

Lecture of Methodology of PMF analysis / Practice of PMF analysis using demo data

Joint Capacity Building through EANET
and SATREPS

4 December 2025
Niigata and Hybrid

Keiichi Sato

Asia Center for Air Pollution Research (ACAP)

Portal Site of Positive Matrix Factorization (PMF) Model



Environmental Topics

Laws & Regulations

About EPA

Search EPA.gov



Related Topics: [Air Research](#)

CONTACT US

SHARE



Positive Matrix Factorization Model for environmental data analyses

What is the Positive Matrix Factorization Model?

EPA's Positive Matrix Factorization (PMF) Model is a mathematical receptor model developed by EPA scientists that provides scientific support for the development and review of air and water quality standards, exposure research and environmental forensics. The PMF model can analyze a wide range of environmental sample data: sediments, wet deposition, surface water, ambient air, and indoor air. EPA's PMF model reduces the large number of variables in complex analytical data sets to combinations of species called source types and source contributions. The source types are identified by comparing them to measured profiles. Source contributions are used to determine how much each source contributed to a sample. In addition, EPA PMF provides robust uncertainty estimates and diagnostics.

How does the model work?

Users of EPA's PMF model provide files of sample species concentrations and uncertainties, and the number of sources. The model calculates source profiles or fingerprints, source contributions, and source profile uncertainties. The PMF model results are constrained to provide positive source contributions and the uncertainty weighted difference between the observed and predicted species concentration is minimized. The PMF model software uses graphical user interfaces that ease data input, visualization of model diagnostics, and

<https://www.epa.gov/air-research/positive-matrix-factorization-model-environmental-data-analyses>

EPA PMF 5.0 Fundamentals and User Guide

EPA Positive Matrix Factorization 5.0 Fundamentals and User Guide

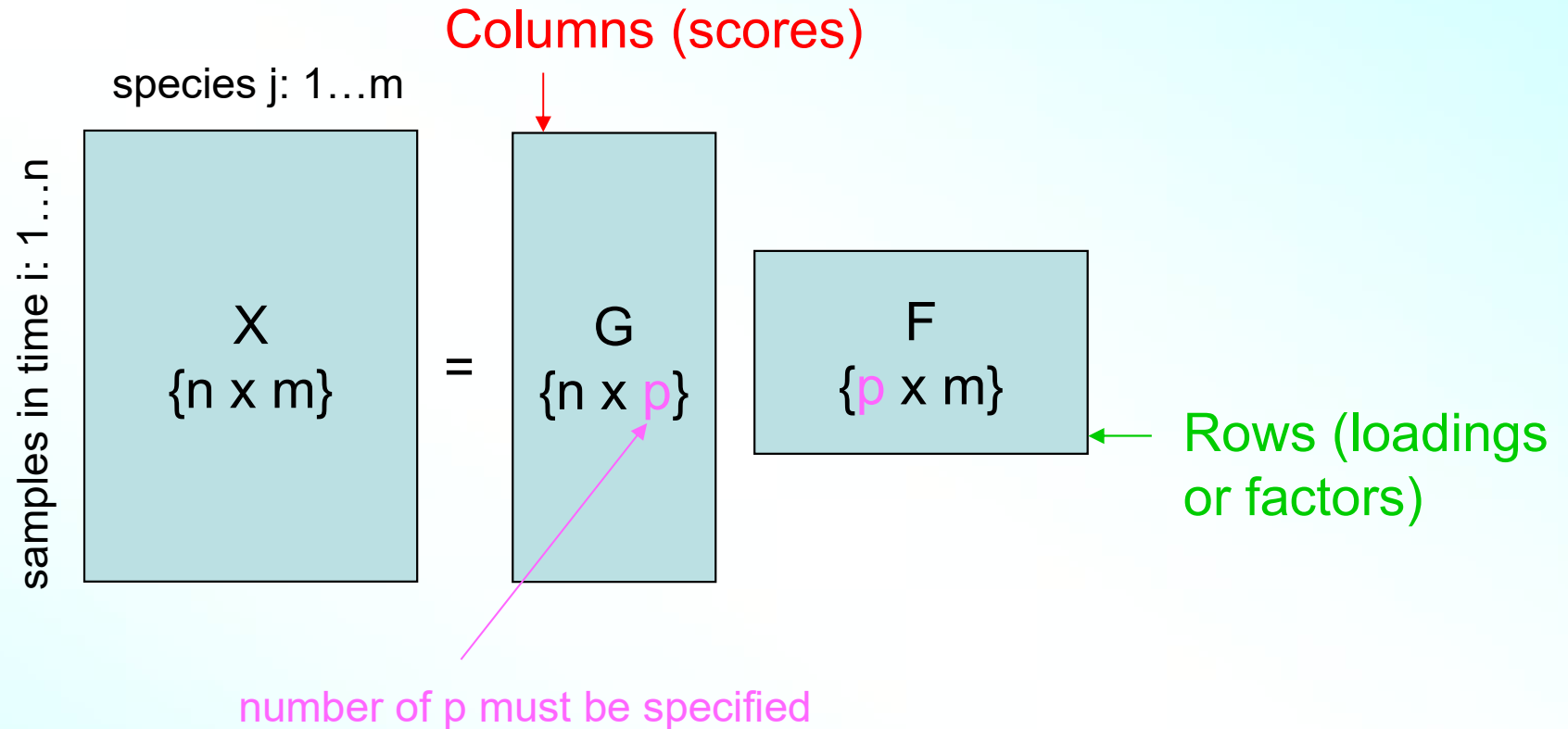
The Positive Matrix Factorization Model is a multivariate factor analysis tool that decomposes a matrix of speciated sample data into two matrices: factor contributions and factor profiles. These factor profiles need to be interpreted by the user to identify the source types that may be contributing to the sample using measured source profile information, and emissions or discharge inventories.

