average amount. A value of 0.5 means that the contribution is half the average amount.

If the user-provided data set includes dates, then the time series will have dates as the labels for the horizontal axis with vertical dashed lines at January 1st of each year. If the user-provided data set does not include dates, then the labels will simply be a counter from 1 to the number of observations and there will be no vertical lines separating years.

Figure 7 shows the graphical display of the summaries of the time series for 4th factor of the 6th random run. Had the input files not included dates, then this display would not have been generated. The display shows the variability in the contributions (where average contribution equals 1) by year, season, and weekend/weekday as box and whisker plots. Seasons are defined as Winter=Dec, Jan, Feb; Spring=Mar, Apr, May; Summer=Jun, Jul, Aug; and Fall=Sep, Oct, Nov. The red line in each box is the median, the box extends from the 25th to the 75th percentile, and the whiskers extend from the 10th to the 90th percentiles. The extremes (less than 10th percentile or greater than 90th percentile) are not shown since the purpose of these plots is to show the central tendencies of the contributions. The numbers in parentheses on the x-axis indicate the number of samples in each category. For the example, factor 4 looks to have slightly higher contributions in 2001, the contributions are highest in the winter, and there appears to be no difference in contributions from weekday to weekend.

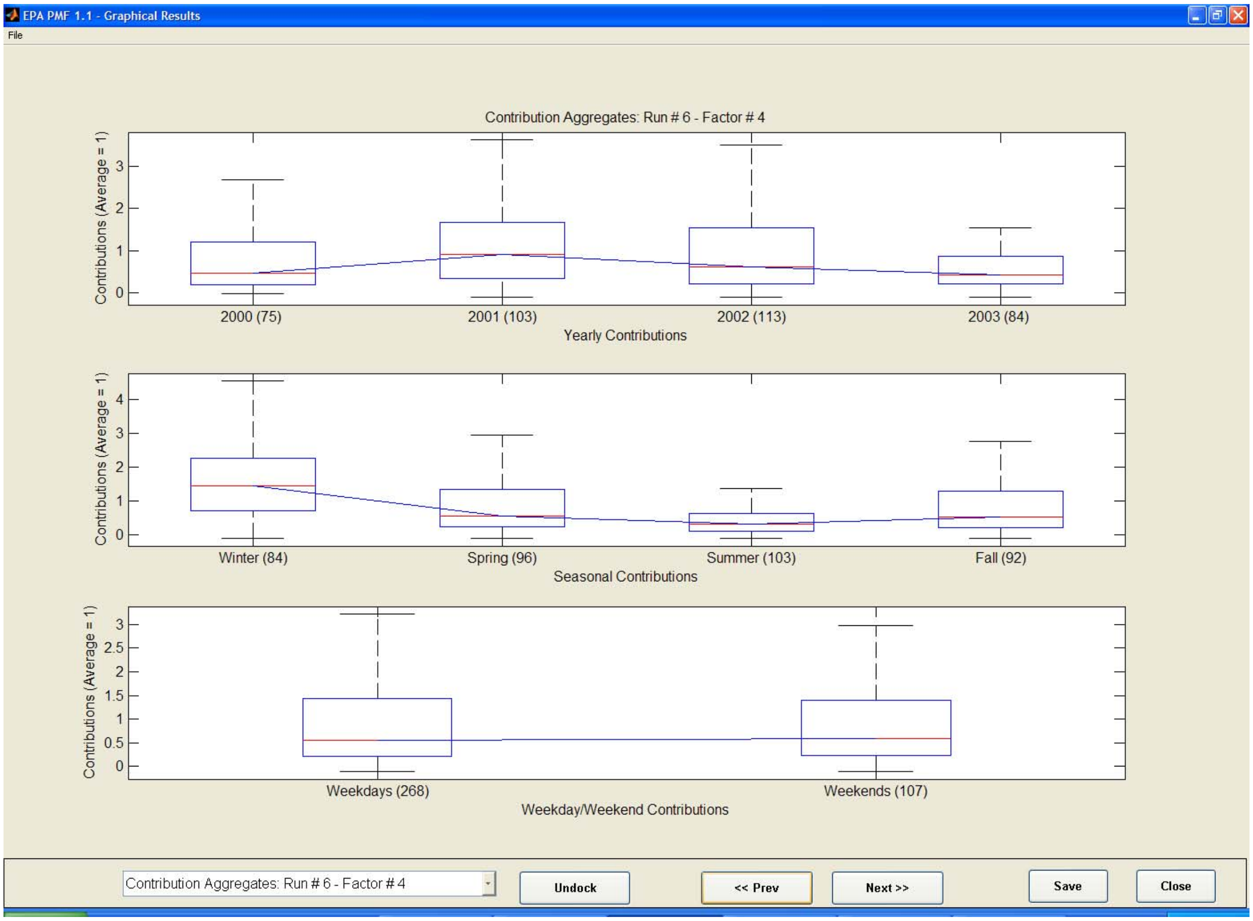

EPA PMF 1.1 - Graphical Results
File
Contribution Aggregates: Run # 6 - Factor # 4
Contributions (Average = 1)
0
1
2
3
2000 (75)
2001 (103)
2002 (113)
2003 (84)
Yearly Contributions
Contributions (Average = 1)
0
1
2
3
4
Winter (84)
Spring (96)
Summer (103)
Fall (92)
Seasonal Contributions
Contributions (Average = 1)
0
0.5
1
1.5
2
2.5
3
Weekdays (268)
Weekends (107)
Weekday/Weekend Contributions
Contribution Aggregates: Run # 6 - Factor # 4
Undock
<< Prev
Next >>
Save
Close

