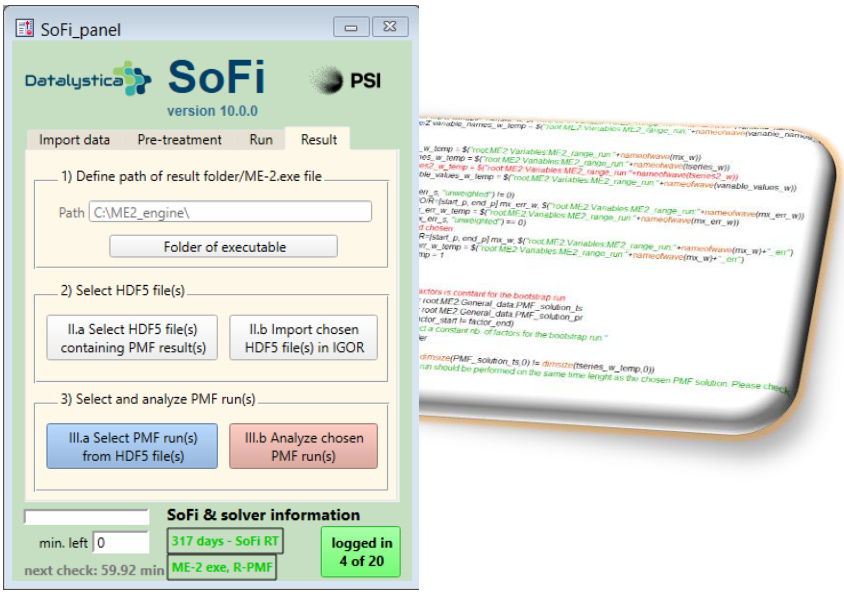
Source Finder (SoFi)

Manual for the software package SoFi (Pro) and SoFi real-time (RT) in IGOR Wavemetrics Inc.



written by

Datalystica Ltd.

at Park InnovAARE

CH-5232 Villigen, Switzerland

last formatted: 5/6/2026

manual for the software version: **SoFi 10.0.1**

Table of contents

1. General considerations 9

1.1. What is SoFi 9

1.2. For whom is it meant? 9

1.3. What is required? 9

1.4. More details about the structure of SoFi 10

1.5. Differences between the SoFi versions? 10

2. PMF theory 12

2.1. The bilinear model – general considerations 12

2.1.1. General bilinear equation 12

2.1.2. Robust mode 13

2.2. Normalizing Q by the expected value of Q (Q_{exp}) 13

2.3. Rotational ambiguity of the model solution 14

2.4. Rotational techniques in ME-2 14

2.4.1. Seed runs 14

2.4.2. Global fpeak 14

2.4.3. Individual fpeak 15

2.4.4. *α*-value approach: Partially constrained matrix F and/or G 15

2.4.5. Pulling approach: Partially constrained matrix F and/or G 16

3. Overview 17

3.1. Installation of the HDF5 interface (Igor 8) 17

3.2. General structure 17

3.3. First steps 18

3.4. SoFi menu 18

3.4.1. Initialize SoFi 18

3.4.2. Delete unused data (e.g. closed plots) 19

3.4.3. Recreate panel, i.e. clise graphs, tables, panels only 19

3.4.4. Recreate settings except folders “External_data” and “Reference_ts/pr” 19

3.4.5. Recreate SoFi, i.e. entire SoFi folder 19
3.4.6. SoFi panels to top 19
3.4.7. Open cluster panel 19
3.4.8. Main settings 19
3.4.9. User-defined families for HR-data 23
3.4.10. Manual 25
3.4.11. History 25
3.4.12. Terms and conditions 25
3.4.13. Credits 25
3.4.14. License information 26
3.4.15. Close SoFi package/key procedure 26
3.4.16. Perform communication tests 26
3.5. SoFi license activation step 26
3.6. Data hierarchy in IGOR 27
3.7. Newer SoFi versions 28
4. Import and define the input data 30
4.1. Import data 30
4.1.1. PMF Import 31
4.1.2. Time series import 33
4.1.3. Profile import 36
4.1.4. Combine matrices 36
4.2. Define PMF input 37
4.2.1. The data and error matrix 37
4.2.2. The time wave 37
4.2.3. The variable vectors 38
4.2.4. Data type 38
5. SoFi – pre-treatment 41
5.1. Treat missing data in PMF input 41
5.2. Treat external data 42

5.2.1. Profiles 42
5.2.2. Time series 43
5.3. Explore PMF input 45
5.3.1. Options for the variable(s) 45
5.3.2. Select based on class / time 50
5.3.3. Options for the graph representation 51
5.4. Select or modify effective PMF input 52
5.4.1. Class(es) 52
5.4.2. Input data 56
5.4.3. Blacklist single points based on MDL 58
6. SoFi – PMF run(s) 60
6.1. Define path, nb. of executables and name of HDF5 file 60
6.1.1. Path 60
6.1.1. nb. of executables 60
6.1.2. HDF5 file name 61
6.2. Define general model options 61
6.2.1. Define number of factors 61
6.2.2. Missing data 62
6.2.3. Down-weighting 63
6.2.4. Error matrix (general) 65
6.2.5. Convergence criteria 67
6.2.6. Pre-PMF evaluation 68
6.2.7. Other various options 69
6.3. Define options for the model matrices 72
6.3.1. Seed 72
6.3.2. fpeak 75
6.3.3. *α*-value constraints 77
6.3.4. Displacement 81
6.3.5. Pulling equations 82

6.3.1. Factor specifications 86
6.3.2. Consideration of the priority of the rotational tools 88
6.4. Estimation of the statistical uncertainty 89
6.4.1. Bootstrap resampling strategy 89
6.4.2. Input perturbation 91
6.5. Rolling mechanism 91
6.6. Call ME-2 and perform PMF 92
7. SoFi – PMF result(s) 93
7.1. Import result files (HDF file(s)) 93
7.1.1. Define path of ME2.exe file 93
7.1.2. Select HDF5 file(s) containing the PMF run(s) 94
7.1.3. Import the chosen HDF5 file(s) into IGOR 94
7.2. Manual selection of PMF run(s) form a HDF file 95
7.2.1. Settings 95
7.2.2. Choose PMF runs(s) 95
7.3. Criteria-based selection 96
7.3.1. Type of criterion 98
7.3.2. Criteria list 99
7.3.3. Active (sorting) versus passive (monitoring) criteria 100
7.3.4. Criterion syntax 101
7.3.5. Examples on how to set up the list of criteria 104
7.3.6. Evaluate the criteria 112
7.3.7. Inspect criteria score 113
7.3.8. Inspect selected PMF runs 117
7.3.9. Inspect criteria stability 118
7.3.10. Different number of factors in one HDF file 119
7.4. File settings and logfile 122
7.5. Options on how to import the chosen PMF run(s) 122
7.5.1. Averaging 123

7.5.2. Averaging (resolution) and data options 123
7.5.3. Factor specification 124
7.6. Technical information on the result folder 125
7.6.1. Structure of the result folders 125
7.6.2. Parameter files 126
7.6.3. Main information 127
8. SoFi – PMF run(s) analysis 129
8.1. Select a PMF run 129
8.2. General information 130
8.2.1. Select PMF run for exploration 130
8.2.2. Info for all PMF run(s) in adjacent list 130
8.2.3. Settings for chosen PMF run from adjacent list 130
8.2.4. Export PMF solution 131
8.3. Subpanels graph and data 131
8.4. Overview plots 132
8.4.1. Time series specific options 133
8.4.2. Profile specific options 134
8.5. Detail plots 135
8.5.1. Fraction 135
8.5.2. Standard 136
8.5.3. Scatter 138
8.5.4. Residuals 142
8.5.5. Image 145
8.5.6. HR plot 151
8.5.7. Category plot 155
8.6. Averaged solution 156
8.6.1. Repeats 156
8.6.1. Mass error estimation 157
8.6.2. Employed *α*-values 158

8.6.1. t-profiles 159

9. Independent Analysis 162

9.1. Highlight interesting periods 162

9.2. Independent bootstrap analysis 163

9.3. Marquee tool 166

9.4. Cluster panel 171

9.4.1. Cluster options 171

9.4.2. Standard plot 174

9.4.3. Scatter plot 174

9.4.4. Silhouette plot 175

9.4.5. Criteria option 176

10. Real-time module in SoFi (SoFi RT) 178

10.1. General considerations 178

10.1.1. What is SoFi RT 178

10.1.2. Working principle 179

10.2. Configuration file 180

10.3. First steps for manual process 183

10.3.1. Launching SoFi and preparing the input data 183

10.3.2. Define main model settings 185

10.3.3. Call SoFi RT 190

10.4. SoFi RT module 190

10.4.1. Define INPUT opt. 190

10.4.2. Define MODEL opt. 191

10.4.3. Define RESULT opt. 195

10.4.4. Report 197

10.5. Launch SoFi RT 199

10.5.1. Start/stop SoFi RT 199

10.5.2. SoFi RT overview results 200

10.6. Inspect SoFi RT results with SoFi Pro 202

10.7. Recommendations for ACSM data 203
10.8. Generate daily BC files from AE33 203
11. List of definitions and abbreviations 205
12. Literature 206

1. General considerations

1.1. What is SoFi

The **Source Finder (SoFi)** is a powerful, user-friendly software package for source apportionment and environmental data analysis. Developed firstly at the Paul Scherrer Institute (PSI) and now being actively maintained and further developed by Datalystica Ltd. in Villigen, Switzerland.

SoFi is a software package that runs within the software environment IGOR Pro (Wavemetrics Inc., USA). SoFi is widely used in atmospheric science and environmental chemistry for initiating and controlling the multilinear engine algorithm (ME-2, (Paatero, 1999)) that solves the positive matrix factorization algorithm (PMF,(Paatero, 1997)), as well as analyzing the PMF results in all its details for extracting mathematically and environmentally reasonable PMF results.

1.2. For whom is it meant?

SoFi is intended to facilitate and accelerate the source apportionment analysis. While SoFi was originally developed for analysis of aerosol chemical speciation monitor (ACSM) data (Ng et al., 2011b), it's versatility extends far beyond. It can be applied to virtually any dataset where dimensionality reduction or factor analysis is of interest.

With its highly panel-oriented and user-friendly structure, SoFi guarantees a seamless source apportionment analysis from the data input to the data output with basically no interaction at all with the software environment IGOR making SoFi fully suitable even for data analysts with no programming or knowledge skills of IGOR. The algorithm can be applied to any kind of data, for which dimensionality reduction analysis is potentially interesting. SoFi is intellectual property of Datalystica Ltd. and can only be used under the terms and conditions of Datalystica, the official distributor of SoFi (consult the homepage of Datalystica (https://datalystica.com) for the terms and conditions). The same terms and conditions are visualized when SoFi is launched and the user must read and agree to them, before being able to use the SoFi software.

1.3. What is required?

Make sure you already possess an official key for ME-2 (must be purchased from the homepage of Datalystica, https://datalystica.com), full administrator rights of your PC (read and write), as SoFi will create, modify and execute script files on the operating system and an official license for the software IGOR 9 or 10 (can be purchased through us in combination with a SoFi Pro/RT purchase, otherwise must be purchased directly from Wavemetrics, https://www.wavemetrics.com).

1.4. More details about the structure of SoFi

SoFi is available as **SoFi Pro** (commercial key that enables all advanced functionalities of SoFi) and **SoFi RT** (real-time module to SoFi for fully automated and in real-time source apportionment solutions, that also includes all SoFi Pro functions). SoFi Pro and SoFi RT keys can be purchased from Datalystica, https://datalystica.com). Consult the homepage of Datalystica for more details. This manual covers all the functionalities of the SoFi software.

This manuscript does not provide a detailed description of the theory behind the PMF algorithm nor on the ME-2 solver and its script code but rather the operation of the abovementioned software. Nonetheless, a small general theoretical introduction on PMF is presented in the next chapter. Interested readers are referred to the original paper of the multilinear engine (Paatero, 1999), the papers introducing the positive matrix factorization algorithm (Paatero and Tapper, 1994; Paatero, 1997; Paatero, 2004), the paper that provides an overview over the different rotational tools of the ME-2 algorithm (Paatero and Hopke, 2009) and the recent methodological papers about ME-2 (Brown et al., 2015; Paatero et al., 2014). In addition, the reader is also referred to Paatero (1997) for a discussion on the robust mode and on (Paatero and Hopke, 2003) for a detailed discussion on the reweighting strategy.

The reader is also referred to Canonaco et al., (2013), (2015) and (2021) for the introduction paper of SoFi and its application on ACSM data, respectively as well as to Crippa et al., (2014), where the *a*-value approach was thoroughly tested for many European AMS case studies, averaged profiles for the primary organic aerosol (POA) were retrieved and a general strategy using the *a*-value approach for AMS data is presented and discussed. A similar study for European ACSM datasets using the rolling PMF technique was presented by (Chen et al., 2022).

1.5. Differences between the SoFi versions?

There are currently two commercial versions of SoFi that are operational in IGOR, i.e., SoFi 9 for IGOR 9 and SoFi 10 for IGOR 10. The descriptions of the features of SoFi contained in this manual refer to the latest version, i.e., SoFi 10. Many of the features presented here are also available in previous SoFi versions. Please note that we officially discontinued from SoFi 6 since mid-2021 and that since then SoFi 6 has not been updated anymore. We are planning to discontinue the support of SoFi 8 somewhen during 2026. We therefore strongly recommend upgrading to SoFi 9 or SoFi 10. If you are still using SoFi 6 or SoFi 8, consult the history file (https://datalystica.com/downloads/) and make sure that you are not suffering from serious unresolved bugs in SoFi 6!

Generally, it is not guaranteed that SoFi 9 for IGOR 9 or SoFi 10 for IGOR 10 or cross-combinations, e.g., SoFi 9 in IGOR 10 will run error-free all the time. There might be cases and conditions that have not been tested or encountered during our test-phases, and that might cause some

unexpected problems on SoFi or IGOR. We apologize for this eventuality and strongly rely on the bug report of all SoFi users to minimize such eventualities but also for improving SoFi for your needs! Thanks already in advance for your feedback! Any feedback, no matter what, is always highly appreciated!

Your Datalystica Team!

2. PMF theory

2.1. The bilinear model – general considerations

2.1.1. General bilinear equation

The measurement of a set of variables (e.g. UMR *m/z* intensities for the ACSM) as a function of time can be represented as a data matrix **X** where the columns *j* are the variables and each row *i* represents one single scan.