-  [PMF 5.0 User Guide \(pdf\)](#) (6.87 MB, April 2014, 600-R-14-108)

<https://www.epa.gov/air-research/epa-positive-matrix-factorization-50-fundamentals-and-user-guide>

Principal of Positive matrix factorization (PMF)

$X(\text{Ambient Conc.}) = G(\text{Source contribution}) \times F(\text{Conc. for each factor}) + e(\text{uncertainty})$
PMF answers “What sources are present?” and “How much did each source contribute?”



(Key Features)

- ✓ Use uncertainty associated with the sample data to weight individual points..
- ✓ Exclude negative contributions out (sources can't emit “negative” pollution).
- ✓ Handle data below detection limits by reducing their influence instead of excluding them.

How does PMF find the optimal solution?

PMF tries to minimize a number called **Q**, which measures how well the model fits the data:

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

There are two types of Q.

- **Q(true)**: Includes all data points.
- **Q(robust)**: Ignores outliers

PMF performs multiple runs of calculations using the Multilinear Engine (ME) to find the optimal solution.

- Run PMF **many times** (20 for initial evaluation, 100 for final solution).
- Solution with the lowest Q(robust) value is considered optimal

Check reliability of the PMF calculation results

PMF provides three error-checking tools:

1. Bootstrap (BS)

- Bootstrap examines how much random variation in the dataset affects the PMF results.
- It randomly selects “blocks” from the original dataset for resampling, creating multiple new datasets. It then runs the PMF on each and compares the results.
- BS can verify whether any abnormal samples are having an excessive impact on the results.

2. Displacement (DISP)

- DISP examines rotational ambiguity and flexibility of solution.
- The solutions by PMF can sometimes be mathematically “rotated.” DISP moves factor profile values up and down to determine how far they can be shifted without significantly worsening the Q-value.

3. BS-DISP (Bootstrap + Displacement)

- BS-DISP is a combination of BS and DISP. It performs DISP for each Bootstrap resample, evaluating both random error and rotational ambiguity.
- BS-DISP enables the most comprehensive and reliable uncertainty assessment and is the recommended method for reporting final solutions.