Figure 7. Example Plot of Summary of Contributions

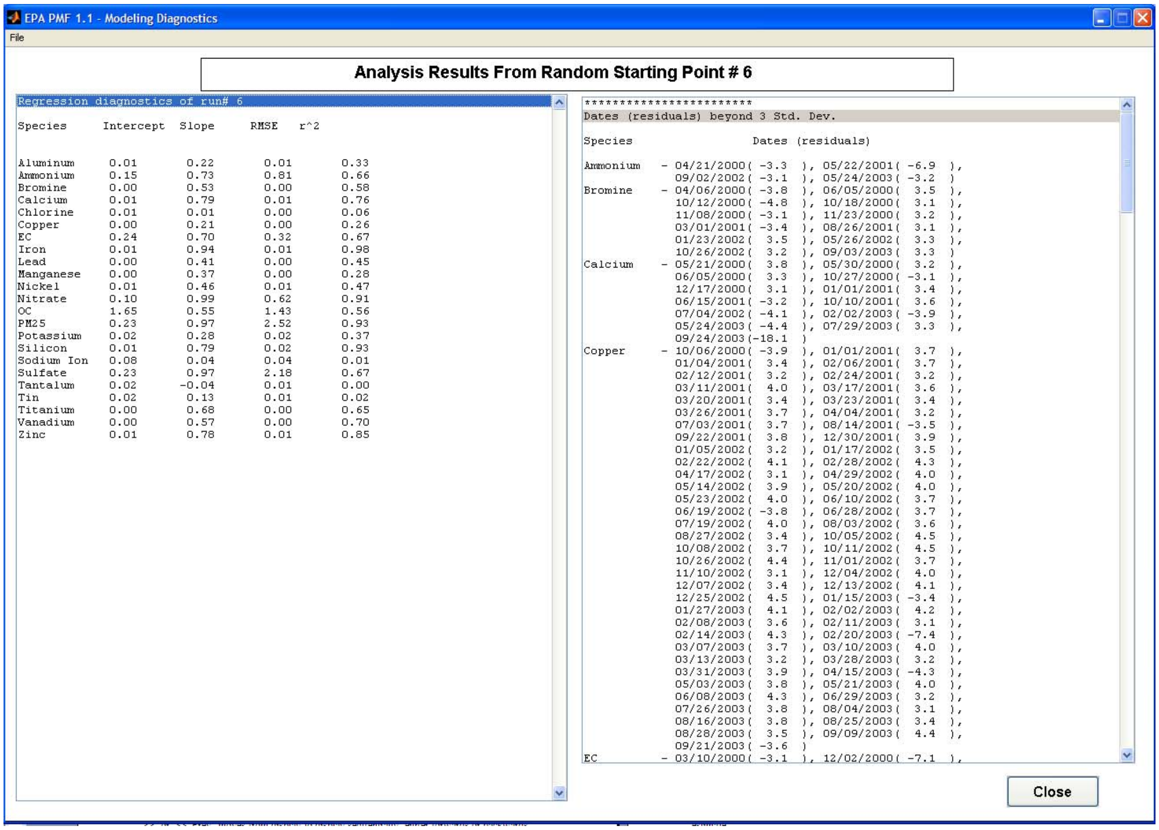
Moving from one graphical display to the next, saving displays, or closing the displays is easily accomplished. On the lower edge of Figures 6 and 7 are several buttons. Pressing "Next >>" or "<< Prev" moves from display to display sequentially, either forwards or backwards. Pressing "Save" gives the user the option to save either all plots or only the current plot in several different formats (JPEG, TIFF, or BMP). Plots are saved using the naming convention shown in section 3.3. Pressing "Close" closes all of the graphical displays.

Moving from display to display can be accomplished another way. The left portion of lower edge of Figures 6 and 7 shows the name of the current display. To the right of the name is a drop down menu that lists all of the available displays. After bringing up the drop down menu, move the cursor to the desired display and press the left mouse button. The selected display will then be showing.

The "Undock" button separates one of the displays from all of the others. This can be useful if the user wants to compare two displays. Select the first display and press the "Undock" button. Then select the second display and press the "Undock button. The two displays are now separate "pages" that the user can resize and put side-by-side.

4.3 Diagnostics for a Single Model Run

When the user selects "Model. Results," the software creates the plots of profiles and time series for each of the factors. Additionally, the software generates a page of summary diagnostics which includes a listing of the observations not fitted well by the model and regression results showing how well the model reproduces the original concentrations. The listing of observations that are not fitted well is assembled by species, to show which species are not being fitted well, and by date, to show which days are not being fitted well. Observations listed are those for which the standardized residual are larger than 3 or less than -3. The regression results show the relationship between the observed and predicted concentrations by species. Statistics include the intercept (hopefully near 0), slope (hopefully near 1), root mean