In so-called factor analytical models, where a dimension reduction occurs, the variables are grouped in a few factors that can be further investigated rather than dealing with all measured variables. A commonly used approach is to group variables into two matrices using the so-called bilinear model. Few examples of bilinear models are the principal component analysis PCA (Wold et al., 1987), chemical mass balance approach (Miller et al., 1972) or the positive matrix factorization PMF (Paatero and Tapper, 1994). A bilinear factor model in matrix notation is defined as:

$$\mathbf{X} = \mathbf{GF} + \mathbf{E} \tag{1}$$

where the measured matrix **X** is approximated by the product of matrices **G** and **F.** *p*, i.e., the number of columns of the modelled matrix **G** and the number of rows of the modelled matrix **F** is called the number of factors. Each column *j* of the matrix **G** represents the time series of each of the *p* factors, whereas each row *i* of **F** represents the corresponding factor profile or fingerprint (e.g. mass spectrum). Note that by definition of (1) factor profiles are static and are not allowed to vary over time. The matrix **E** represents the model residual. The differences between PCA and PMF are mainly due to the restrictions of these models. PCA imposes orthogonality on the factor profiles, i.e., the scalar of two different rows of **F** is zero and does not require non-negative entries. In contrast, PMF require non-negative entries throughout **G** and **F**. PMF2 and ME-2 are solvers capable of solving the PMF algorithm. Both solvers fit the entries in **G** and **F** using a least squares algorithm that minimizes iteratively the quantity Q main Q^m, i.e., the total sum of the squared model residuals e_{ij} weighted by the uncertainty σ_{ij} for all input points *ij* (2):

$$\mathbf{Q^m} = \sum_{\mathbf{i}=1}^{\mathbf{m}} \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{n}} \left(\frac{\mathbf{e_{ij}}}{\sigma_{\mathbf{ij}}} \right)^2 \tag{2}$$

Here, the uncertainties σ_{ij} typically contain the measurement uncertainty for the input point *ij*. Data points for which the ratio e_{ij} / σ_{ij} >> 1 constitute a large fraction of Q^m and consequently these points will have a high impact during the model iteration. Normally this condition ensures that data with high signal-to-noise ratio is considered more than data from measurements close

to the detection limit. However, for some data points a case when the ratio e_{ij} / σ_{ij} >> 1 may also occur due to dominant local events or electronic noise within the measurement equipment, neither of which should be considered by the model.

2.1.2. Robust mode

To prevent a solution to be driven by these events, the model is generally run in a ”robust” mode, in which pulling of the solution by outliers is reduced. At each step of the iteration process, outliers are defined based on the ratio of residuals to uncertainties:

$$\textbf{outlier} := \left|\frac{\mathbf{e_{ij}}}{\sigma_{\mathbf{ij}}}\right| > \alpha \tag{3}$$

α is the user-defined threshold value. A value of four is recommended as a defining criterion for outliers within the robust mode (Paatero, 1997). The residuals are reweighted dynamically to remove the dependence of the change of Q^m with respect to the change of the outlier residuals.

$$\frac{d\mathbf{Q^m}}{d\mathbf{E_{outliers}}} = const. \tag{4}$$

This prevents the model solution to be driven by few dominant points in the datasets. It is recommended to always run the model in the “robust” mode, when treating data prone to e.g. rare transient sources, spikes, electronic noise or low signal to noise ratio, typically the case for ambient data.

2.2. Normalizing Q by the expected value of Q (Q_{exp})

Normally, monitoring and comparing total Q among various PMF runs is not meaningful because the expected value depends on the size of the data matrix and on the number of chosen factors. One therefore normalizes Q^m by the remaining degrees of freedom that is both a function of the size of the data matrix and of the number of factors and is called Q_{exp}.

$$Q_{exp} = n \cdot m - p \cdot (n + m) \tag{5}$$

In past, AMS studies reported typical values for the ratio of Q/Q_{exp} between 1 and 5. Sometimes values below 1 were reported. Note that when running PMF in the robust mode with a robust mode value of 4 to highest value of Q/Q_{exp} is ~16. For an ideal PMF run the Q/Q_{exp} would decrease down to ~1 only if the employed uncertainty considers both the measured as well as the modeled uncertainty. The latter is very difficult to assess and it remains often unknown, especially for ambient data. In addition, downweighting procedures (see 6.2.3) may have an impact on the value of Q, depending on the amount and magnitude of weak / bad variables. Therefore, the absolute value cannot be strictly used as a metric for judging model results. Instead, one should

investigate the relative change of Q/Q_{exp} across different model runs as guidance in choosing reasonable model solutions.

2.3. Rotational ambiguity of the model solution

General bilinear models possess a huge number of solutions, as **G** and **F** can be transformed with **T** and $\mathbf{T}^{-1}$, respectively (see equation (6)). Bilinear models with restrictions as orthogonality or non-negativity decrease this number of rotations (also called degree of rotational ambiguity), but it remains high (Paatero et al., 2002). There are two different kinds of rotations that are allowed occur, the pure and the approximate rotations. For pure rotations, the object function Q^m stays invariant with respect to rotations in space:

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

where **T** is a nonsingular matrix of dimension p x p, $\mathbf{T}^{-1}$ is its inverse and $\bar{\mathbf{G}}$ and $\bar{\mathbf{F}}$ are the rotated matrices. The matrix multiplication of $\bar{\mathbf{G}}$ and $\bar{\mathbf{F}}$ leads to the same product as for multiplication of **G** and **F**, and therefore Q^m remains unchanged. If the transformation matrix **T** does not fulfill equation (6), the rotation is called an approximate rotation, in case of which Q^m changes.

2.4. Rotational techniques in ME-2

2.4.1. Seed runs

Initializing the PMF model matrices with random or pseudo random data is called a seed run, as the model matrices are sowed / populated with data. This is typically performed when looking for the global minimum in the Q-space. However, as seed runs have various starting points, they will also fall at different locations at the global minimum, given that the global minimum is very likely to occur and the data has a high degree of rotational ambiguity. Therefore, seed runs also allow for a partial exploration of the rotational ambiguity. However, this technique does not allow for a systematic exploration and it is recommended for this purpose to use one of the techniques described below.

2.4.2. Global fpeak

For the PMF2 and the ME-2 engine, there is a user-specific parameter called fpeak, denoted by ϕ for the global control of such rotations. For positive ϕ, elementary rotations or a series of elementary rotations are performed that increase columns of the matrix **G** and decrease rows of the matrix **F** while conserving mass. The opposite occurs for negative ϕ. However, the fpeak tool explores only rotations in one dimension of the multidimensional space and if the entries of **G** and **F** are positive and more than one factor is chosen then the rotational space is

multidimensional and the corresponding ambiguity can be very large (e.g. for three factors, the rotational space is nine-dimensional).

For PMF2 users: Note that the fpeak value in ME-2 is proportional to the fpeak of PMF2. They are not necessarily equal (Paatero and Hopke, 2009).

2.4.3. Individual fpeak

An advantage of the ME-2 solver compared to the PMF solver is improved rotational control, e.g. selected factors can be summed / subtracted together rather than transforming the entire matrix. Thus, the rotations can be studied in a more controlled environment.

The analyst has to explore the solution space to provide the range of environmentally reasonable PMF runs upon the use of rotational tools.

Rotational ambiguity can be very high, in particular for large datasets with many variables, long measurement times, high number of factors, low signal to noise ratios, etc. and hence, it is advantageous to introduce *a priori* information into the PMF model to reduce the amount of rotational ambiguity. This can happen in form of known rows of **F** (factor profiles) and/or of known columns of **G** (factor time series) (Paatero and Hopke, 2009). This *a priori* information prevents the model matrices to rotate, resulting in ~unique solutions. In ME-2 two different main approaches are implemented, the *a*-value and the pulling techniques which will be briefly discussed in the following subchapters.

2.4.4. *a*-value approach: Partially constrained matrix F and/or G

In *a*-value approach the elements of the **F** matrix (factor profiles) or of the **G** matrix (factor time series) can be constrained by the user. The user inputs one or more factor profiles (rows of **F**) / factor time series (columns of **G**) and a constraint defined by the scalar *a* that can be applied to the entire profile / time series or to elements of the profile / time series only. The *a*-value determines the extent to which the output $\mathbf{F}_{solution}$ / $\mathbf{G}_{solution}$ is allowed to vary from the input $\mathbf{F}_{initial}$ / $\mathbf{G}_{initial}$, according to:

$$f_{j,\ solution} = f_j \pm a \cdot f_j \tag{7}$$

$$g_{i,\ solution} = g_i \pm a \cdot g_i \tag{8}$$

where f_i and g_i represent the row and the column of the matrices **F** and **G**, respectively. The index *j* varies between 0 and the number of variables and *i* varies between 0 and the number of measured points.

The scalar *a* ranges between 0 and 1. Setting the *a*-value to zero for all the factor profiles, one can fully constrain the factor profile matrix **F,** similar to the method of a chemical mass balance

(CMB). On the contrary, setting the *a*-value to one for all entries of a profile or time series allows it to similar as in an unconstrained situation.

2.4.5. Pulling approach: Partially constrained matrix F and/or G

The user has the possibility to introduce pulling equations into the model that pull profile factor elements towards target values, so-called anchors (here shown for the row j of the matrix **F** only):

$$\overline{a_j} = \overline{f_j} + r_j \tag{9}$$

In equation (9), $\overline{a_j}$ represents the anchor to which the model has to pull the iterative value $\overline{f_{ij}}$ and r_{ij} represents the residual. The anchor is some known value introduced as *a priori* information by the user. The pulling equations create an additional auxiliary term Q^{aux} that is added apart from Q^m. Thus, if pulling equations are introduced, the model will minimize the argument of Q with respect to all entries in the matrices **G** and **F**:

$$\arg\min_{G,F}(Q) = \arg\min_{G,F}(Q^m + Q^{aux}) \tag{10}$$

The term of Q^{aux} has a conceptually similar aspect to Q^m:

$$Q^{aux} = \sum_{ki=1}^{K} \left(\frac{r_k}{s_k} \right)^2 \tag{11}$$

The index j from equation (9) has been replaced by k, since k denotes the index of the pulling equations added to the model (over many factor profiles / time series). The pulling parameter s_k specifies the "softness" of the pull. The smaller s_k becomes, the higher the impact of Q^{aux} of the k-th pull during the iterative process. The pulling approach is a sensitive technique in that if the pulling equation is not compatible with the specific data matrix, i.e., a decrease of Q^{aux} obtained as $\overline{f_{ij}}$ reaches its anchor value $\overline{a_j}$ (9) will be negligible compared to a larger increase of Q^m, then the pull is neglected. Generally, the user provides the acceptable limits of Q^m denoted as *dQ*. Changing *dQ* and the pulling parameter s_k allows monitoring a change in Q^m and judging about its acceptability.

<u>The anchor can be any entry of a factor profile / time series, a mathematical combination of one or more entries of a factor profile / time series and also involving external data.</u>

3. Overview

3.1. Installation of the HDF5 interface (Igor 8)

The storage of the model results is based on the object oriented HDF5 file system. Since the IGOR software does not load the procedures for the HDF5 file system by default before IGOR 9, the user needs to link the required procedures manually in the operating system (OS) for IGOR 8 (not needed for IGOR 9 or IGOR 10).

Create a shortcut of the following files

I. C:\Program Files\WaveMetrics\Igor Pro Folder\More Extensions\File Loaders\HDF5.xop

II. C:\Program Files\WaveMetrics\Igor Pro Folder\More Extensions\File Loaders\HDF5 Loader.ipf

III. C:\Program Files\WaveMetrics\Igor Pro Folder\WaveMetrics Procedures\File Input Output\HDF5 Browser.ipf

Place the shortcuts in

I. C:\Program Files\WaveMetrics\Igor Pro Folder\Igor Extensions\HDF5.xop

II. C:\Program Files\WaveMetrics\Igor Pro Folder\Igor Extensions\HDF5 Loader.ipf

III. C:\Program Files\WaveMetrics\Igor Pro Folder\Igor Procedures\HDF5 Browser.ipf

3.2. General structure

SoFi handles all ME-2 implementations discussed in the previous chapter. The operational framework is presented in Figure 1.

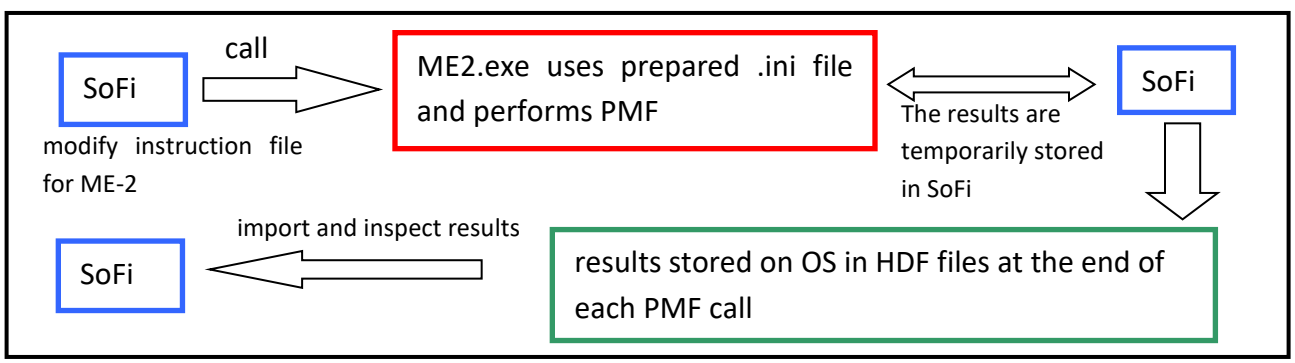


Figure 1 The general process from the model call to the storage of the model output. Blue rectangles describe SoFi tasks, red rectangle represents the ME-2 execution. After all model runs finished, the results are stored in an HDF5 file (green color).

3.3. First steps

The first step when initializing the SoFi software is to open a new IGOR experiment. Thereafter, the user needs to import the SoFi procedure file (double-click on the SoFi procedure files or drag and drop the files into IGOR) with the name:

SoFi-10.ipf (main file) and **SoFi_key.ipf** (SoFi Pro/RT license key, do not change the name)

Once the SoFi procedure is compiled, a new tab appears in the main menu in IGOR.

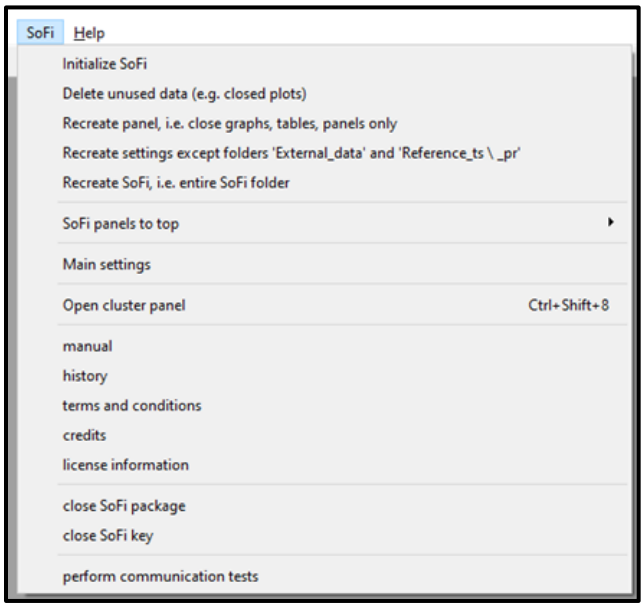


Figure 2 The tab “SoFi” appears once the procedure is compiled.