Utilization of PMF

PMF has been applied to a wide range of environmental and air quality data sets shown as follows:

1. 24-hour data of PM_{2.5} compositions

PMF helps determine potential sources (e.g., traffic, coal plants, biomass burning) contributed to the measured PM_{2.5} during a 24-hour period.

2. Size-resolved aerosol data

Different sources emit particles of different sizes. PMF can identify which sources dominate in each size range (e.g., ultrafine, fine, coarse).

3. Deposition Studies

PMF can identify sources contributing to acid rain or heavy metal deposition (Ex. Whether sulfate in rain comes from coal combustion or volcanic activity.)

4. Air Toxics Monitoring

By using measurement data of hazardous air pollutants like benzene, formaldehyde, or heavy metals, PMF evaluates apportion these pollutants to sources such as industrial plants, traffic, or solvent use.

5. High Time-Resolution Measurements

The data are collected at very short intervals (minutes or seconds) by using Aerosol Mass Spectrometers (AMS). PMF can track rapid changes in source contributions during events like rush hour or wildfire smoke episodes.

Installing EPA PMF 5.0

1. Verify System Requirements

- Operating System: **Windows 7, or higher (Windows 11 is feasible).**
- Ensure that the data can be written to the C drive.

2. Get the Software

Download USEPA Website (<https://www.epa.gov/air-research/positive-matrix-factorization-model-environmental-data-analyses>)

3. Run the Installer

- Locate **EPA PMF 5.0 Setup.exe**.
- Double-click it and follow the installation directions on the screen.

4. Complete Installation

- Finish the setup process.
- The installation program creates an EPA PMF subfolder in the Program Files folder for the software and an EPA PMF subfolder in the Documents folder for data files.

5. Start the Program

- Double-click the **EPA PMF 5.0 icon** on your desktop.

Procedure of source analysis by using PMF Model

Screening of observational data (Remove outliers, Missing data handling etc.)

Preparation of observational data matrix table and uncertainty dataset

Check of calculation settings

Run base model of PMF

Estimate uncertainties (Bootstrap, Fpeak, DISP, BS-DISP)

Repeat calculations until the optimal solution is found

Interpret PMF Results (Identify potential sources by using indicators)

Outlier Removal

1. Mass Closure Model Check

- Compare measured PM_{2.5} mass with reconstructed mass from major species; remove samples with large discrepancies.

$$M=1.375[\text{SO}_4^{2-}]+1.29[\text{NO}_3^-]+2.5[\text{Na}^+]+1.4[\text{OC}]+[\text{EC}]+[\text{SOIL}]$$

$$[\text{SOIL}]=9.19[\text{Al}]+1.40[\text{Ca}]+1.38[\text{Fe}]+1.67[\text{Ti}]$$

- If the difference between **reconstructed mass (M)** and **measured PM_{2.5}** exceeds $\pm 20\%$, consider as outlier and exclude.

(Note) The above equation may vary according to the countries.

2. Ion Balance Check

$$R1 (\%) = (C-A)/(C+A) \times 100,$$

C and A: Sum of cation and anion concentrations (meq/L),

EANET manuals show the criteria of ion balance.

3. Other items

- Identify extreme values using a time series plot (e.g., sudden increase in K⁺ concentration due to fireworks events).
- Identify abnormal correlation patterns using a concentration scatter plot.
- Residual analysis: If scale residuals exceed ± 3 , potential outliers exist.

EANET Criteria for Ion Balance (Technical Manual for Wet Deposition Monitoring in East Asia, p.77-79)

Ion balance among all ions analyzed

■ $R1 (\%) = (C-A)/(C+A) \times 100$

C: Sum of cation concentration (meq/L),

A: Sum of anion concentration (meq/L)

- R1 is an item to evaluate analytical data quality assuming that there is no other ion than analyzed ions in a sample
- R1 is relative standard deviation calculated from two data, C and A.