Analysis Results From Random Starting Point # 6

```
Regression diagnostics of run# 6

Species      Intercept  Slope      RMSE    r^2


Aluminum      0.01       0.22       0.01       0.33
Ammonium      0.15       0.73       0.81       0.66
Bromine       0.00       0.53       0.00       0.58
Calcium       0.01       0.79       0.01       0.76
Chlorine      0.01       0.01       0.00       0.06
Copper        0.00       0.21       0.00       0.26
EC            0.24       0.70       0.32       0.67
Iron          0.01       0.94       0.01       0.98
Lead          0.00       0.41       0.00       0.45
Manganese     0.00       0.37       0.00       0.28
Nickel        0.01       0.46       0.01       0.47
Nitrate       0.10       0.99       0.62       0.91
OC            1.65       0.55       1.43       0.56
PM25          0.23       0.97       2.52       0.93
Potassium     0.02       0.28       0.02       0.37
Silicon       0.01       0.79       0.02       0.93
Sodium Ion    0.08       0.04       0.04       0.01
Sulfate       0.23       0.97       2.18       0.67
Tantalum      0.02      -0.04       0.01       0.00
Tin           0.02       0.13       0.01       0.02
Titanium      0.00       0.68       0.00       0.65
Vanadium      0.00       0.57       0.00       0.70
Zinc          0.01       0.78       0.01       0.85
```

```
*************************
Dates (residuals) beyond 3 Std. Dev.

Species                  Dates (residuals)

Ammonium     - 04/21/2000( -3.3  ), 05/22/2001( -6.9  ),
               09/02/2002( -3.1  ), 05/24/2003( -3.2  )
Bromine      - 04/06/2000( -3.8  ), 06/05/2000(  3.5  ),
               10/12/2000( -4.8  ), 10/18/2000(  3.1  ),
               11/08/2000( -3.1  ), 11/23/2000(  3.2  ),
               03/01/2001( -3.4  ), 08/26/2001(  3.1  ),
               01/23/2002(  3.5  ), 05/26/2002(  3.3  ),
               10/26/2002(  3.2  ), 09/03/2003(  3.3  )
Calcium      - 05/21/2000(  3.8  ), 05/30/2000(  3.2  ),
               06/05/2000(  3.3  ), 10/27/2000( -3.1  ),
               12/17/2000(  3.1  ), 01/01/2001(  3.4  ),
               06/15/2001( -3.2  ), 10/10/2001(  3.6  ),
               07/04/2002( -4.1  ), 02/02/2003( -3.9  ),
               05/24/2003( -4.4  ), 07/29/2003(  3.3  ),
               09/24/2003(-18.1  )
Copper       - 10/06/2000( -3.9  ), 01/01/2001(  3.7  ),
               01/04/2001(  3.4  ), 02/06/2001(  3.7  ),
               02/12/2001(  3.2  ), 02/24/2001(  3.2  ),
               03/11/2001(  4.0  ), 03/17/2001(  3.6  ),
               03/20/2001(  3.4  ), 03/23/2001(  3.4  ),
               03/26/2001(  3.7  ), 04/04/2001(  3.2  ),
               07/03/2001(  3.7  ), 08/14/2001( -3.5  ),
               09/22/2001(  3.8  ), 12/30/2001(  3.9  ),
               01/05/2002(  3.2  ), 01/17/2002(  3.5  ),
               02/22/2002(  4.1  ), 02/28/2002(  4.3  ),
               04/17/2002(  3.1  ), 04/29/2002(  4.0  ),
               05/14/2002(  3.9  ), 05/20/2002(  4.0  ),
               05/23/2002(  4.0  ), 06/10/2002(  3.7  ),
               06/19/2002( -3.8  ), 06/28/2002(  3.7  ),
               07/19/2002(  4.0  ), 08/03/2002(  3.6  ),
               08/27/2002(  3.4  ), 10/05/2002(  4.5  ),
               10/08/2002(  3.7  ), 10/11/2002(  4.5  ),
               10/26/2002(  4.4  ), 11/01/2002(  3.7  ),
               11/10/2002(  3.1  ), 12/04/2002(  4.0  ),
               12/07/2002(  3.4  ), 12/13/2002(  4.1  ),
               12/25/2002(  4.5  ), 01/15/2003( -3.4  ),
               01/27/2003(  4.1  ), 02/02/2003(  4.2  ),
               02/08/2003(  3.6  ), 02/11/2003(  3.1  ),
               02/14/2003(  4.3  ), 02/20/2003( -7.4  ),
               03/07/2003(  3.7  ), 03/10/2003(  4.0  ),
               03/13/2003(  3.2  ), 03/28/2003(  3.2  ),
               03/31/2003(  3.9  ), 04/15/2003( -4.3  ),
               05/03/2003(  3.8  ), 05/21/2003(  4.0  ),
               06/08/2003(  4.3  ), 06/29/2003(  3.2  ),
               07/26/2003(  3.8  ), 08/04/2003(  3.1  ),
               08/16/2003(  3.8  ), 08/25/2003(  3.4  ),
               08/28/2003(  3.5  ), 09/09/2003(  4.4  ),
               09/21/2003( -3.6  )
EC           - 03/10/2000( -3.1  ), 12/02/2000( -7.1  ),
```

squared error (to show an estimate of the variability in the concentrations after accounting for the linear relationship), and r-square (hopefully larger than 0.6).

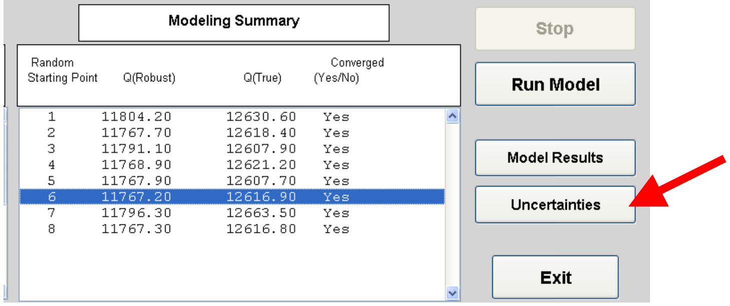
The information in the Modeling Diagnostics window is automatically written to the diagnostics file. If the user wants to save the information to an additional location, move the cursor somewhere onto the page and press the left-mouse button so that the cursor is actively in the left or right pane of the diagnostics window. When the left-mouse button is pressed, a line on the page will be highlighted. Select the "File" button from the upper left of the diagnostics window. Select "Save All" and specify the location and file name for where the information is to be stored. Note that the regression results ("left pane") and the large residuals ("right pane") must be saved separately.

Reviewing these diagnostics may suggest to the user species that should not be used in the analysis or should be down-weighted. The diagnostics may also reveal anomalous days that possibly should be removed from the analysis. To remove or downweight species, all the user need do is return to the main menu, select "View/Edit Species Category," and change the category for the poorly fitted species from "Strong" to "Weak" or from "Weak" to "Bad." After changing the categories, simply rerun the model. Note, that the poor fitting of a species (as indicated by large residuals or poor regression results) may also be an indication that the user-specified uncertainties for that species are not appropriate. Do not modify the uncertainties simply to make the model fit, but do review the uncertainties to be sure they reflect a reasonable estimate.