3.4. SoFi menu

The tab “SoFi” contains several main options:

3.4.1. Initialize SoFi

“initialize SoFi” launches the SoFi code and sets all its variables and strings

Before launching SoFi the first time in a new Igor experiment, the user must read the user’s agreement and agree to the terms and conditions. Only then SoFi can be used. The terms and conditions can be downloaded as a pdf or be revisited anytime under “terms and conditions”.

3.4.2. Delete unused data (e.g. closed plots)

SoFi deletes all waves from previous plots that are not in use anymore.

3.4.3. Recreate panel, i.e. clise graphs, tables, panels only

“recreate panel, i.e….” closes all graphs and restarts the main panel. This button leaves the PMF input, external data and the imported PMF results data untouched.

3.4.4. Recreate settings except folders “External_data” and “Reference_ts/pr”

“Recreate settings except folders “External_data” and “Reference_ts/pr”” additionally, resets all variables and strings related to SoFi. This button leaves the imported PMF input data, external data and PMF results untouched.

3.4.5. Recreate SoFi, i.e. entire SoFi folder

This option is similar to “Initialize SoFi”. Everything will be recreated, all data will be deleted.

3.4.6. SoFi panels to top

A helpful option when many graphs, tables and panels are open. Open panel can easily be brought back to the front via this menu or via the indicated short cuts. The panel must be open in the back or being minimized, it cannot be built through this menu.

3.4.7. Open cluster panel

Opens the cluster panel. The cluster panel cannot be opened from the main SoFi panel or any other panels but only here. Note, Ctrl+Shift+8 opens and closes the panel, Ctrl+8 brings the panel to the front if already open.

3.4.8. Main settings

Allows the user to modify main settings. Currently, the following settings are available and modifiable from here:

“GENERAL options”	enable/disable the progress bar, enable/disable pop-up about the windows position, colors of the PMF factor positions, default unit for the PMF results, RT settings
“DATA options”	signal to noise used in SoFi (S/N), downweighting and robust value for results, S/N threshold for the (un)explained variation plot, resolution for the daily cycle plot, averaging threshold for cycle plots
“FAMILY wave”	user-defined families for HR-data. Details can be found in 3.4.9.

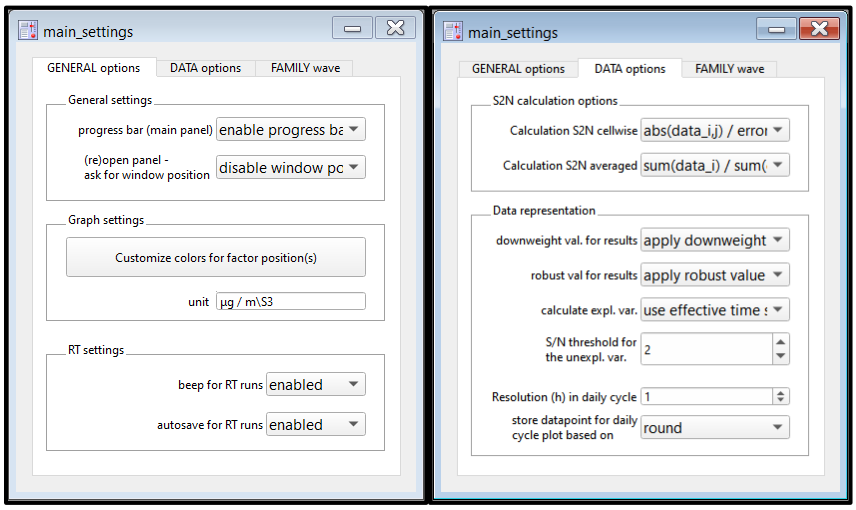


Figure 3 Main settings panel for SoFi

3.4.8.1. **General settings**

There are two main settings the user has to option to enable or disable. The progress bar in the main SoFi panel is enabled by default. It has been shown, that this can also use some computational power, therefore there is the option to disable it.

By default, SoFi reopens closed panels at their last position (as long as they are not connected to another panel). If experiments are opened on a different screen, this might get messed up and the panel is opened outside of what can be seen on the screen. In this case, the (re)open panel should be enabled so the panels will be opened at the position as they were opened in a new experiment.

3.4.8.2. **Customize colors for the factor position(s)**

The user has the possibility to define colors for the factor positions up to 100 positions. Pressing the “SHIFT” key selects adjacent factor positions. Pressing the “CTRL” key allows the user to add non-adjacent factor positions. Once the factors are chosen, pressing on “Start defining” pops the color palette, in which the colors for the chosen factor positions are defined. This process is additive, i.e., the user can define a second or third step other colors or existing factors or new colors for other factors.

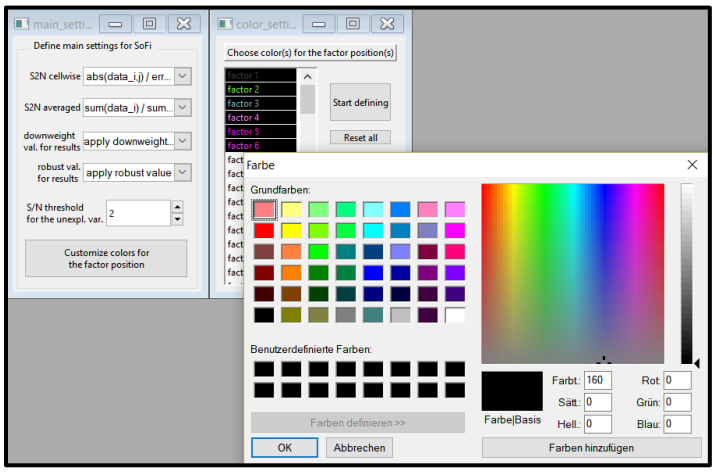


Figure 4 Options present for "change main settings" panel on the left. The right panel is used to select user-specific colors for the various factor positions.

3.4.8.3. **RT settings**

By default, SoFi beeps for a RT run and also autosaves the experiment. The user can disable these options.

3.4.8.4. **Signal to noise (S/N)**

The user defines which method of S/N calculation should be employed in SoFi for the cell-wise or averaged approach. This affects the variable S/N plot (see 5.3) or when estimating the (un)explained variation in the result plots (see 8.5). In case of activation of the cell-wise calculation, S/N ratios are estimated for each cell/entry of the PMF input separately, whereas the averaged treatment provides an averaged (over time) S/N ratio per variable.

<u>Cell-wise</u>

abs(data_*i,j*) / error_*i,j*

data_*i,j* / error_*i,j* See supplement in Brown et al. (2015)

<u>Averaged</u>

sum(data_*i*) / sum(error_*i*)

sum(abs(data_*i*))/ sum(error_*i*)

sqrt(sum(data_*i*^2))/sqrt(sum(error_*i*^2)

data_*i* / error_*i*

sum(data_*i*) / sum(error_*i*) See supplement in Brown et al. (2015)

"*i*" being the index running over time and "*j*" being the index for the variables.

3.4.8.5. **Downweight values for results**

The error matrix for a PMF result can either contain the downweighted values (default case) or be shown without the downweighted values (see 6.2.3). The former has the advantage that the user sees what the ME-2 solver tried to minimize during the iteration. The latter shows the residuals, without the downweighted values that is relevant (and gives also more realistic values for Q), in case many variables were classified as weak or bad and downweighted.

3.4.8.6. **Robust values for result**

The robust mode truncates the dependency of the iterative process from the scaled residuals (see 2.1.2), preventing e.g. outliers to drive a PMF solution, by rescaling and hence by modifying the corresponding errors. The user has the possibility to see the result including this modification (default case) or to remove this effect from the error matrix. The advantage of applying the robust mode is that the minimization process as solved by ME-2 can be inspected. Removing this option reveals which variables or points in time are better or worse explained among all points that hit and remained until convergence at the robust limit.

<u>The last two points (3.4.8.5 and 3.4.8.6) can be performed while inspecting the PMF results. This allows the user to compare the result with and without the application of these threshold limits.</u>

3.4.8.7. **Calculate explained variation**

By default, the explained variation is calculated over the effective time series (excluding blacklisted time points). If the whole time series should be used for the calculation of the explained variation (e.g. for rolling PMF), this can be changed here.

3.4.8.8. **S/N threshold for the unexplained variation**

Paatero (2010) introduced a metric based on the estimation of the measurement variation explained by the factors. The explained variation (EV) is a dimensionless quantity that indicates how much variation in time or variation in each variable is explained by each factor. As an example, the equation for the explained variation EV for the ith point in time for the kth factor is given by:

$$EV_{ik} = \frac{\sum_{j=1k}^{m}(|g_{ik} \cdot f_{kj}|/\sigma_{ij})}{\sum_{j=1k}^{m}((\sum_{h=1}^{p}|g_{ih} \cdot f_{hj}|+e_{ij})/\sigma_{ij})} \tag{12}$$

Similar equations can be formulated for the unexplained variation (UEV) by replacing the product $g_{ik} \cdot f_{kj}$ in the numerator with e_{ij}. Expressing the explained and the unexplained variation for a variable *j* as EV_{jk} is done by simply replacing in the expression the sum over *j* with the sum over *i*. If all variation is explained by the model, then EV = 1. According to Paatero (2010) a variable should be regarded as explained only if the UEV for that variable is less than 25 %.

The UEV is further separated into the real UEV for data possessing a high S/N value and UEV for noisy data. This separation is performed using the S/N threshold defined in this section (default value is 2).

3.4.8.9. **Resolution in daily cycle**

The resolution for the daily cycle plots supported in SoFi can be changed here. The default value is one that leads to the classical daily cycle resolution between 0 and 24. This option is interesting and necessary, in case the PMF input data or external data possess a lower resolution than one hour.

3.4.9. User-defined families for HR-data

For HR-data, the user can define custom families based on a list of criteria. This is not limited to a specific instrument type but it requires as input the chemical formula information. The user-defined families can be created after the initialization of the data (5.1). By clicking on “Add new family” an interactive dialogue is being popped, where the user is asked to enter a family name (what should be shown in the legend later, starting with a letter, no space or special characters), a criterion (see below for details) and a color. SoFi automatically creates a family for ions where the name consists of numbers (“not_identified”) and ions that are not assigned to any family (“other”). The column “number of variables” will be empty until the speciesfamilytext wave with the family assignment is created.

Families can be updated (name, criterion and color) through the “Update family” button. The user is asked for which family that should be updated and is asked to provide the position (row, starting from 0), afterwards an interactive dialogue is popped. The user can also delete single entries in the table or the whole table. Furthermore, the list of families and their criteria and color can be saved and loaded (via .itx files).

Clicking on the “CREATE ‘speciesfamilytext’” button will then create the speciesfamilytext wave with the family assignment for each ion. The ions assigned to a specific family can be inspected by selecting the family in the dropdown menu and then clicking on the “inspect ‘…’” button.

3.4.9.1. **Define a family based on a chemical formula**

One way to define families, is to categorize chemical formulas according to their elemental composition (similar to the default for HR-AMS data). The criterion is essential to determine the family assignment. It follows the format of an element (case sensitive) followed by round brackets. If the round brackets are left empty, it means that any number of that specific element is accepted. If a number is specified within the round brackets, it means that only the exact specified number of that element is accepted. SoFi also accepts > and < before the number in the brackets if the number of that specific element should be larger or smaller than the following number, respectively. Note: only chemical formulas with only the specified elements will be

assigned to the family. However, it is possible to define multiple criteria for one family. If a certain element should be ignored throughout the whole family assignment process (e.g. Na for EESI data), it can be added to the list of ignored element(s) through the “ignore element” button. To remove ignored elements, click on the “check ignored element(s)” button which will pop a table, delete the element therein.

If the user defined the data as AMS data in the “define PMF input” selection, the user will be asked whether SoFi should check for some ions in the list that are assigned by the Aerodyne default to a family, even if the don’t fully fullfill the family criterion (e.g. CN being assigned to CHN, CO being assigned to CHO1 or CO_2 being assigned to CHOgt1). If the user wishes to do so, a window is popped where the critical ions are listed and the family can be assigned via a dropdown menu. SoFi will try to suggest families but it will fail if the default Aerodyne names are different from the user-defined families. In this case the user has to select the family him/herself.

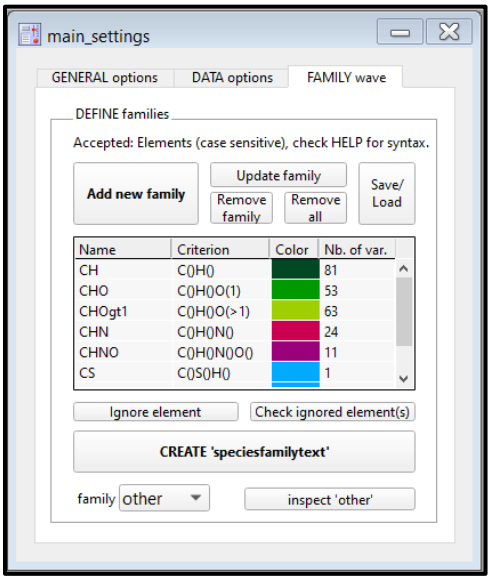


Figure 5 Sample criteria list for defining families based on the chemical formula. For the criterion, an element should always be followed by round brackets. If the brackets are left empty, any number of this element is accepted (e.g. C()H() will include all ions with the formula C_xH_y). If a specific number of on element is expected, that number should be written inside the brackets (e.g., C()H()O(1) for $C_xH_yO_1$ or C()H()O(>1) for $C_xH_yO_{>1}$).

3.4.9.2. **Define a family based on a number of a specific element**

Another way to define families, is to categorize based on the number of a specific element in the chemical formula. The syntax for this kind of defining the families, the syntax is hashtag (#) followed by the specific element and in brackets the number of this element. For example, assigning families based on the numbers of nitrogen in the chemical formula could look like Figure 6. Note that the family “other” is automatically created by SoFi.

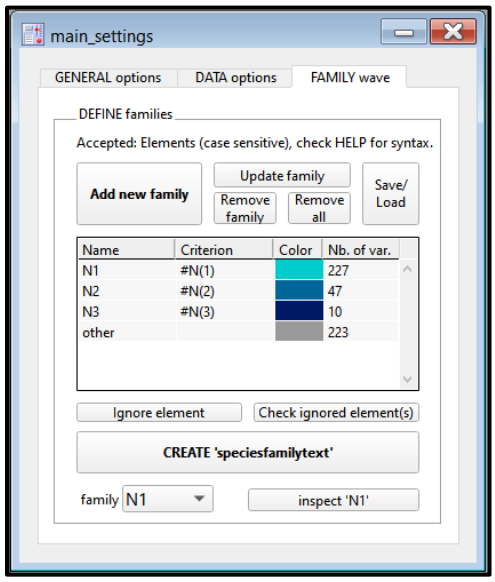


Figure 6 Sample criteria list for defining families based on a number of a specific element. In this case the families should be assigned by the number of nitrogen (N) in the chemical formula. For the criteria use # followed by the element and a number in brackets (e.g., #N(1) for all ions that contain exactly one nitrogen).

3.4.10. Manual

The option “Manual” downloads the latest SoFi manual as a pdf from the Datalystica server. An internet connection is required for this.