R_1 : Ion balance

- Table Required criteria for R_1

<u>(C + A), $\mu\text{eq/L}$</u>	<u>$R_1, \%$</u>
<50	± 30
50-100	± 15
>100	± 8

Missing Data Handling

1. Values below detection limits

- Concentrations: Replace with 1/2 of the detection limits
- Uncertainty: Use the detection limit as is

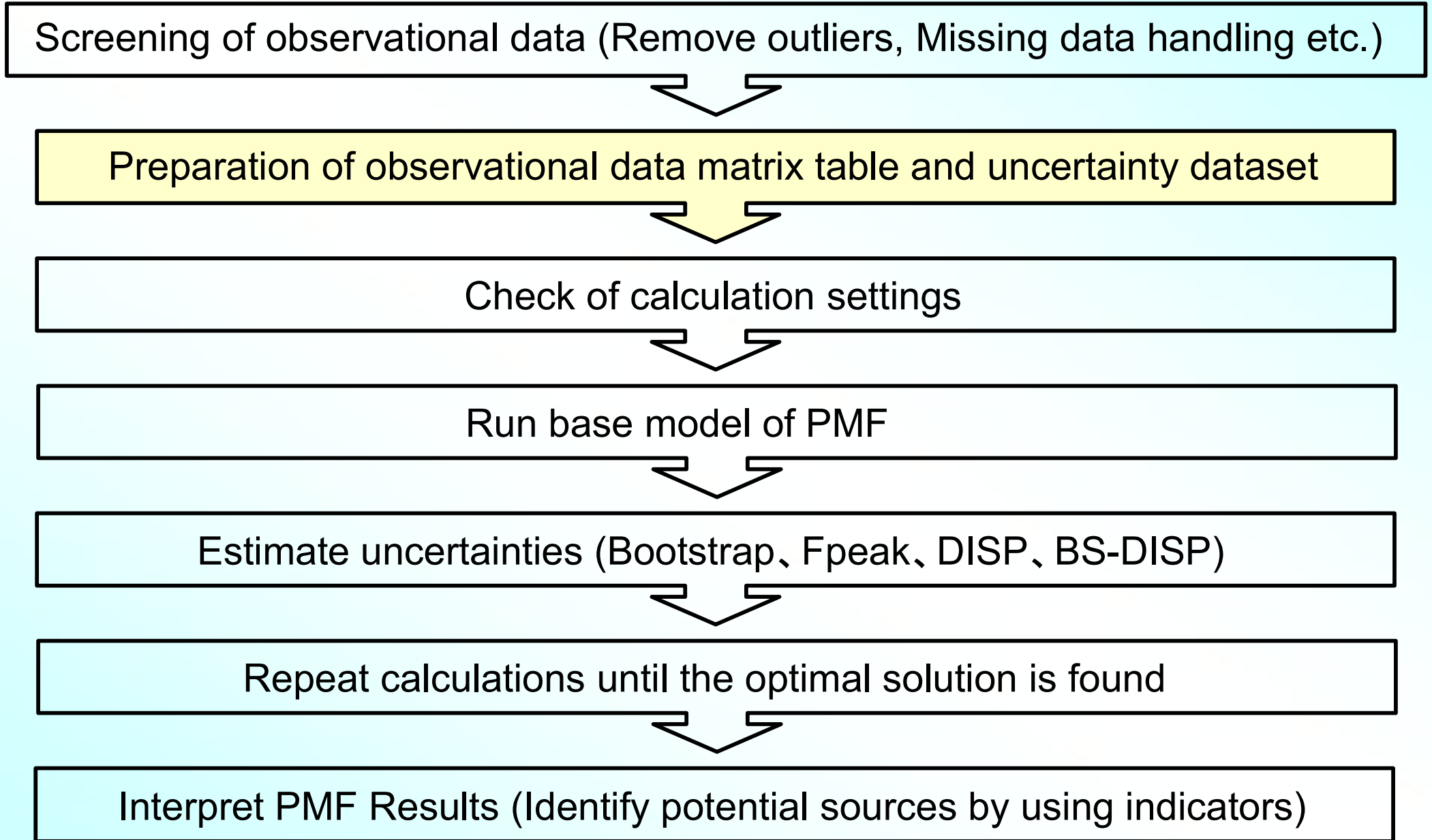
2. Completely missing for a specific species