If the diagnostics indicate that there is a day or there are multiple days that are anomalous compared to the other days, then the user must generate new input files of concentrations and uncertainties where the anomalous days have been removed. There is no way in the current version of EPA PMF to downweight or remove days from the input files.

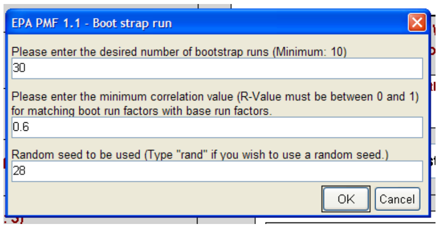
4.4 *Estimates of Uncertainties Based on Bootstrapping*

The profiles presented in Section 4.2 do not show error bars, that is, they show only point estimates but not how much reproducibility there is in those point estimates. Uncertainties in the EPA PMF solution are estimated using a bootstrapping technique combined with a method to account for the rotational freedom in the solution. To estimate the uncertainties in a solution, highlight (using the cursor) the solution of interest in the Model Summary area of the Main Menu. Once a solution is highlighted, press the "Uncertainties" button located to the right.



Random Starting Point	Q(Robust)	Q(True)	Converged (Yes/No)
1	11804.20	12630.60	Yes
2	11767.70	12618.40	Yes
3	11791.10	12607.90	Yes
4	11768.90	12621.20	Yes
5	11767.90	12607.70	Yes
6	11767.20	12616.90	Yes
7	11796.30	12663.50	Yes
8	11767.30	12616.80	Yes

After pressing the "Uncertainties" button, a menu pops up asking the user to specify the number of bootstraps desired, the minimum required correlation to say a bootstrapped factor is similar to a factor from the random run, and the seed to use to start the random number generation to be used for the bootstrapping. Regarding the number of bootstraps, the default value is 30. This is small enough that the bootstrapping runs reasonably fast and yet is large enough that the results are generally indicative of the underlying uncertainties. For publications



Please enter the desired number of bootstrap runs (Minimum: 10)
30

Please enter the minimum correlation value (R-Value must be between 0 and 1) for matching boot run factors with base run factors.
0.6

Random seed to be used (Type "rand" if you wish to use a random seed.)
28

or decisions requiring a high level of confidence in the uncertainty estimates, it is recommended that minimally 100 bootstraps be run, and preferably, more like 200-500 be run. This large number of bootstraps will take a long time to run, so use a smaller number (like 30 or 50) while in the preliminary stages of investigation of how well the model is representing the input data. Regarding correlation, see the next the section for an explanation of the correlation. The default value is 0.6 The appropriate correlation will depend on the type of data being analyzed and the decision being supported by the analysis. Regarding the seed for the random number generator, just as with the random runs, the idea is that if you need to exactly reproduce results, specify a numeric seed. The seed need not be related to the seed selected for the random runs. If exact reproducibility is not essential, then let the seed be randomly generated by specifying "rand" as the seed.

Press "OK" and the software reiterates the user's specifications to a dialog box. Press "OK" if the values are as desired or press "Cancel" to cancel the bootstraps. Depending on the size of the input files, the number of factors in the solution, the number of bootstraps requested, and the computer on which the software is operating, bootstrapping will take several minutes or longer. If you want to watch the progress of the bootstraps, view the DOS window that was opened when EPA PMF first started. Information about the modeling is outputted to this window. Look for lines such as "when solving task number 22." This means that it is on the 22nd bootstrapping run. The total number of bootstraps executed is 1 more than what the user specified.