3.4.11. History

SoFi downloads the newest history file from the Datalystica server (internet connection required). The history file contains all bug fixes and improvements since SoFi 6.5 (2018). It is strongly recommended to always use the newest SoFi version. If older versions are being used, it is the user’s responsibility to make sure results are not compromised by bugs that are solved in newer versions.

3.4.12. Terms and conditions

This option downloads the terms and conditions as a pdf from the Datalystica server. It is the same that the user must read and accept before launching SoFi.

3.4.13. Credits

“Credits” option provides an e-mail address and a phone number for further support.

3.4.14. License information

Get information about what type of license and for how long it is active. Furthermore, the licensee as well as the license serial number are displayed.

3.4.15. Close SoFi package/key procedure

The SoFi procedure files are hidden and cannot be closed (killed) manually. If there are e.g., problems with SoFi and you wish to reload the procedure files, the SoFi procedures can be closed with this option.

3.4.16. Perform communication tests

SoFi checks the communication between Igor and the operating system as well as between Igor and the Datalystica server. If there are problems with the activation of the SoFi license, these tests should be performed and the results sent to our support (support@datalystica.com). The results are printed in the Igor history (Ctrl+J).

3.5. SoFi license activation step

Pressing on any tab or button on the main SoFi panel will automatically connect to the Datalystica server and validate your SoFi Pro license. Make sure you are connected to the Internet. Otherwise, the validation process fails, and you will be asked to use SoFi without the license, i.e. very limited options, as long as the license cannot be verified.

The license button turns green for successful license authentication and shows how many licenses of your SoFi Pro package are currently in use. On the left side, the number of days left for the SoFi Pro license is shown in green. Below in gray the panel reports the remaining time prior to the next server connection (Figure 7).

To log off from the server press on the license button (green button). This option is useful in case only a few licenses are purchased, and SoFi Pro should be sequentially used on several PCs. **Thereafter, IGOR must be recompiled manually. The simplest approach is to save, close and reopen the IGOR experiment**.

SoFi requires the Internet for the license validation and will connect to the server every 24 hours, once being logged in. During these 24 hours SoFi Pro works perfectly offline. Hence, if SoFi Pro must be used offline, make sure you validate your license first while being connected to the Internet.

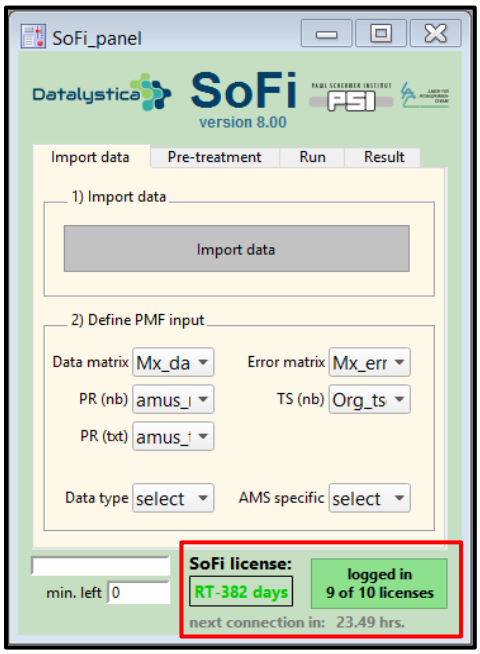


Figure 7 SoFi license panel with one SoFi Pro license activated (green panel). The example shows 61 days left for the SoFi Pro license (green text on the left of the button) and SoFi will attempt the next connection to the server for the license validation in one hour (gray text below).

3.6. Data hierarchy in IGOR

Once SoFi is launched the folder SoFi with all its subfolders are created under the root folder (see Figure 8). The data folders contain:

"External_data"	Raw data loaded by the user.
"Raw_input"	The raw PMF input.
"Raw_pr"	External reference profiles together with the associated variable bases (must be in txt format, see 5.2.1 for more details)
"Raw_ts"	The time series of external data together with the associated time stamp (see 5.2.2 for more details).
"Reference_pr"	The reference profiles that contain the same variables present in the PMF input matrix.
"Reference_ts"	The external information that matches the time of the PMF input (default resolution), **SoFi Pro:** in addition, hourly, daily, weekly, monthly and yearly resolution are also evaluated.

<u>Note: The user should not be concerned about the length (time series) / amount of variables (profiles) of the external data. There is a matching algorithm (see 5.2) that makes sure that the external information is harmonized to that of the PMF input.</u>

“Eff_input”	The PMF input used for the current PMF run.
“Results”	Once the model results are imported by the user, they will be stored in the folder “root:SoFi:Results:XXX”, with “XXX” being the name of the user-defined HDF5 file.
“Preanalysis”	Data related to the pre-treatment part (tab 2 in the main SoFi panel).
“Variables”	All variables are necessary for the software. The user is not supposed to modify anything here!!!
“Indep_Analysis”	All variables and waves required for independent analysis panels (bootstrap, marquee and clustering panel).

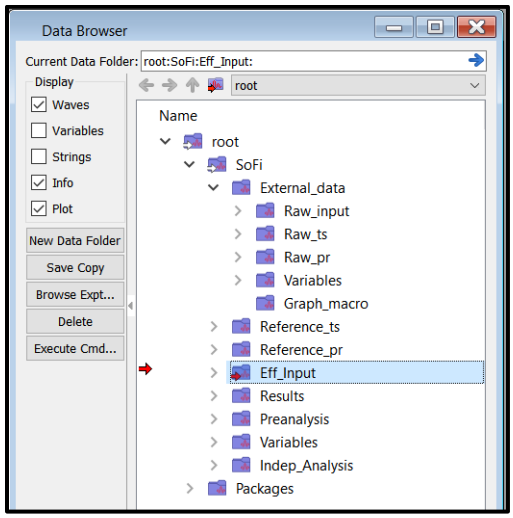


Figure 8 Data structure for SoFi.

3.7. Newer SoFi versions

SoFi checks for newer SoFi versions on our homepage (https://datalystica.com/downloads/) every time the SoFi software package has been successfully compiled. This is typically the case when, for example the Igor experiment is opened for the first time. Whenever a newer SoFi version is found (for SoFi 9 and SoFi 10 independently) on the homepage, the user gets a unique warning message (see Figure 9).

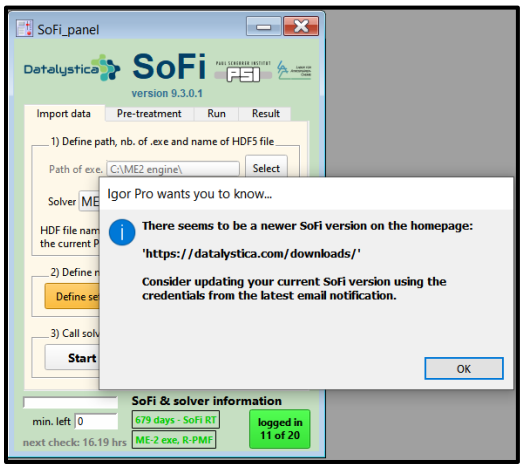


Figure 9 Warning for a newer SoFi version available on our homepage.

There are several different ways to update your experiment to a newer SoFi version. If all SoFi experiments are closed, it is easiest to simply replace the SoFi-X.ipf file in the original location. When the experiment is opened again, it will automatically read the newest version. While working with SoFi, the current SoFi main package can be closed via the main menu and the new SoFi-X.ipf can simply be dragged, dropped and compiled in the open experiment.

4. Import and define the input data

SoFi is structured in a task-dependent panel system. The main panel is shown in Figure 10. This is the panel that controls all major tasks in SoFi, such as defining the PMF input, creating various PMF cases, launching the ME-2 solver, importing and visualizing the results.

Subpanels should be opened and closed by clicking on the respective "open" buttons. Closing panels and/or graphs using X (top right) could cause unexpected behavior of some features.

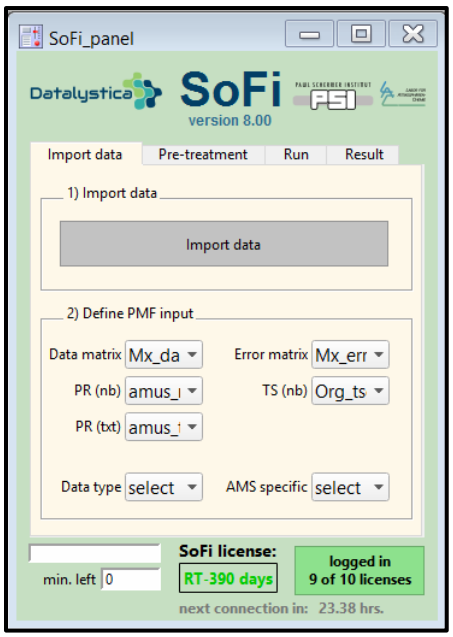


Figure 10 Main SoFi panel, import tab. The "Import data" button allows automatic import of the PMF input and external time series and profiles. Under "Define PMF input" the user has to define the PMF input.

4.1. Import data

The first step is to import the data into the proper folder of the current IGOR experiments. This can be done manually or automatically. SoFi currently supports automatic data import for the PMF input from IGOR text files (.itx) and excel files (.xls and .xlsx), as well as external time series and external profiles from IGOR text files (.itx), excel files (.xls and .xlsx), comma-separated values files (.csv), data files (.dat) and text files (.txt). Using the button "Import data" opens the import panel where the details for the automatic import can be specified. For the manual import, please consult the subchapters for the locations where the data has to be copied to.

Please contact support@datalystica.com if you experience problems with importing your data. We are always happy to expand the import function.

4.1.1. PMF Import

The PMF import needs to be stored under **"root:SoFi:External_data:Raw_input".** The PMF input should contain at least the data and error matrix, as well as the time series and profile (variable) vector. The error matrix can also be calculated by SoFi if the user imports from an excel file (Figure 11, left) details below.

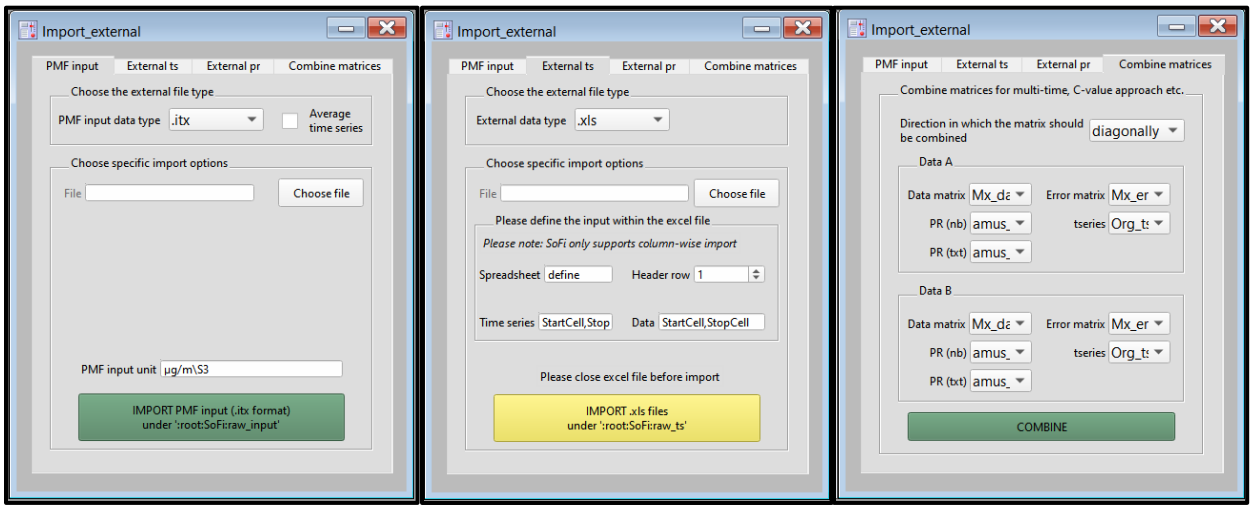


Figure 11 Import panel for the PMF input (left) from an IGOR txt file and for external time series from an Excel file (middle) and for combining matrices (right).

By the drop-down menu on the top, the user can select the data file type from which SoFi should import the PMF input. For .itx files, the required matrices should be stored as separate 1D or 2D waves, so that it can be imported automatically, and no further information is required. The "Import" button opens a file dialog to select the .itx file.

For PMF inputs from an excel file, the user first needs to select the file (via the "Choose file" button) and then specify the cells in which the data matrix, error matrix, time series and profiles can be found. In addition, the spreadsheets have to be specified. Data matrix and error matrix are allowed to be in separate spreadsheets. For excel files, SoFi imports the profile vectors based on the row names as a text wave and generates automatically a corresponding numeric wave, as needed to define the PMF input. If the field for the error matrix is left blank, SoFi assumes that there is no error message and gives the option to let the error matrix (later named "Error_Mx_SoFi") be calculated automatically. For this feature, SoFi supports two types of error matrix calculation:

"Polissar et al., 1998" In the approach described by Polissar et al. (1998), data matrix and error matrix are composed as fixed fractions of the method detection limit (MDL):

$$For\ conc. > DL: conc. = \ measurements, \sigma = \frac{1}{3} MDL \quad (14)$$

$$For\ conc. \leq DL: conc. = \frac{1}{2} MDL, \sigma = \frac{5}{6} MDL \quad (13)$$

"Norris et al., 2014" In this approach uses an equation-based uncertainty and an error fraction for concentrations above the MDL. For concentrations below the MDL, the approach from Polissar et al. (1998) is followed. This error matrix calculation is also used in EPA PMF.

$$For\ conc. > DL: \sigma = \sqrt{(error\ fraction \cdot conc.)^2 + (MDL)^2} \quad (16)$$

$$For\ conc. \leq DL: \ \sigma = \frac{5}{6} MDL \quad (15)$$

Both error matrix calculations require the input of the method detection limit (MDL) for each element/species/ion/etc. as illustrated in Figure 12. The equation-based uncertainty according to Norris et al. (2014), additionally requires the input of the error fraction. SoFi accepts an error fraction for each element or one error fraction for all elements. Both, input values and error fraction, need to be separated by semicolons.

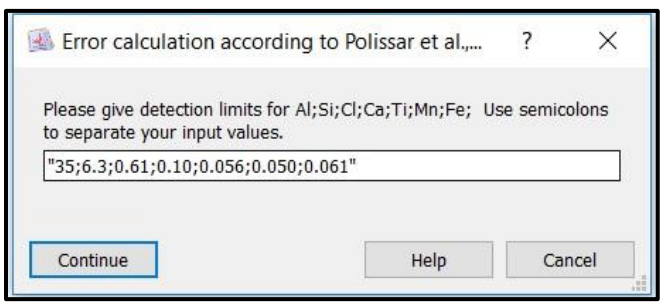


Figure 12 Interface to enter the detection limits for each element/species/ion/etc.

By checking the box "Average time series", SoFi averages the imported PMF input to the desired time resolution (in minutes), storing both, the original input and the averaged input under "root:SoFi:External_data:Raw_input". After the import of the original PMF input, a window pops up where the user has to define the desired time resolution (in minutes). For the import from an .itx file, the user has to additionally define the data matrix, error matrix and time series.