- Concentrations: Enter a missing value indicator (e.g., -999), then replace with the species median in PMF
- Uncertainty: Use 4 times of the median

3. Species missing in all samples

- Automatically categorized as “Bad” and excluded from analysis

Procedure of source analysis by using PMF Model



Observational Data Matrix Preparation (1)

Structure

- Rows: Individual samples (e.g., daily or hourly measurements)
- Columns: Chemical species (e.g., ions, metals, carbon fractions)
- Header: Species names; optional second header for units
- First column: Date/time or sample ID for traceability

[CSV, TXT, Excel format]

	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	0.0321	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	0.1993	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	0.0067	1.8440	0.0010	0.0386	0.0013	0.0005	1.4180	0.0867	0.0046	0.0007	0.0663	1.3380	4.2680

Observational Data Matrix Preparation (2)

1. Select species to include in the matrix

- Major ions: Na⁺, NH₄⁺, K⁺, Mg²⁺, Ca²⁺, Cl⁻, SO₄²⁻, NO₃⁻
- Metals: Al, V, Cr, Mn, Fe, Ni, Cu, Zn, As, Sb, Pb, etc.
- Carbon: OC, EC
- Organic components (Levoglucosan, Malonic acid, etc.)
- If there are region-specific sources, additional components may be added.

2. Handle missing values

- If a sample lacks all species → remove in the matrix table
- If a species is missing in a sample → enter a missing value indicator (e.g., -999).
- If below detection limit → replace with 1/2 of detection limit.

3. Check consistency

- Ensure consistency for the component names and the column order between the concentration file and uncertainty file.
- Remove blank cells (PMF does not allow empty cells).

Uncertainty Dataset Preparation (1)

Structure

- Same dimensions as the concentration matrix
- First row: Detection limit for each species (Uncertainty)
- Second row: Species-specific Method Detection Limit (MDL)
- If there are multiple detection limits, the highest one is adopted.
- No units in header; species names must match concentration file
- No negative or zero uncertainties
- No blank cells

[CSV, TXT, Excel format]

	A	B	C	D	E	F	G	H
1	unc	Aluminum	Ammoniu	Arsenic	Barium	Bromine	Calcium	Chlorine
2	2	0.00419	0.0125	0.00098	0.0068	0.0016	0.0038	0.002635
3	10	10	10	10	10	10	10	10
4								

Uncertainty Dataset Preparation (2)

1. Assign uncertainty for each species and sample

- If the concentration is less than or equal to the MDL:

$$\text{Uncertainty} = 5/6 \times (\text{MDL})$$

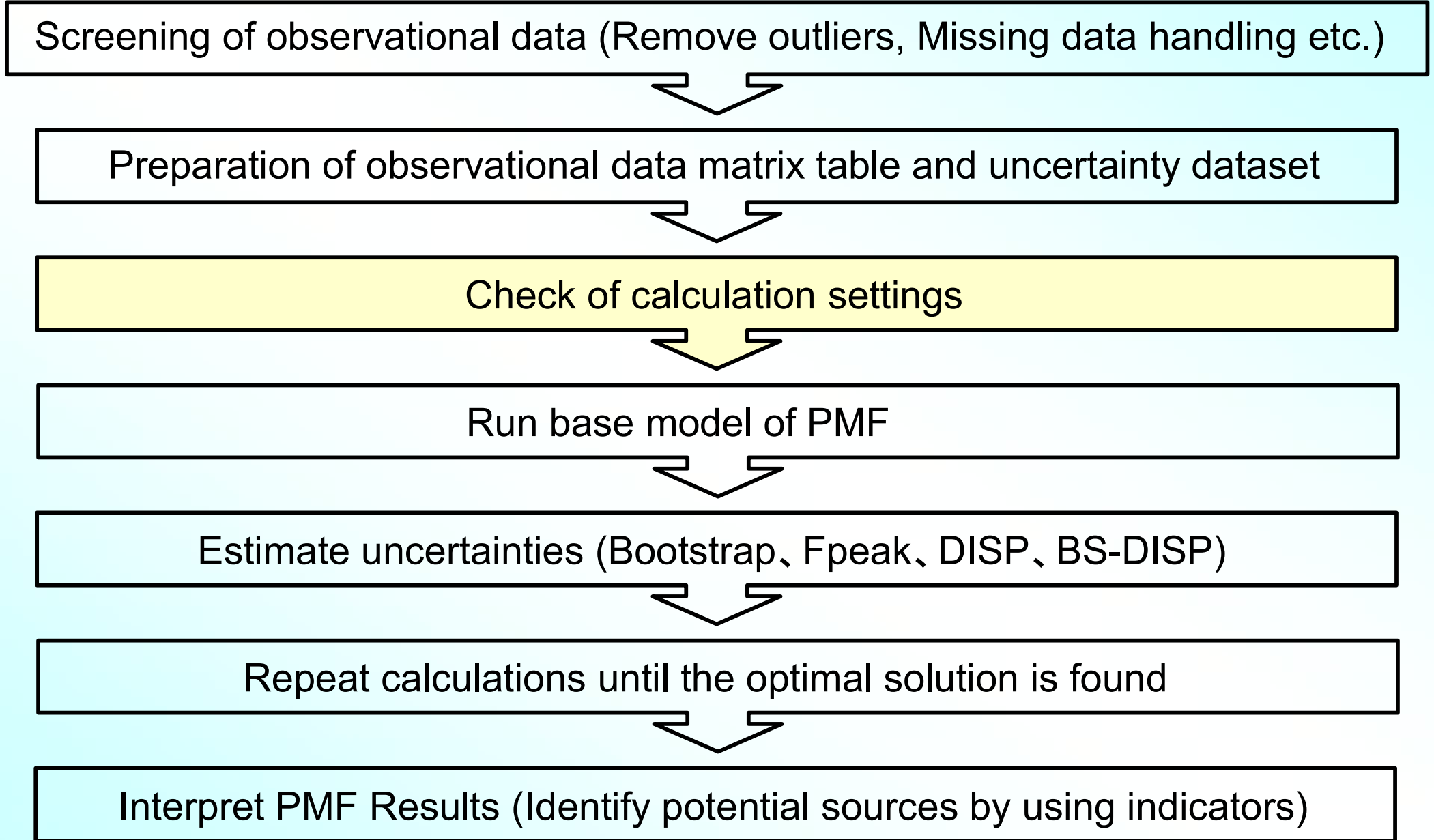
- If the concentration is above the MDL:

$$\text{Uncertainty} = \sqrt{(\text{Error Fraction} \times \text{concentration})^2 + (0.5 \times \text{MDL})^2}$$

2. Missing values

- For -999 entries in concentration file, set uncertainty to $4 \times$ species median.

Procedure of source analysis by using PMF Model



Check of calculation settings

1. Input Data Verification

- Ensure concentration and uncertainty files match in species names and number of rows/columns
- No blank cells or negative/zero uncertainty values
- Date/time column correctly specified
- Missing values handled properly

2. Species Weighting

- Strong / Weak / Bad categorization based on Signal-to-Noise ratio (S/N)
Bad: $S/N < 0.5 \rightarrow$ exclude
Weak: $0.5 \leq S/N < 1 \rightarrow$ uncertainty $\times 3$
Strong: $S/N \geq 1 \rightarrow$ keep as is