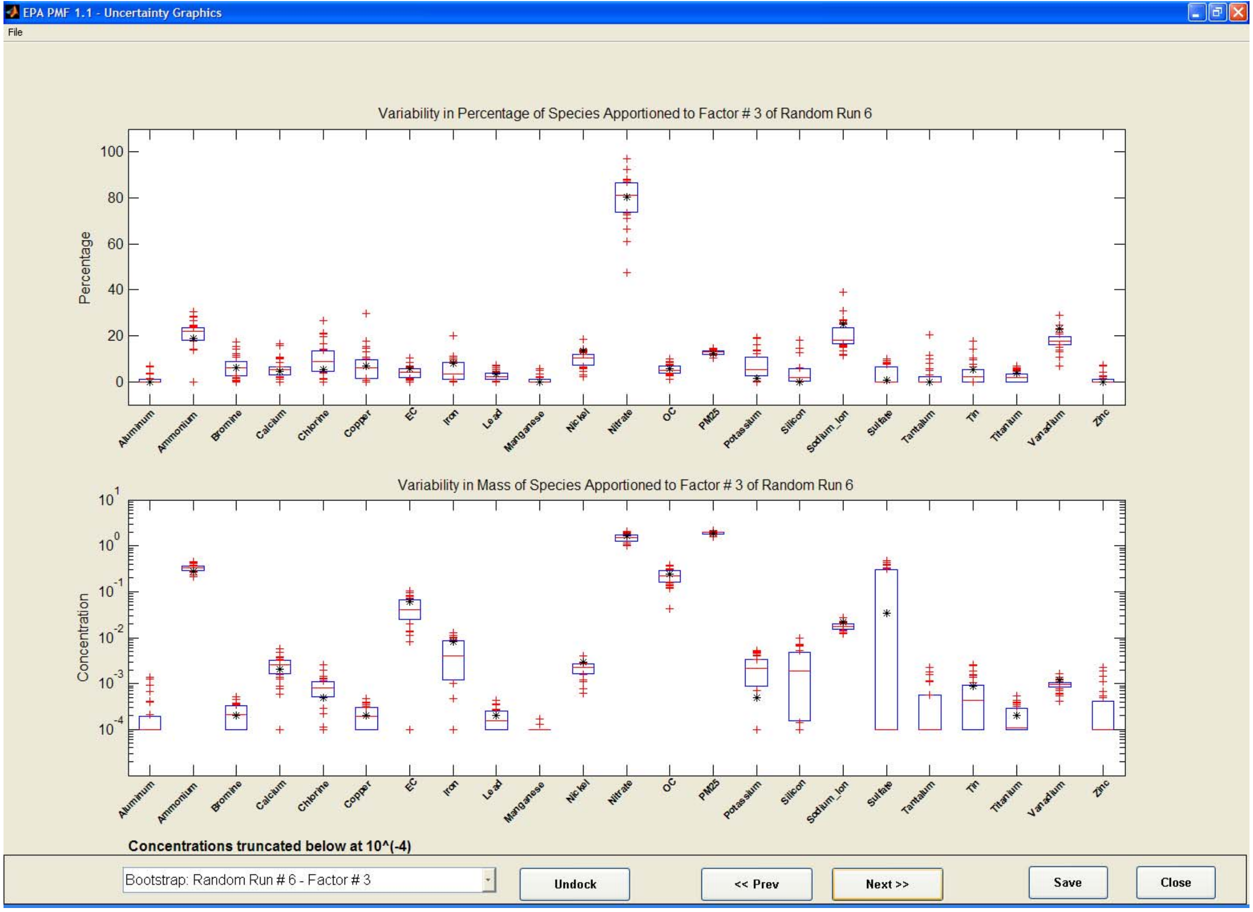
Once the bootstrapping has started, there is no way to interrupt the execution of all of the bootstraps. Once the bootstrapping has finished, a dialog box will appear stating that statistics are being calculated for the bootstrapping results. Once the statistics have been completed, EPA PMF produces 2 types of displays. One page, titled Uncertainty Summary, shows textual information about what was requested for the bootstrapping and the resulting uncertainties. The second page is a scrollable window, titled "Uncertainty Graphics," that contains graphical displays of the uncertainties in the profiles. For both displays, only those bootstraps that converged in the allowable number of steps are summarized. The number that converged are reported in the Uncertainty Summary.

NOTE: Once you close the bootstrapping graphical displays, they can not be regenerated without rerunning all of the bootstraps. So be careful that you do not close the graphical displays from bootstrapping until you have either saved them or decided that they are not needed.

4.4.1 Graphical Displays of Uncertainties from Bootstrapping

To summarize the estimated uncertainties, the bootstrap factors are mapped to the base case factors where the base case is the one the user selected on the Main Menu, like random run number 6 displayed in Figure 4.2. To do this mapping, each of the bootstrap factors is correlated with each of the base case factors where the correlation is based on the time series of contributions. The pair with the highest correlation is retained, assuming that correlation is as high or higher than that specified by the user. In this way, each of the bootstrap factors is mapped to exactly one of the base case factors. For a bootstrap run, more than one bootstrap factor may be mapped to the same base case factor. For a robust model, the number of bootstrap factors paired with each base case factor should be approximately equal to the number of bootstraps specified by the user. The number of bootstrap factors mapped to each base case factor is reported on the Uncertainty Summary page and is also written to the diagnostics file.

After pairing the bootstrap factors to the base case factors, plots of the profiles are produced. The upper panel shows the uncertainties in the percentage of species mass. To create this plot, imagine a plot like Figure 4.2 being produced for each of the bootstrap factors, gather all the red stars for aluminum, and then create a box-and-whisker plot from those red stars. Continue to do this for each species. The lower panel shows the uncertainties in the species mass and is created by making a box-and-whisker plot from the height of the blue bars. The box of the box-and-whisker plot shows where 50% of the bootstrap values lie: the narrower the box, the more consistent the results across the bootstraps and the wider the box, the less consistent the results. Twenty-five percent of the bootstrap values lie above the box and 25% below the box. These more extreme bootstraps are shown as plus signs. Again, if the plus signs are clustered and are close to the box, then there is good reproducibility. The black star that is overlaid is the value from the base case run. Ideally, the black star should lie within the box. If this is not the case, then the base case likely has some observations that are atypical and these observations are influencing the factorization.



Variability in Percentage of Species Apportioned to Factor # 3 of Random Run 6

Variability in Mass of Species Apportioned to Factor # 3 of Random Run 6

Concentrations truncated below at 10^(-4)

Bootstrap: Random Run # 6 - Factor # 3

All the bootstrap factors that do not correlate well (as specified by the user) with any base case factor are lumped together and plotted as a group. Sometimes this plot of unmapped factors can provide insights into the data if there are patterns in the unmapped factors, but often there are so few unmapped factors that no patterns can be discerned.

Uncertainties in the profiles based on bootstrapping are displayed for each factor on a separate page of a scrollable display, just as the original factors from the random runs are displayed. Thus, the user can scroll forward or backward, save the plots, close all the plots, or undock individual plots, as described extensively in Section 4.2.

Uncertainties are shown for both ways of displaying the profiles, as a percentage of species mass on the upper panel and as mass for each species on the lower panel. These two types of display are discussed in Section 4.2. The graphical display of the uncertainties associated with the nitrate factor of the 6th random run is shown in the figure above. This shows that the bootstraps support the conclusion that the majority (70-90%) of nitrate is ascribed to this factor. On a mass basis, nitrate and ammonium are present in a ratio that matches what is expected for ammonium nitrate. Also, the amounts of nitrate, ammonium, and PM2.5 ascribed to this source are very consistent, as indicated by the tight box and whisker plots in the lower plot.