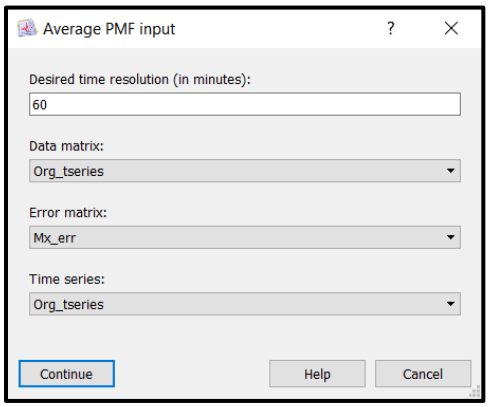


Figure 13 Interface to average the PMF input. The desired time resolution needs to be entered in minutes. The definition of data matrix, error matrix and time series is only required for import from .itx files.

If present, the minerr1ion wave will be automatically averaged too. However for PMF, SoFi will take and apply to the error matrix the wave with the name "minerr1ion" only. Hence, the user must remove a resulting suffix manually and make sure that the main minerr1ion wave with default resolution has a different name or is removed from the raw input folder.

4.1.2. Time series import

The external time series need to be stored under **"root:SoFi:External_data:Raw_ts"**, which is done automatically if the "Import external ts" option is chosen. The minimum data that is required, is a time series and a corresponding numeric value series. In a first step, the user has to choose the file type from which the data should be imported. In the following the specifications for the different file types are explained:

itx files: no further information is required, the "Import" button will open a file dialog to select the .itx file and all waves in this file will be imported.

.xls/.xlsx files: in a first step, the user has to select the file by clicking on "Choose file". The user needs to define where to find the time series (only one column at the time), by giving the start and end cell, e.g. A2,A40. Similar for the data (concentrations). The user has to again give the start and end cell, however here more than one column is supported, e.g. B2,H40. The spreadsheet name also has to be defined. SoFi will import the data column-wise.

Row-wise import is not supported.

.csv,.dat., txt files: the user has to choose the file by clicking on "Choose file". First, the user has to define in which column the time series can be found.

<u>Please note, the column number starts from 0</u>. In the following three options, the user should define the time format that is used. "Format" defines the date format, while "*" is the date separator (e.g. ".", "/", "-") which has to be

specified in “Date separator (*)” (without “”). Whether the time series contains a time information and if so in which format, has to be selected under “Time”. In the next step, the columns which contain data need to be specified. “1st column of data” is the column from where the data should be imported (note, column numbers start from 0). “Nb of columns” defines how many columns should be imported. It can be either set to a certain number or to import all columns after the 1st data columns, set it to 0. To name the imported waves correctly, set the “Header (row)” on where to find the names (note, row numbers start from 0). If there is no header, Igor will name the concentrations waves “waveX”, with X being an increasing number. “Nb of rows to skip” defines how many rows should be skipped. The header is not automatically included. In “Column separator” the column separator needs to be defined. In case the columns are separated by tabs, please use \t. “Decimal character” is pre-defined as dot but can be changed into comma.

If data waves with the same name are being imported at different steps, they are not over-written, but a continuous number is added to the end of newly imported wave.

4.1.2.1. **AE33 data import**

SoFi offers the option to read multiple AE33 files automatically, extract the valid data and perform source apportionment (Sandradewi et al., 2008). In a first step, the raw data has to be imported. SoFi offers the option to only import (a) the date and total eBC (BC6), (b) date, all wavelengths and some diagnostics and (c) all available data. Note that with option (a) post processing is not possible (as the status is not imported) and depending on the amount of imported files, option (c), although recommended, can take quite some time. The imported waves will be saved under root:SoFi:External_data:Raw_ts:AE33_raw, a separate folder with all imported raw AE33 waves. All waves and folder names should be kept as imported otherwise further processing of the data might fail.

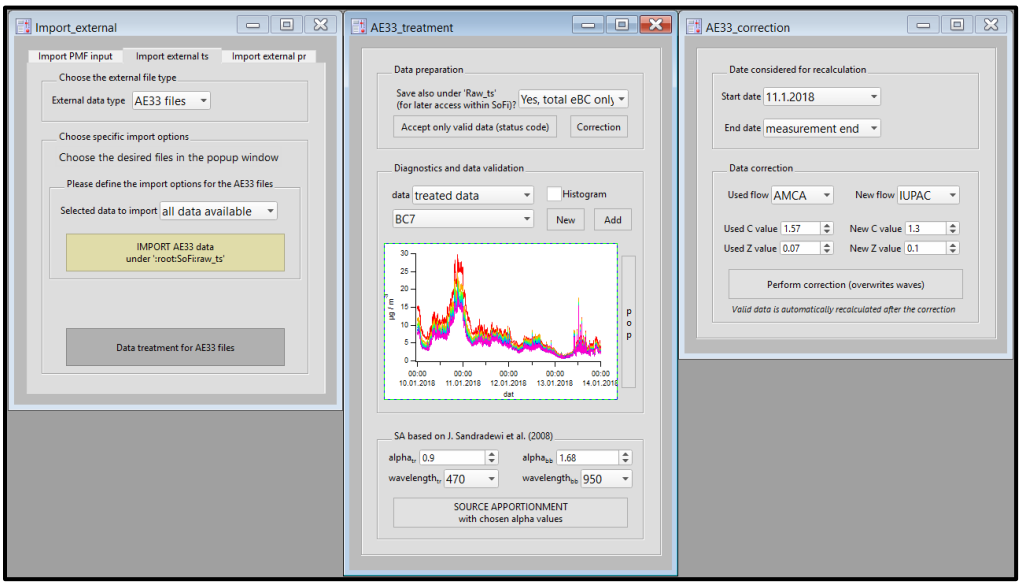


Figure 14 Interface to import and treat AE33 files. AE33 files can be imported, treated (valid measurements extracted) and corrected and displayed for a short checkup.

The data treatment is performed in a separate panel that opens through the "Data treatment for AE33 files" button. Typically, the raw AE33 data consists of the full time series, also including also filter tape advances, etc. calibrations besides valid measurements. For the data preparation, SoFi filters out all valid measurements (status = 0, 128, 256) and changes the units to what is commonly used. The valid data is saved under root:SoFi:External_data:Raw_ts:AE33_treated. As this folder is not accessed by SoFi when treating external data (yellow panel), the user has the option to save a copy of (a) total eBC (BC6), (b) all imported waves or (c) no waves under root:SoFi:External_data:Raw_ts for later.

Under diagnostics and data validation, some key parameters (eBC from all wavelengths, total flow, k parameter from all wavelengths as well as BCx vs ATNx_1) can be plotted for quick visual validation (normally or as histogram). The user can access the raw as well as the treated AE33 waves. The attenuation for the BCx vs ATNx_1 plot are ATN0 corrected. The AAE can also be plotted after source apportionment is performed.

SoFi also offers the option to perform source apportionment on the eBC. For more information on the source apportionment approach please consult the work published by (Sandradewi et al., 2008) or (Zotter et al., 2017). SoFi performs the source apportionment using the user-defined α values and the corresponding wavelengths. The source apportionment results are directly saved under root:SoFi:External_data:Raw_ts so that they can be easily used in SoFi later on.

It is possible to correct the filter tape settings (multiple scattering parameter (C) and filter leakage (Z)), as well as to recalculate the flow. These calculations can be performed over the whole imported data set or from/to specific dates. The dates refer to filter exchange dates (based on status code for tape change procedure). For the flow, the user can select from the five flow

reporting standards in the AE33 (Table 1). For the used flow reporting standard, multiple scattering parameter (C) and filter leakage (Z) value, the user should check the setup files ("AE_SETUP_AE33-[…]").

Table 1 flow reporting standards in the AE33.

Standard	Pressure	Temperature
AMCA	101325 Pa	21.11 °C
EPA	101325 Pa	25 °C
ISO	101325 Pa	20 °C
NIST	101325 Pa	0 °C
IUPAC	100000 Pa	0 °C

By pressing "perform correction (overwrites waves)", the recalculation overwrites all imported waves in the AE33_raw and AE33_treated data folder that are affected by the correction. The original .dat files are not changed.

4.1.3. Profile import

The external profile series need to be stored under **"root:SoFi:External_data:Raw_pr"**, which is done automatically if the "Import external pr" option is chosen. The minimum data that is required, is a profile series (variable vector) and a corresponding intensity series. In a first step, the user has to choose the file type from which the data should be imported. The required specification for the import is similar to the time series import, except that there are not information about the time format required. Please consult the information under "time series import" on how to set up the import and replace time series by profile series. SoFi imports the profile variable vector as a text wave, so that it can be directly used in the "treat external data" option.

4.1.4. Combine matrices

SoFi can combine matrices diagonally (e.g., for multi-time PMF), horizontally (e.g., for combining organic and inorganic matrices from the same instrument) and vertically (e.g., combining data from different seasons from the same instrument). The matrices have to be imported first, afterwards the user has to provide the name of the two matrices and the corresponding profiles/timeseries that should be combined. If more than two matrices should be combined, the procedure has to be repeated several times.

SoFi will automatically get rid of empty row/columns like during the initialization process as long as the missing value is either nan or zero. This is most important for diagonally combined matrices as these rows/columns would not be caught during the initialization process. If SoFi find a

minimum error wave with the name “minErr1ion”, SoFi will ask if the minimum error wave should also be applied during the combination of the matrices.

4.2. Define PMF input

The internal string variables in SoFi must be linked to the just imported PMF data. This is done using the dropdown menus “Data” (for the data matrix), “Error” (for the error matrix), “PR (nb)” (for the profile (variables) in numeric format), “PR (txt)” (for the profile (variables) in text format), and “TS (nb)” (for the time series in numeric format) (Figure 15).

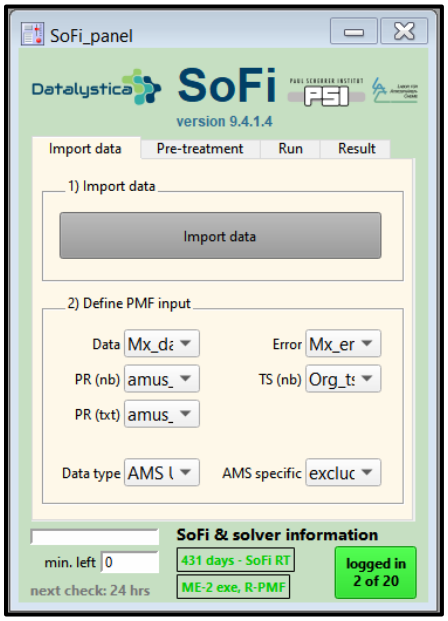


Figure 15 Main SoFi panel. Under “Define PMF input” the user has to define the PMF input.

4.2.1. The data and error matrix

The main matrices with the measured data and the corresponding error matrix. SoFi can calculate simple error matrices during the data import (see 4.1.1). The error matrix must contain the word “_err” in the name. This allows for an internal differentiation between the data and error matrix.

Large names, e.g. more than 15-20 characters should be avoided. This is a restriction given by the ME-2 solver.

4.2.2. The time wave

“TS(nb)” contains the numerical information of the time series. The dates do not have to be continuous but it is advantageous for the plots later on to avoid jumping back and forth when choosing line styles. In case of thermogram data, set define the temperature wave (either continuous or thermogram cycles) here are time wave, as PMF will performed over this dimension (i.e. treating it similar as the time wave in traditional PMFs).

4.2.3. The variable vectors

"PR (nb)" contains the numerical information of the variables. "PR (txt)" must be of text format and typically contains the text information of the variable, e.g. the name of the elements for elemental analysis or the name of the ions for HR-AMS analysis. However, even for unit mass resolution variables both vectors are required. In such a case the information is in both vectors the same but with a different representation, i.e., numerical for "PR (nb)" and text information for "PR (txt)". After selecting a wave for PR (nb), SoFi will ask if a text wave with the same information but in text format should be generated. This is useful for UMR data.

4.2.4. Data type

The dropdown menu allows specifying the details of the PMF input data, such as the type of data: Q-ACSM/Q-AMS (UMR), ToF-ACSM/ToF-AMS (UMR), HR-AMS, Xact, Figaero TD, others. This is important to define, if the CO_2-related variables should be excluded (ACSM/AMS) (see 4.2.4.1 and 4.2.4.2), for correct access to the spectral database (where available), as well as to specify units of the PMF input. For ACSM and AMS the unit "µg/m^3", for Xact the unit "ng/m^3" and for Figaero the unit "counts/s" and "° C" is automatically passed, respectively. For the other cases the user is prompted to define the units manually. The unit information is used in SoFi any time the contribution is visualized.

Currently, SoFi offers the following options for AMS/ACSM data:

Table 2 Data type options available in SoFi.

Option type	Comments
Q-ACSM / Q-AMS (UMR)	*m/z* 16, 17, 18, 19, 20 are based on m/z 44.
ToF-ACSM / ToF-AMS (UMR)	*m/z* 16, 17, 18, 19, 20 and 28 are based on m/z 44.
AMS HR / ToF-ACSM X (CO=CO2)	*m/z* 16, 17, 18 and 28 are based on CO_2 signal.
AMS HR (CO fitted)	*m/z* 16, 17, 18 are are based on CO_2 signal, CO is supposed to be fitted!
Xact	-
Figaero TD	PMF over Figaero-CIMS thermograms
other	-

4.2.4.1. Fragmentation table for UMR ACSM/AMS

Over time, different fragmentation tables have been used, leading to inconsistent data structures. To make sure corrections, especially those involving the exclusion of CO_2-related variables, are properly applied, the input data needs to be linked with the correct fragmentation table. According to Aerodyne, only one of the two fragmentation tables that are shown in Table 3 should be used, everything else is outdated.

Table 3 Fragmentation tables confirmed by Aerodyne. For HR-AMS the corresponding ions are shown in the brackets.

m/z	Q-ACSM/AMS	ToF-ACSM/AMS
16 (O)	0.04*frag_organic[18]	0.04*frag_Org[18]
17 (HO)	0.25*frag_organic[18]	0.25*frag_Org[18]
18 (H_2O)	1*frag_organic[44]	0.225*frag_Org[44]
19	0.000691*frag_organic[18],0.002*frag_organic[17]	0.000691*frag_Org[18],0.002*frag_Org[17]
20	0.002*frag_organic[18]	0.002*frag_Org[18]
28 (CO)		frag_Org[44]
44 (CO_2)	44	44

The user has the option to choose between the two fragmentation tables. SoFi checks automatically whether the fragmentation table in the input data matches the instrument type that has been defined by the user. If there's a mismatch, SoFi alerts the user and adjusts the input data, so it aligns with the selected instrument type. This can also be used for better comparison of data from different ACSM so that they all use the same fragmentation table.

4.2.4.2. Exclude CO_2-related variables (for ACSM/AMS)

Based on the fragmentation table (see Table 3), several ions are calculated based on the CO_2 measurement and are not real measurements. Excluding those CO_2-related variables is essential for obtaining a physically meaningful and stable PMF solution. Including them in the PMF input introduces redundancy, which violates the PMF assumption that each variable provides independent information.