NOTE: For the mass presentation of the profiles (lower panel), the masses are truncated below at 10^{-4}. That is, any masses less than 10^{-4} are set to 10^{-4} for this plot..

4.4.2 Textual Summaries of Uncertainties from Bootstrapping

The textual summary of the uncertainties from bootstrapping are shown on the "Uncertainty Summary" page generated by EPA PMF and are written to the diagnostics files. The summary shows (a) how many bootstraps were requested, required correlation, random seed, and how many bootstraps did not converge and are therefore not summarized, (b) how many bootstrap samples were paired to each of the base case factors and how many could not be paired, (c) the distribution of the robust Q's, (d) variability in the strengths of each factor, and (d) estimates of uncertainties for each species in each factor of the base case.

While the information in the Uncertainty Summary window is automatically written to the diagnostics file, it is possible for the user to save the information to an additional file. To do this, move the cursor somewhere onto the page and press the left-mouse button so that the cursor is actively in the window. When the left-mouse button is pressed, a line on the page will be highlighted. Select the "File" button from the upper left of the Uncertainty Summary window. Select "Save All" and specify the location and file name for where the information is to be stored.

4.5 References

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, Florida.

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

Section 5. More Advanced Running of EPA PMF

This section describes some advanced options for the running of EPA PMF. These options are not necessarily difficult but they do require a solid understanding of what EPA PMF is solving.

Most of the advanced options are implemented by making modifications to a file that contains a few parameters that control the model that EPA PMF is fitting. There is one file that contains the parameters for the random runs and another file that contains the parameters for the bootstrapping runs. *If modifications are made, be sure to change the parameters in both of these files!* The two files are located in the directory in which EPA PMF was installed (default being c:\Program Files\EPA PMF 1) and are named "iniparamsRandRun.txt" for controlling the random runs and "iniparamsBootstrap.txt" for controlling the bootstrapping.

A few of the advanced options are implemented by making modifications to a file that contains all the details about the model, minus those in the iniparams files. This file is also located in the directory in which EPA PMF was installed and is named "PMF_bs1d.ini." There is only one of these files, whether running random runs or bootstrapping.

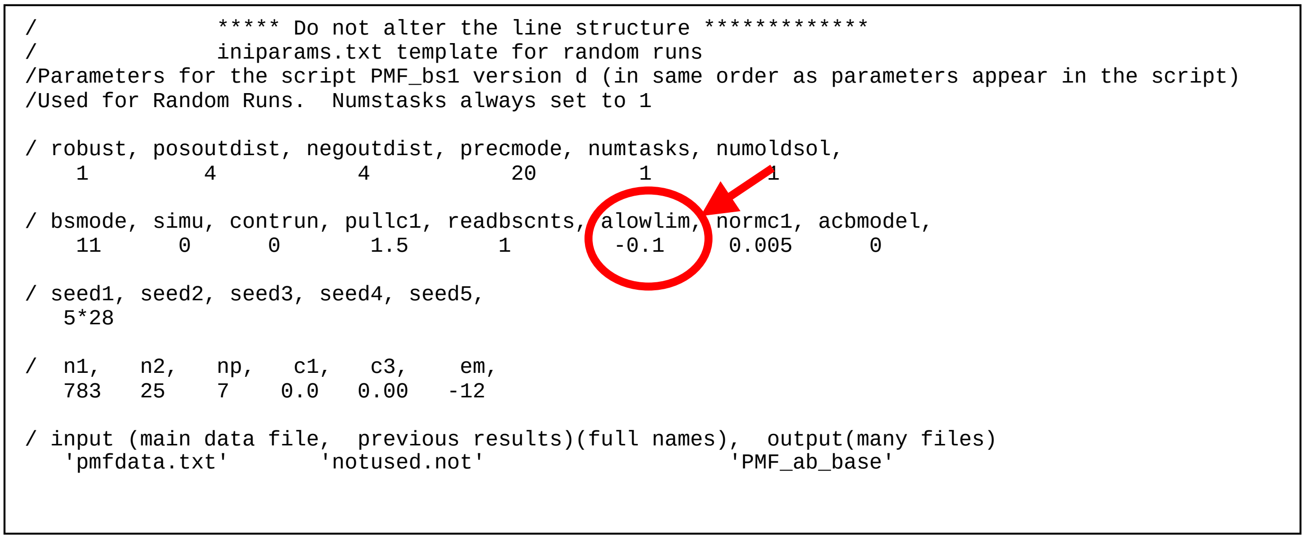
In the discussions below, when the user is instructed to change a value in a file, the user should open the specified file(s) using a text editor like Notepad or Ultraedit. Do not use a word processing package like Word or Word Perfect. Once the edits have been made, save the file back to the same name, close the editor, and restart EPA PMF. Please be careful when editing these files. Only edit the specific areas identified unless you have read the more detailed user's manuals and understand the impact of your edits.

5.1 Comments for Those Used to Running PMF2

If you are used to running PMF2, one of the options you may have used often is FPEAK. FPEAK is a simple way to rotate the solution such that the entries in the time series matrix are made more extreme (driven to zero or large values) or entries in the profile matrix are made more extreme (again, driven to zero or large values). The concept of FPEAK does not currently exist in EPA PMF because EPA PMF is based on ME-2 and ME-2 does not have FPEAK. There is no way to control or specify rotations in this version of EPA PMF, although it is possible to see how much rotational freedom there is in the solution by using the pulling parameter in the bootstrapping, described below. It is hoped that future versions will allow either "automatic" rotations, such as what FPEAK does in PMF2, or "user-specified" rotations, to be used if the user knows something about the profiles or contributions.