This issue becomes especially critical when PMF is run in robust mode, which is standard practice. During the iterative optimization, individual copies of m/z 44 can be treated differently: one copy may be downweighted when its residual-to-uncertainty ratio exceeds the robust threshold, while other copies are not. As a result, identical signals are handled inconsistently, even though they represent the same physical quantity.

For this reason, SoFi removes all CO_2-related variables prior to PMF and adds them back afterwards using the appropriate fragmentation table. This ensures that the CO_2 signal is treated consistently and influences the solution in a controlled and physically meaningful way.

Simply downweighting the CO_2-related variables is not sufficient. Even with reduced weights, these variables remain in the PMF input. So the problem of redundant information and inconsistent treatment of identical signal persists. Furthermor, it can bias the solution toward explaining variability in CO_2, particularly affecting oxygenated factors at the expense of less intense but chemically important ions.

To guarantee that this exclusion is applied correctly, the fragmentation table in the data must match the selected instrument type. SoFi therefore enforces instrument-consistent fragmentation tables based on the instrument that the user selected, to ensure reliable and reproducible analyses.

4.2.4.3. **Minimum error for ACSM and AMS data**

For the ACSM and AMS data the user can further apply the minimum error to the error matrix, if available. It is important to pass the wave with the exact name “minErr1ion” (generated from the previous data analysis software SQUIRREL/PIKA or Tofware (ACSM)) under “root:SoFi:External_data:Raw_input:”. Do not change the name of this wave!

4.2.4.4. **Figaero TD**

Use this option to analyze thermal desorption aerosol mass spectrometry data from FIGAERO-CIMS, to gain insights into sources and processes based on desorption temperature profiles (volatility). Prior to the PMF analysis in SoFi, the data must be processed with **MaS-PET** (curtesy of Dr. Angela Buchholz) first to ensure correct data structure, calibration, and uncertainty treatment. Export following data from MaS-PET.

Import all PMF relevant waves (data and error matrix, temperature, numeric and text wave with variable information as well as idxSample) into “root:SoFi:External_data:Raw_input:”. When selecting this option, the user also has to give SoFi the time stamp of the original filters. This will later be used as the time the thermograms are averaged to. It needs to be a numerical wave with the same length as idxSample.

5. SoFi – pre-treatment

5.1. Treat missing data in PMF input

Every cell of the PMF data matrix represents an equation that is used during the ME-2 minimization process. Missing information in the PMF input must be treated accordingly and excluded from the set of equations. Two cases can be distinguished, a) entire row / column has no information, b) few missing data.

The two cases should be treated in this order a) than b) and the first step is to define the symbol for missing information under “Define symbol for missing data”.

a) Entire row / column has no information

Note that entire rows / columns are removed directly from the main imported input under “root:SoFi:External_data:Raw:input”, once the button “initialize PMF input & check for missing data” is pressed. The process is repetitive, in that if various symbols are present for missing information, e.g. NaN and 0 for AMS data, the symbol can be changed using the corresponding button and the initialization process can be repeated. This action also creates a copy of the current PMF input under “root:SoFi:Eff_input”.

<u>If more symbols representing missing data are present, the order matters and the user must start with the one that is uninterruptedly present throughout an entire row / column. This allows to correctly remove entire missing rows / columns.</u>

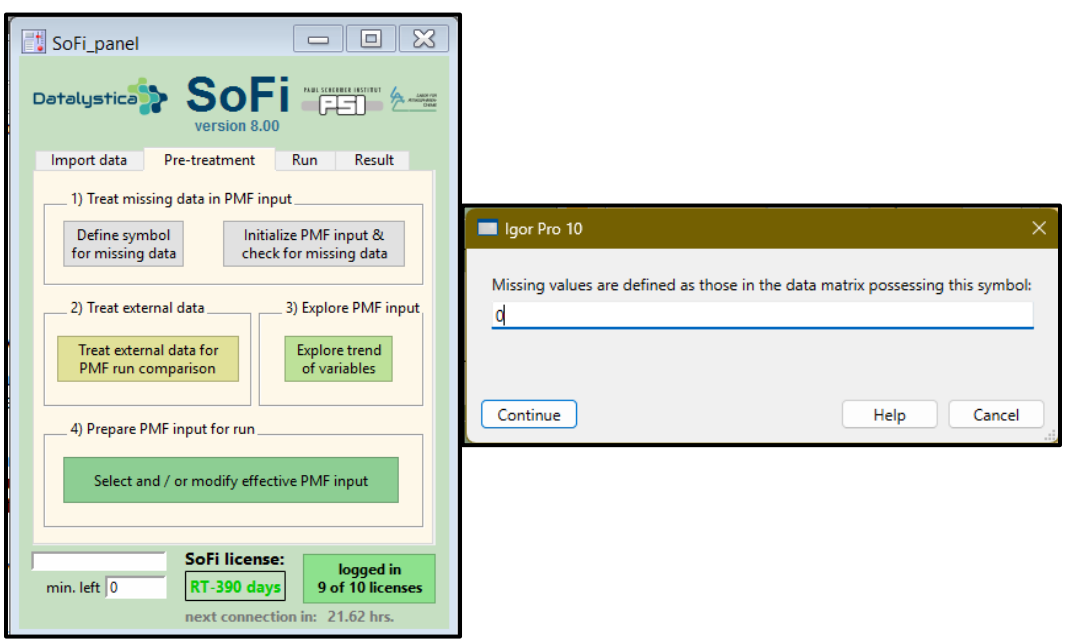


Figure 16 Main SoFi panel, pre-treatment tab, with the menu to define missing data symbols.

b) Few missing data points

At this stage, only completely empty rows and columns are removed. For the tagging and the exclusion of single missing data points, please consult 6.2.2 “II. Missing data”.

5.2. Treat external data

The external information is harmonized to match the PMF input data using the yellow panel by clicking on “Treat external data for the PMF run/result comparison” (Figure 16).

5.2.1. Profiles

The algorithm extracts the PMF variables present in the external vector(s) and stores the copy of the external vector(s) with the same length as the number of variables in the PMF matrix under “root:SoFi: Reference_pr”. Missing variables in the external vector(s) are defined as “NaN”s. The extraction works by comparing the text information, e.g. name of ions for HR-AMS of the PMF input with that of the external data. Therefore, the x-axis for the external data must be linked to the text information (Figure 17 left panel). Pressing the “SHIFT” key allows to select adjacent entries while the “CTRL” key allows the user to add non-adjacent entries.

It is possible that the reference profile does not include all variables that are present in the PMF input, especially in the case of HR data. Since SoFi performs an exact text match, these “missing ions” are left blank in the re-gridded profiles. If such missing variables (i.e., NaNs) are detected, SoFi asks the user whether a missing ion wave should be created. This missing ion wave will have values for all variables that are NaN. The value is based on a fraction of the variable with the highest intensity. The user can define the fraction him/herself.

5.2.1.1. **Online spectral database**

Datalystica provides a platform for reference spectra that can be directly accessed from within SoFi. Obviously, this requires an active internet connection. Based on the data type defined under “2) Define PMF input” in the “Import data” tab, the option “from database” is available if reference profiles for this datatype exist on the spectral database. Once the option “from database” is selected, the list of available reference spectra is shown in the table. The user can select the reference profiles that he would like to import and press on “Extract external information […]” and the selected profiles will automatically be downloaded, stored under “root:SoFi:External_data:raw_pr” (original reference profile), while at the same time, the reference profile with the same length and number as the PMF input (and the same fragementation table in case of AMS/ACSM) is stored under “root:SoFi:Reference_pr”.

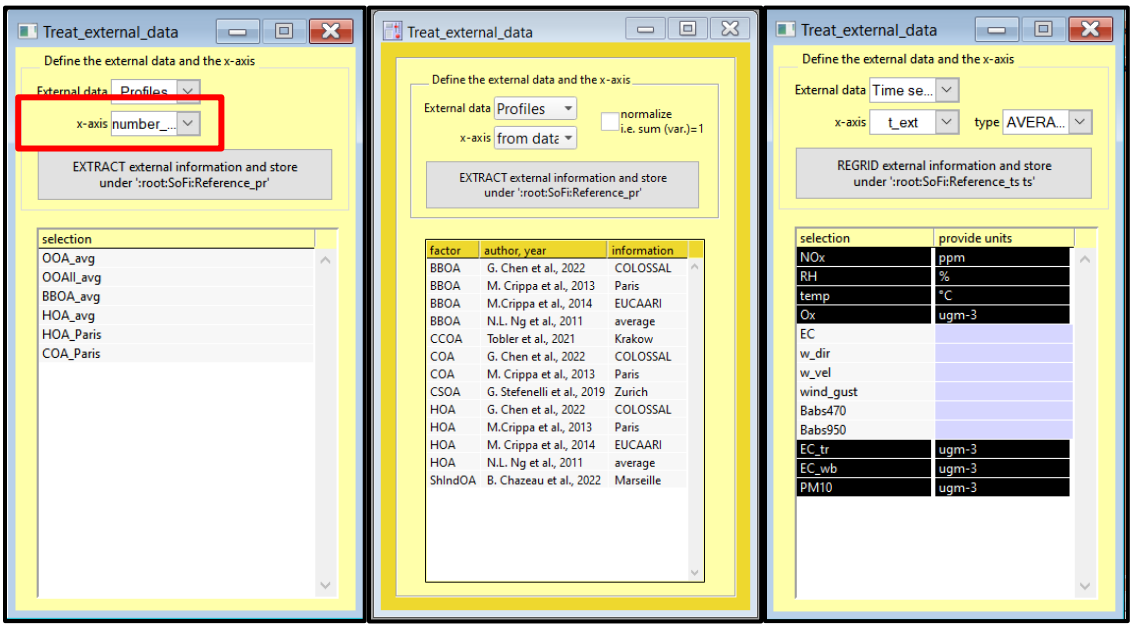


Figure 17 Left panel: extraction of reference profiles. Middle panel: reference profiles from database. Right panel: re-grid panel for collocated external time series.

5.2.2. Time series

The algorithm re-grids the time series of the external data. The following options are available:

"EXTRACT (same timep.)"	In case the time stamp of the external data equals that of the PMF input with the presence or absence of certain points.
"INTERPOLATE"	In case time resolution of the external wave is similar but not equal to that of the PMF input.
"AVERAGE"	In case the time resolution of the external wave is higher, e.g. PMF input resolution 30 min and external resolution 5 min.
(start/middle/end sav. t.)	defines where the data should be stored, either as the start / middle / end saving time. **Choose the option based on the saving time of your PMF input data.** The time series of the PMF input serves as the basis for the averaging, therefore the averaging option should be chosen to match the way that the PMF input data was measured.

The hourly, daily, weekly, monthly and yearly averages (as shown in Figure 18) are automatically evaluated. By default, the averages are not plotted. If for sanity purposes, the user wishes to get the plots for all averages, the checkbox "plot sanity check" should be checked. This option is very useful, when

- comparing PMF results with hourly, daily, weekly, monthly and yearly resolution
- quick inspections of PMF results, where the analysis with a high temporal resolution would require too much waiting time.
- comparing PMF results to external data that has a lower time resolution, e.g. PMF on AMS and 24 hours filter data as external data. Then the options hourly or daily can be used for comparison.

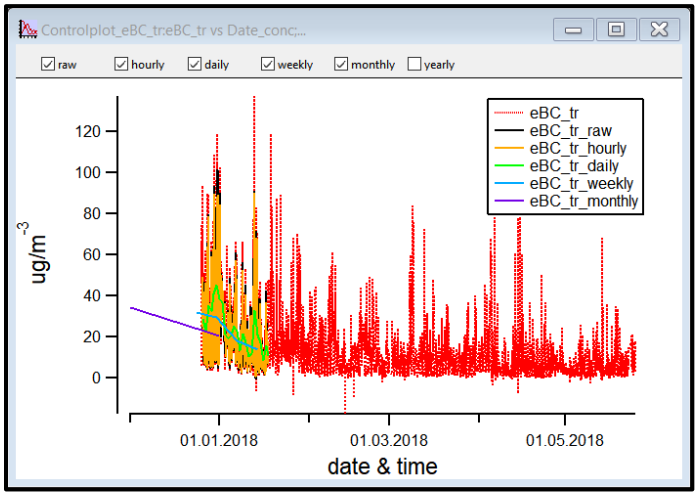


Figure 18 Sanity plot of temperature over time for default resolution (black, instrument resolution), averaged using the resolution of the PMF input (red, raw resolution), hourly resolution (green) and daily resolution (blue). Various resolutions can be enabled/disabled from the control bar on top.

The algorithm stores a copy of the external wave(s), with the same number of points as in the PMF input matrix under "root:SoFi:External_data:Reference_ts:raw" as well as the corresponding hourly, daily, weekly, monthly and yearly copies under "root:…:hourly", "root:…:daily, etc.". Missing variables are defined as "NaN"s. The x-axis is the regular time series information in numerical representation.

It is convenient to add the units of the external time series (Figure 17 right panel). This information is later used when the corresponding external time series is plotted in the result section. Pressing the "SHIFT" key allows to select adjacent entries while the "CTRL" key allows the user to add non-adjacent entries. After saving the external wave(s), a sanity plot pops and compares the raw time series and the re-gridded time series (Figure 18). This is performed for every chosen external time series.

5.2.2.1. **Wind data**

Wind data (wind direction and wind speed) are treated specially when averaged in SoFi. Wind direction is averaged by averaging the x and y components of the angles separately and merging them back together after having averaged (*direction dependent averaging*). For wind speed

information the normal average as well as the direction dependent average as described above for the wind direction is applied.

Note: wind direction is only averaged direction dependent when the name of the wave with the wind direction information is "**wind_dir**". Wind speed is averaged direction dependent when the name of the wave with the wind velocity information is "**wind_vel**". Moreover, the direction dependent averaging for the wind_velocity is not directly popped on the "Sanity_plot" graph (this graph shows the normally averaged wind velocity), but the wave is generated and stored under the corresponding Time series folder with the name "**wind_vel_degdep**".

5.3. Explore PMF input

It is highly recommended to perform a variable exploration prior to the PMF analysis. The user should get a feeling, especially the key variables, how they vary over time. For this reason, an additional panel was developed that pops after clicking on "Explore trend of the variables" (Figure 16). The explore data panel (green panel, Figure 19) has two adjacent subpanels that govern the options on the graph and on the data. The following options are available:

5.3.1. Options for the variable(s)

"plot"	point plot, image and HR-family (for HR-AMS data only)
"y-var."	a list of all variables of the PMF input matrix including external data
"x-var."	time, a list of all variables of the PMF input matrix including external data
"color- var."	normal (no wave), time, hours, any variable of the PMF input matrix including external data
"data"	all data, effective data (exclude blacklisted data = effective PMF input data)
"t-plot"	normal time series, diurnal cycle, weekly cycle, monthly cycle, yearly cycle
"type"	abs, fraction (a variable normalized by the sum of all input variables), signal to noise (S/N) (consult 3.3 for the S/N calculation)
"select based on class / time"	see 5.3.2 for more information

“moving window” only working when col-t is different than normal. The user can use the slider to move through to better understand the relationship at certain days, hours or concentrations.