5.2 Requiring Contributions in Time Series Be Strictly Non-Negative

EPA PMF 1.1, as distributed, allows the predicted contributions to be slightly negative, down to -0.1 to be exact. Based on simulated data sets, it has been found that allowing slightly negative values results in more exact replications of the simulated data, especially with respect to confidence levels from the bootstrapping. If the user wants the contributions to be strictly non-negative, this can easily be achieved by changing a parameter in the "iniparams" files mentioned in the introduction to this section. Set the parameter "alowlim" to lowest value that you want the contributions to be. Thus if you want them to be non-negative, change -0.1 to 0, save the file, close your editor, restart EPA PMF, and reanalyze your data. You will see that the predicted contributions are no smaller than what you specified for "alowlim."



```
/              ***** Do not alter the line structure *************
/              iniparams.txt template for random runs
/Parameters for the script PMF_bs1 version d (in same order as parameters appear in the script)
/Used for Random Runs.  Numstasks always set to 1

/ robust, posoutdist, negoutdist, precmode, numtasks, numoldsol,
    1         4           4           20        1         1

/ bsmode, simu, contrun, pullc1, readbscnts, alowlim, normc1, acbmodel,
    11      0      0       1.5       1        -0.1     0.005      0

/ seed1, seed2, seed3, seed4, seed5,
   5*28

/  n1,   n2,   np,   c1,   c3,    em,
   783   25    7    0.0   0.00   -12

/ input (main data file,  previous results)(full names),  output(many files)
   'pmfdata.txt'       'notused.not'                   'PMF_ab_base'
```

5.3 *Robust Mode and Changing the Definition of Outlier for Robust Calculations*

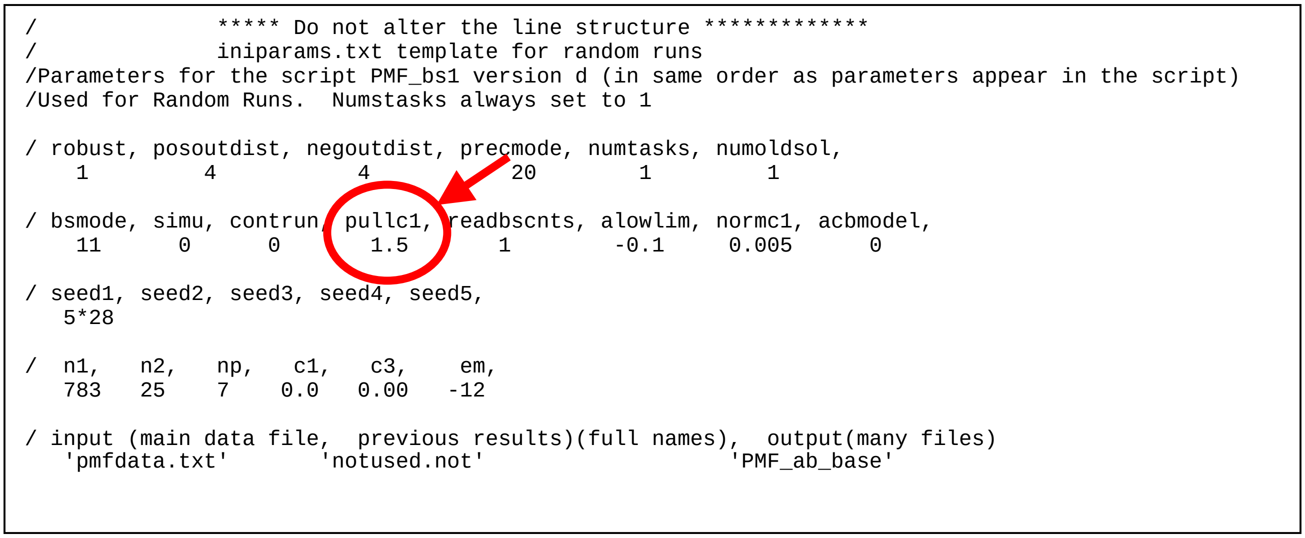
It is possible to run EPA PMF either in robust mode or in non-robust mode. It is recommended for environmental data that robust mode only be used. In robust mode, predicted values that are extremely far from the observed values are not permitted to unduly influence the fitting. [Need to include equation for outlier.]

To operate in robust mode, the "robust" variable in the iniparams files (both for the Random Runs and for the Bootstrapping) should be set to 1. To operate in non-robust mode, the variable should be set to 0.

To change the definition of an outlier for the robust calculations, change the variables "posoutdist" and "negoutdist" in the iniparams files (both for the Random Runs and for the Bootstrapping). The definition of outlier can be different for the positive direction versus the negative direction, if the user so wishes. The default value specified in EPA PMF is 4 for both the positive and negative direction.

5.4 *Increasing Uncertainty due to Rotational Ambiguity*

The method for estimating uncertainty due to rotational ambiguity has to do with randomly selecting some entries from the fitted time series and/or the profiles and seeing if it is possible to pull these entries to smaller values or to larger values. If pulled very hard, the entries will certainly change but the penalty for such a change will be large. Alternatively, if not pulled hard enough, the full potential for rotation may not be realized. The amount of pulling is controlled by the variable “pullc1” in the file “iniparamsBootstrap.txt” located in the folder in which EPA PMF 1.1 was installed. The value of pullc1 has been set to 1.5 which indicates mild pulling. Smaller values of “pullc1” pull stronger. Meaningful trial values might be between 0.6 and 1.5, but in all cases, “pullc1” must be greater than 0. It is unknown at this time as to how strongly a rotation should be pulled to estimate the uncertainty due to rotational ambiguity.