5.3.1.1. **Point plots**

The point plots allow the user to explore the data, either as standard plot, i.e. variables or externals vs date/time or as scatter plots. If a color-scheme is chosen for scatter plots, a dynamic plot analysis is available, where one can scroll through the date by moving the slider on the left-hand side. By default, the effective input data (without the blacklisted data) will be plotted, however the user has the option to display all data.

Plotting the variables uses a string search algorithm in the background and therefore the names of the variables (in the variables text wave) must differ. If for example elements from various size fractions are present, then the element name must contain the information of the size fraction too, e.g. “Na” for the element in both size fractions is not sufficient, it should be “Na_fine” and “Na_coarse”, or a similar specification.

If fractional ACSM m/z 44 vs m/z43 is plotted, “Sally’s triangle” (Ng et al., 2011a)can be added automatically to the plot (Figure 19).

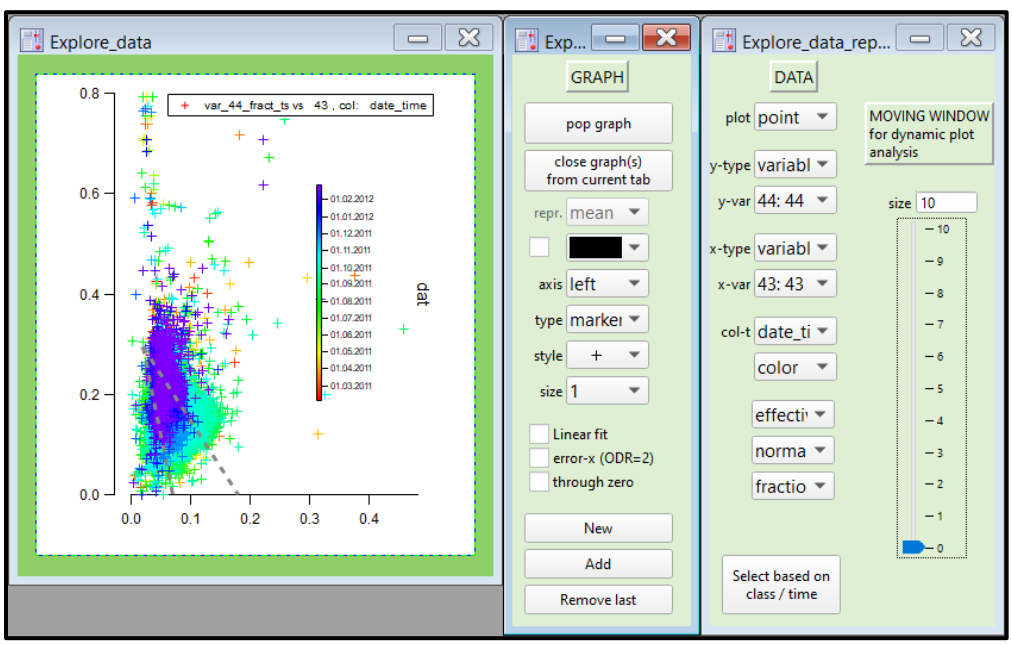


Figure 19 Panel for exploration of the input variables. Left part shows the graph and the subpanels on the right govern options of the graph and data.

By default, cycle plots in SoFi report the information in “start saving time”. The data of e.g. a daily cycle stored under 2 pm is representative for the data between 2 and 3 pm. This can be changed in the main settings.

5.3.1.2. **Image plots**

The user can also plot the entire data matrix using the option “image” from the very first drop-down menu (Figure 20). In case the data contains several outliers, the user can modify the color-scale using the slider (red rectangle) on the subpanel “GRAPH”. Moreover, the user has the possibility to select a region of interest from the image plot by either directly drawing a marquee or by using the marquee limits from the variable options (blue rectangle) and then popping this information using the options from the right mouse button (

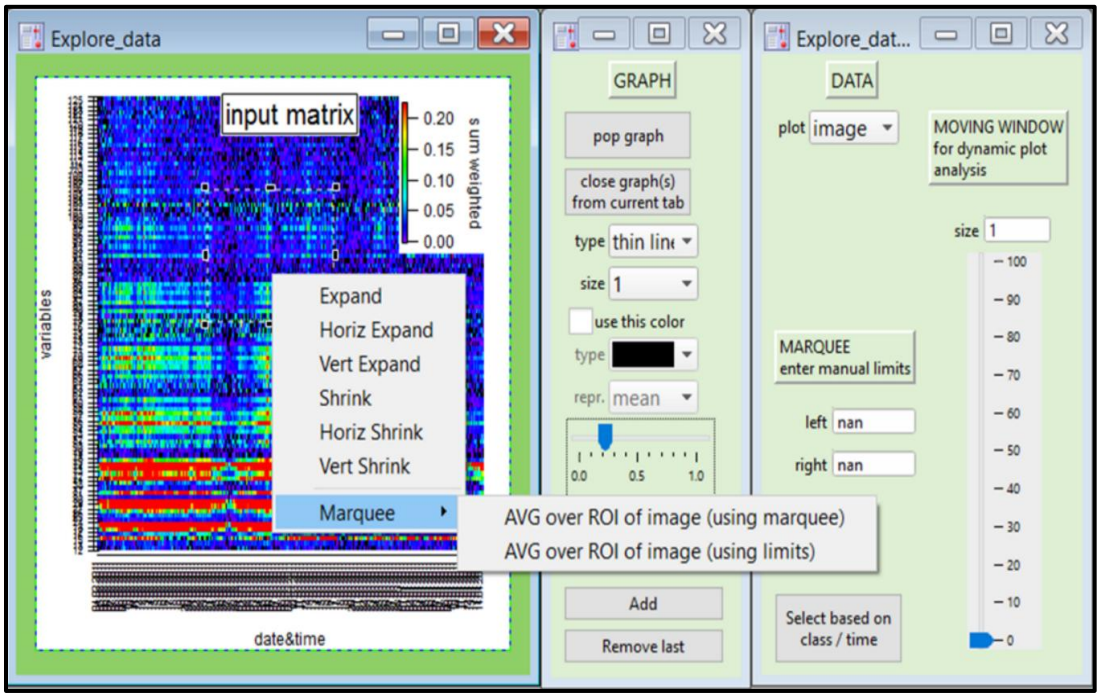


Figure 21).

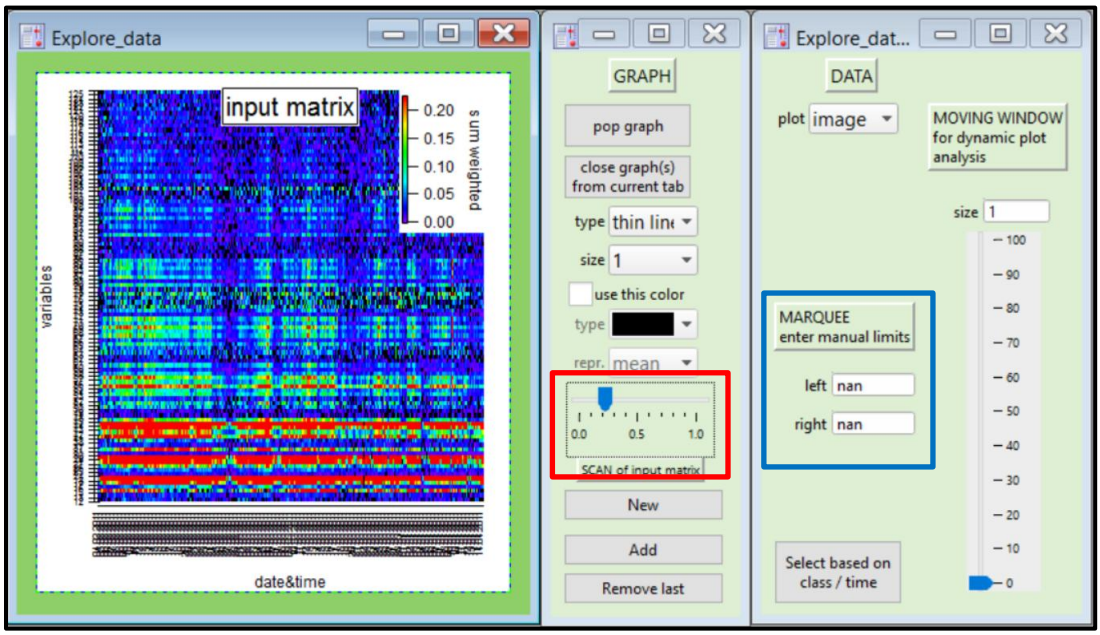


Figure 20 Image plot of the entire input matrix. The slider (red rectangle) can be used to vary the size of the color-scale.

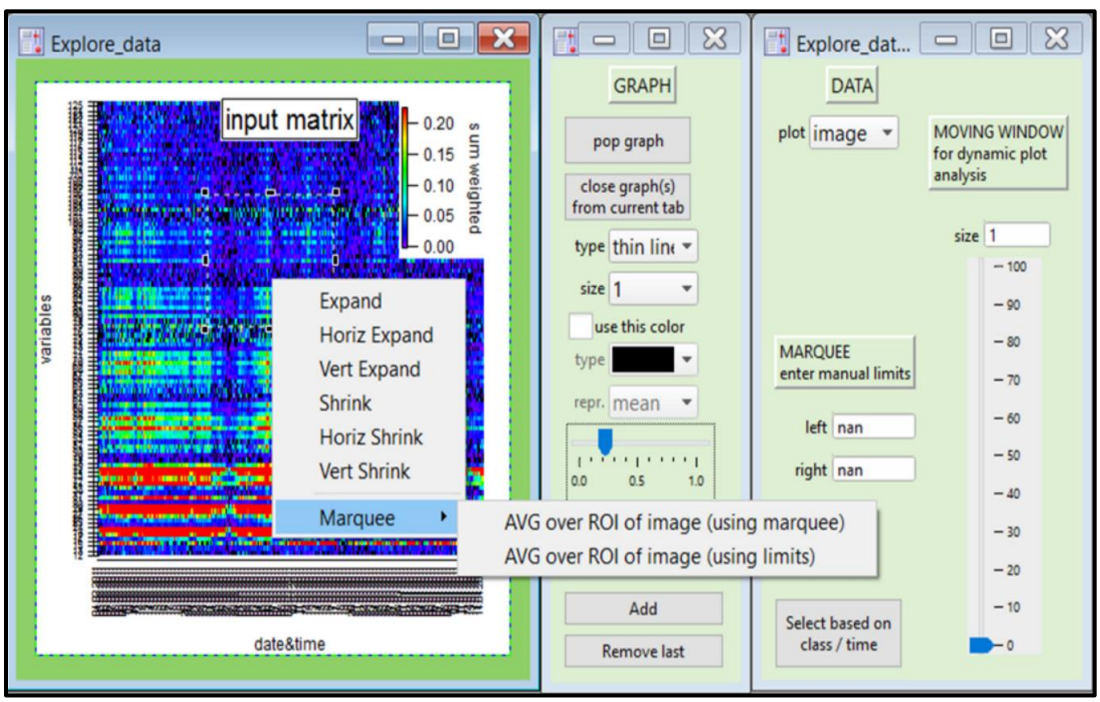


Figure 21 Image plot of the entire input matrix including popped information over the region of interest.

5.3.1.3. **HR-family plots**

For HR-AMS data, the user can inspect the OM:OC, O:C, H:C, N:C and S:C of each measurement point (Figure 22) to easily identify special events or measurement periods before performing PMF. As with the points plots, this figure can also be plotted as cycle plot.

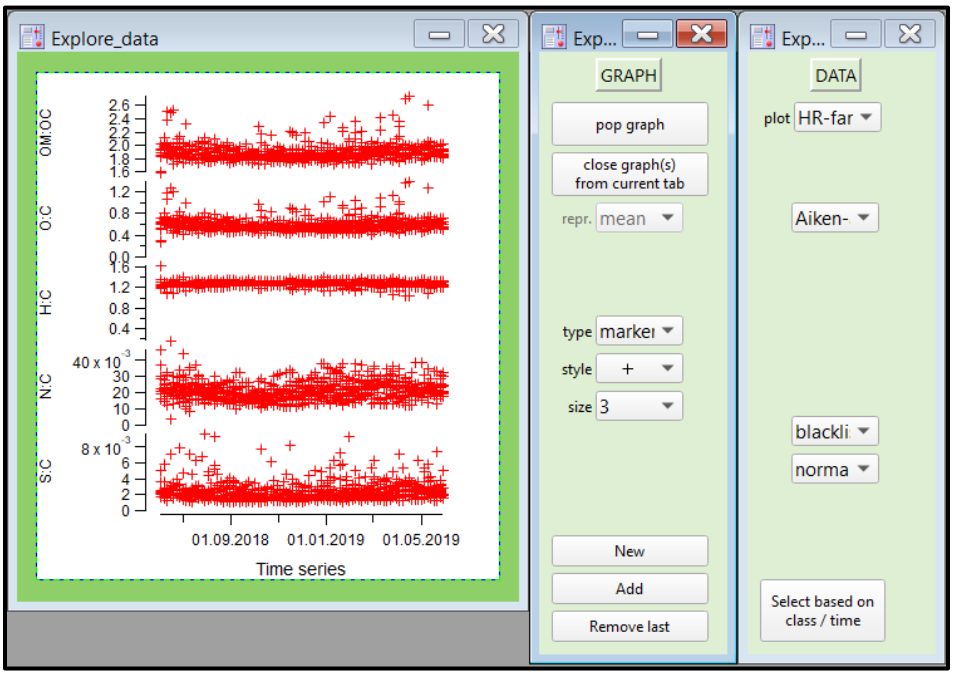


Figure 22 For HR-AMS the OM:OC, O:C, H:C, N:C and S:C of each measurement point is plotted.

Furthermore, the mass defect plot can be inspected. More details on the mass defect plot can be found in Sect. 8.5.6.2. Please note that if the user chose to exclude the CO_2-related variables, they are not added back into the matrix on which the calculation is performed, therefore the plot here and in the analysis panel (where the variables are added back to the matrix) might differ.