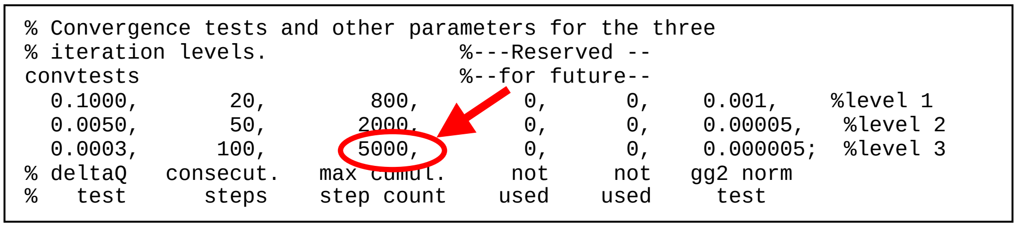


```
/               ***** Do not alter the line structure *************
/               iniparams.txt template for random runs
/Parameters for the script PMF_bs1 version d (in same order as parameters appear in the script)
/Used for Random Runs.  Numstasks always set to 1

/ robust, posoutdist, negoutdist, precmode, numtasks, numoldsol,
    1         4           4           20        1         1

/ bsmode, simu, contrun, pullc1, readbscnts, alowlim, normc1, acbmodel,
    11      0      0       1.5        1         -0.1     0.005      0

/ seed1, seed2, seed3, seed4, seed5,
   5*28

/  n1,   n2,   np,   c1,   c3,    em,
   783   25    7    0.0   0.00   -12

/ input (main data file,  previous results)(full names),  output(many files)
   'pmfdata.txt'       'notused.not'                    'PMF_ab_base'
```

5.5 *Changing the Maximum Number of Steps Allowed for Convergence*

It is possible for the user to change the maximum number of steps allowed for convergence. To change this number, the user must edit the file “PMF_bs1d.ini” located in the folder in which EPA PMF 1.1 was installed.

To change the maximum number of steps, open “PMF_bs1d.ini” using a text editor like Notepad or Ultraedit. Do not use a word processing package like Word or Word Perfect. Find the section specifying the convergence tests, as shown to the right. The default maximum number of steps is 5000. If additional steps are needed, overwrite 5000 with a number large enough to reach convergence. Be sure not to delete the comma preceding or following the number 5000. Save the INI file and exit the text editor software. All future executions of EPA PMF will use the revised maximum number of steps.



```
% Convergence tests and other parameters for the three
% iteration levels.                   %---Reserved --
convtests                             %--for future--
  0.1000,       20,        800,        0,       0,     0.001,    %level 1
  0.0050,       50,       2000,        0,       0,     0.00005,   %level 2
  0.0003,      100,       5000,        0,       0,     0.000005;  %level 3
% deltaQ   consecut.   max cumul.     not      not    gg2 norm
%   test      steps    step count     used     used     test
```

Section 6. PMF Publications

Chueinta, W., Hopke, P.K., Paatero, P., 2000. Investigation of sources of atmospheric aerosol at urban and suburban residential areas in Thailand by positive matrix factorization. Atmospheric Environment 34, 3319-3329.

Huang, S., Arimoto, R., Rahn, K., 2001. Sources and source variations for aerosol at Mace Head, Ireland. Atmospheric Environment 35, 1421-1437.

Kim, E., Hopke, P.K., Edgerton, E.S., 2004. Improving source identification of Atlanta aerosol using temperature resolved carbon fractions in positive matrix factorization. Atmospheric Environment 38, 3349-3362.

Kim, E., Hopke, P.K., Edgerton, E.S., 2003. Source Identification of Atlanta Aerosol by Positive Matrix Factorization. Journal of Air & Waste Management Association 53, 731-739.

Kim, E., Hopke P.K., 2004. Source Apportionment of Fine Particles in Washington, DC, Utilizing Temperature-Resolved Carbon Fractions. Journal of Air & Waste Management Association 54, 773-785.

Kim, E., Hopke, P.K., Larson, T.V., Maykut, N., Lewtas, J., 2004. Factor Analysis of Seattle Fine Particles. Aerosol Science and Technology 38, 724-738.

Kim, E., Hopke, P.K., Paatero, P., Edgerton, E.S., 2003. Incorporation of parametric factors into multilinear receptor model studies of Atlanta aerosol. Atmospheric Environment 37, 5009-5021.

Lee, P.K.H, Brook, J.R., Dabek-Alotorzynska, E., Mabury, S.A., 2003. Identification of the Major Sources Contributing to PM2.5 Observed in Toronto. Environmental Science and Technology 37, 4831-4840.

Maykut, N.N., Lewtas, J., Kim, E., Larson, T.V., 2003. Source Apportionment of PM2.5 at an Urban IMPROVE Site in Seattle, Washington. Environmental Science and Technology 37, 5135-5142.

Poirot, R.L., Wishinski, P.R., Hopke, P.K., Polissar, A.V., 2001. Comparative Application of Multiple Receptor Methods to Identify Aerosol Sources in Northern Vermont. Environmental Science and Technology 35, 4622-4636.

Polissar, A.V., Hopke, P.K., Paatero, P., Malm, W.C., Sisler, J.F., 1998. Atmospheric aerosol over Alaska. 2. Elemental composition and sources. Journal of Geophysical Research 103, 19045-19057.

Polissar, A.V., Hopke, P.K., Poirot, R.L., 2001. Atmospheric Aerosol over Vermont: Chemical Composition and Sources. Environmental Science and Technology 35, 4604-4621.

Ramadan, Z., Eickhout, B., Song, X.-H., Buydens, L.M.C., Hopke, P.K., 2002. Comparison of Positive Matrix Factorization and Multilinear Engine for the source apportionment of particulate pollutants. Chemometrics and Intelligent Laboratory Systems 66, 15-28.

Song, X.-H., Polissar, A.V., Hopke, P.K., 2001. Sources of fine particle composition in the northeastern US. Atmospheric Environment 35, 5277-5286.

Zhou, L., Kim, E., Hopke, P.K., Stanier, C.O., Pandis, S., 2004. Advanced Factor Analysis on Pittsburgh Particle Size-Distribution Data. Aerosol Science and Technology 38, 118-132.