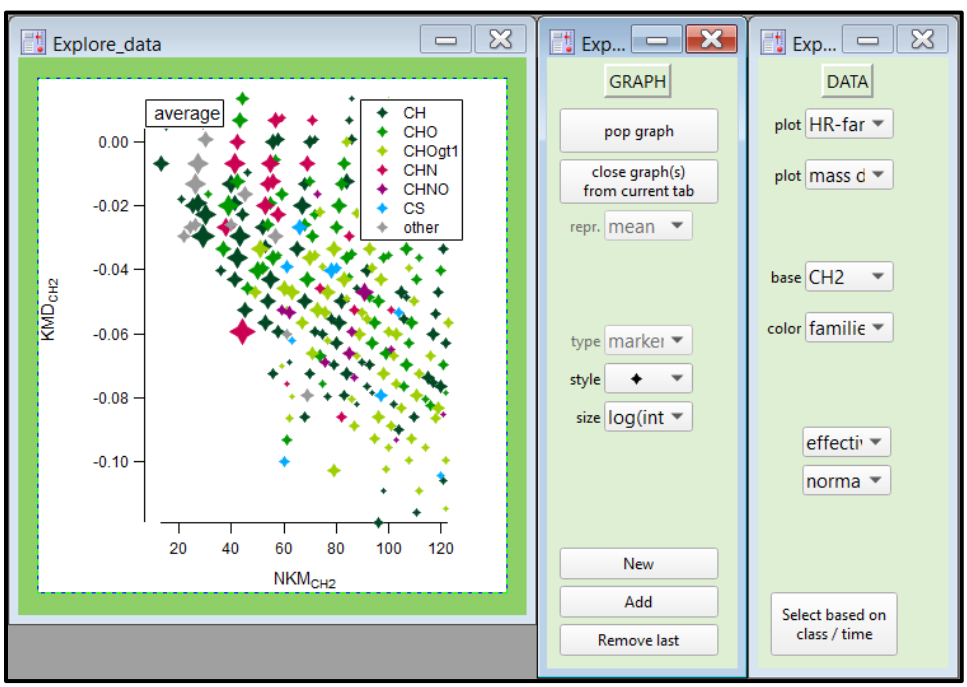


Figure 23 mass defect plot for HR-data.

5.3.2. Select based on class / time

This selection panel allows the user the easily and efficiently inspect data based on a multitude of options summarized in Figure 24. Here the options are discussed based on the three tabs separately.

"time-selection"	selection of hours, days, months and relevant years
"overview over time"	user can select several regions of interest with the marquee (draw a marquee around the period of interest (only points with both x- and y-coordinates falling in the marquee will be considered), click on the right mouse button and confirm under the menu "marquee". Note: this option is additive, i.e., the user can select and confirm various regions several times with the marquee tool. **The user can select under "time series wave" also external waves. This allows selecting points not only based on the PMF input, but also on external waves, e.g. investigating variables during episodes with high concentration of a tracer.**
"class"	All entries are selected as default, i.e., all points are used. The user can specify points to belong to certain classes and select among these when present from this tab (see 5.4.1 for more details on the class)

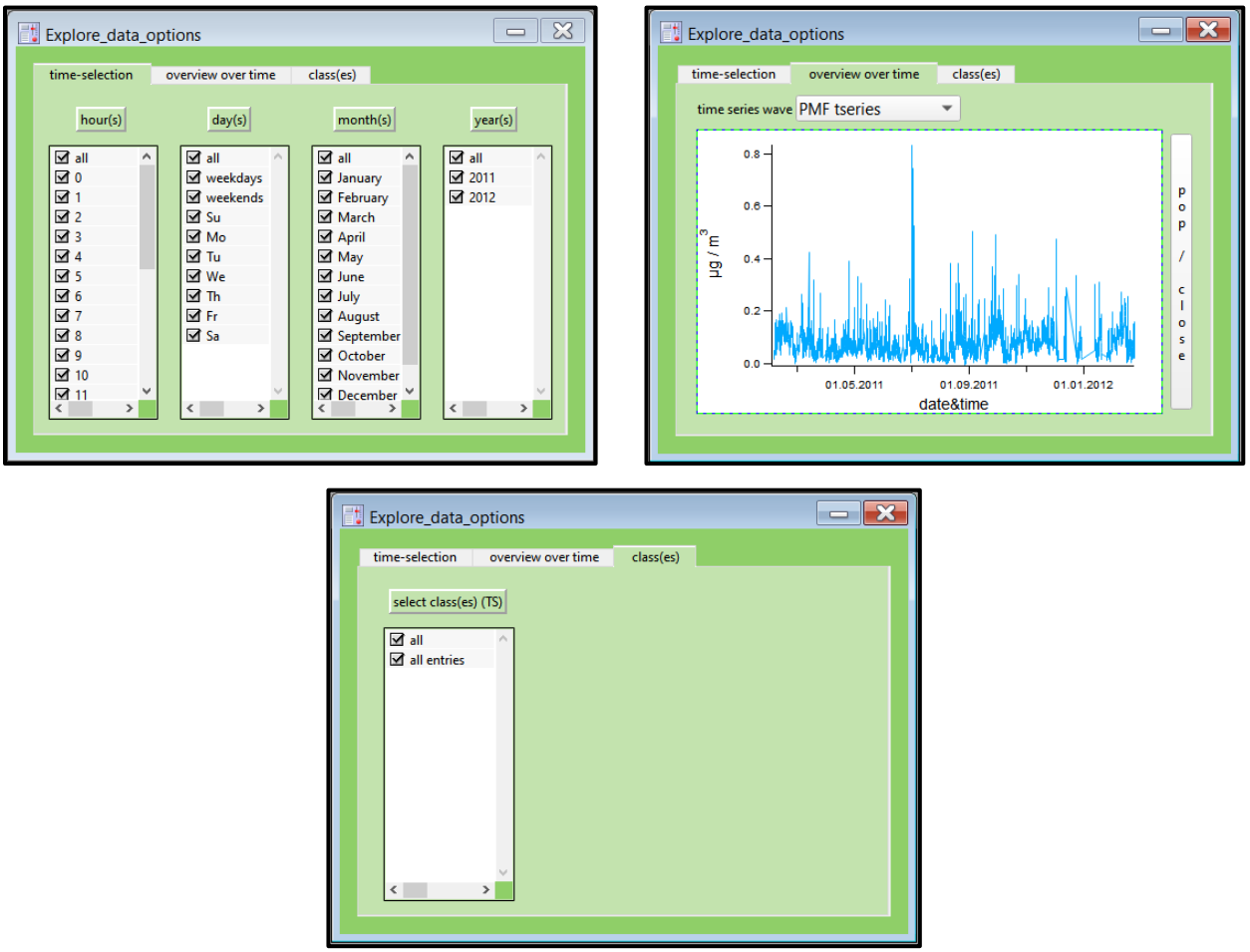


Figure 24 The three tabs of the selection subpanel based on time / class.

5.3.3. Options for the graph representation

"Pop graph"	pops the current representation from the panel
"Close graph(s)…"	closes all graphs related to the exploration panel
"type"	user can choose between various default trace representations, e.g. solid line, dashed line, etc.
"size"	size of the trace
"repr."	Option only available for cycle plots, where mean or median and the corresponding spreads (standard deviation, interquartile range (IQR), IQR with 10th and 90th %, IQR with percentiles and min/max) can be added to the representation
"use this color/define color"	user can define color for the current trace. If nothing is specific, SoFi uses colors from the color palette starting with red and

	moving row-wise and using every 2nd color (avoids using similar colors twice).
“linear fit”	performs a linear fit, makes only sense when performing a scatter plot
“error-x..”	error also assumed for the x-wave and least squares are not vertical but rather orthogonal to the fit (ODR = 2 in IGOR)
“through zero”	fit is forced to go through zero, i.e., intercept = 0
“New” / “Add”	Chosen data can be plotted using the button “New”, or added to an existing graph using the button “Add”.
“Remove last”	The last trace is removed from the graph

Two graphical examples for this exploration panel are shown in Figure 25.

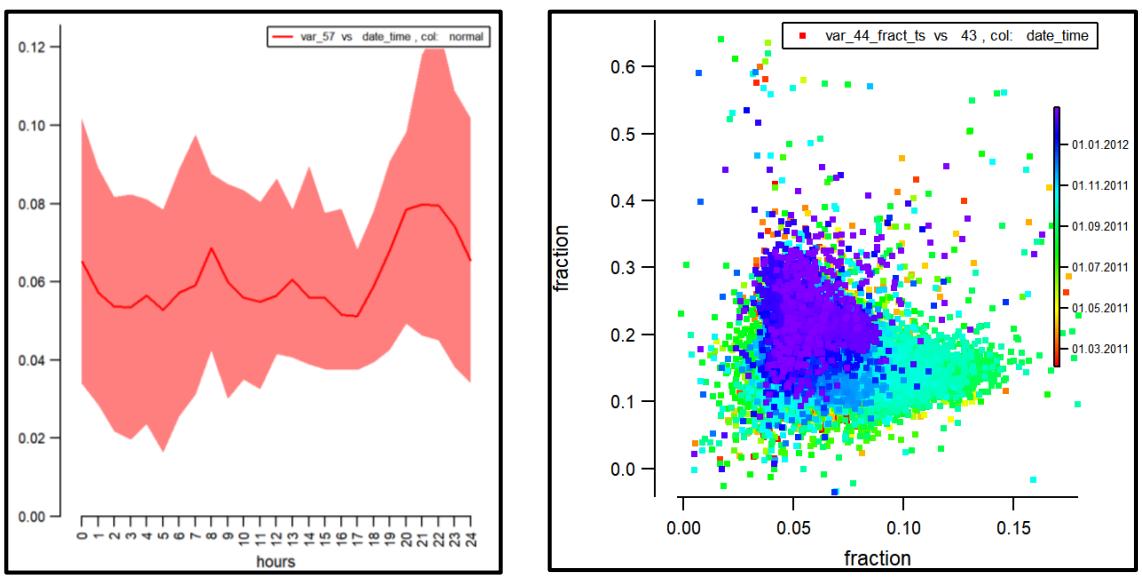


Figure 25 Daily cycle of *m/z* 57 mass (left), scatter plot of *f*44 against *f*43 colored over time.

5.4. Select or modify effective PMF input

5.4.1. Class(es)

Variables or time points can be grouped together into classes. Classes are a powerful tool for:

“PMF input”	PMF can be performed over certain classes of variables or time points only, e.g. PMF over air masses affected by continental air

	(back-trajectory data would be a class wave), firework period vs non-firework period.
“C-value”	classes can receive different relative errors when performing PMF (see 6.2.4.4) for more details). This option is only available for variable classes.
“Result inspection”	PMF results over Profile / Time series can be shown for defined classes only. This can be interesting if e.g. PMF is run over filter data from PM_{10} and $PM_{2.5}$ filter data and some results should be inspected in detail for the respective size fraction.

Classes can be defined in two different ways in SoFi.

<u>Import from preexisting text file</u>

Classes have already been defined beforehand and need to be read and linked in SoFi. Class waves can exist for variables or time points and the user governs the option based on the “Profile, Time series” dropdown menu on the “PMF_input_sel” panel (upper red rectangle on Figure 26). The dropdown menu “class names (txt wave)” reports the list of waves contained under “root:SoFi:External_data:Raw_input:”, where the unprocessed PMF input is stored (lower red rectangle on Figure 26). The class waves must be in text format and possess the same length as the PMF input (over variables / time points). It is important that all variables / time points belong to a specified class, if a class wave is read. If not, the user must complete defining the untagged variables / time points directly in SoFi.

Figure 26 Create classes based on text wave. The text wave has the same dimension as the time series and assigns each time point to a class.

Define classes directly in SoFi

If no classes are defined from external class waves, all variables / time points belong to the default class “all entries”. The user can define classes for variables in SoFi by entering into “type index”:

Index based blocks: use “–” for a range and “;”to separate single indices or ranges and press "create new block” to create the blocks and name them in the popped-up window (red rectangle on Figure 27).

Monthly blocks: use the keyword “monthly” to create blocks month-wise. The blocks will automatically be named after the month and year of the corresponding time point.

Hourly blocks: use the keyword "hours" followed by the hours that should be used for the blocks. E.g. “hours17-20” will create one block that includes all points that are between 5 and 8 pm.

Minutely blocks: use the keyword "minutes" followed by the hours that should be used for the blocks. E.g. “minutes10-20” will create one block that includes all points that are between minutes 10 and 20.

Fixed length blocks: use the keyword “block” followed by the number of point you want to have grouped together. E.g. “block10” will create new blocks every 10 points. Please note that this format is only supported for data with low time resolution.

Instrument-wise blocks: use the keyword “instruments” followed by the number of instruments you have in your data. The data of the instruments must be added together along the diagonal. E.g. “instrumets2” will create two blocks in the time series and the profiles automatically. For the instrument names a window pops up, asking for the name of the instrument, i.e. group name.

Thermogram-wise blocks: for thermogram data, where PMF is performed over a series of thermograms instead of a time series, use the keyword “idxSample”. Based on the provided filter index, a new group for each thermogram is created.

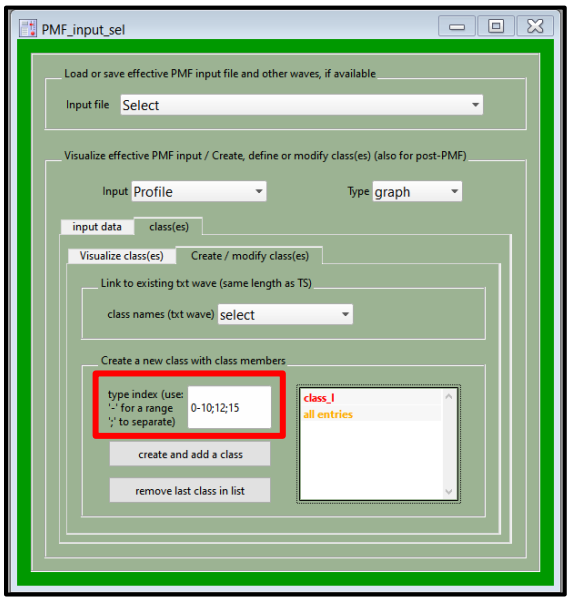
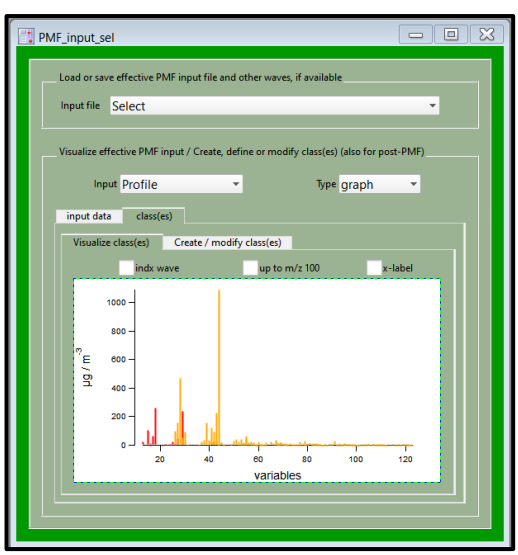


Figure 27 Define class for the variables at position 0-10 including variables at position 12 and 15. This class received the class name “class_I” and a red color. Colors can be changed by right clicking on the entry for which the color should be changed.

Clicking on “create and add a class” prompts the user to define the name and the color for the new class. In the example (Figure 27) the class “all entries” is still present, as only some variables have been classified so far. As soon as all variables or time points are members of user-defined classes, the default class “all entries” disappears. The user can inspect the grouped variables or time points under the tab “Visualize class(es) (see Figure 28).

The colors are assigned by SoFi based on the number of classes. They can be easily changed by clicking right on the class where the color should be changed and from the color pallet a new color can be chosen. In case there are typos or wrongly chosen names, double clicking on the class name allows changing the name.



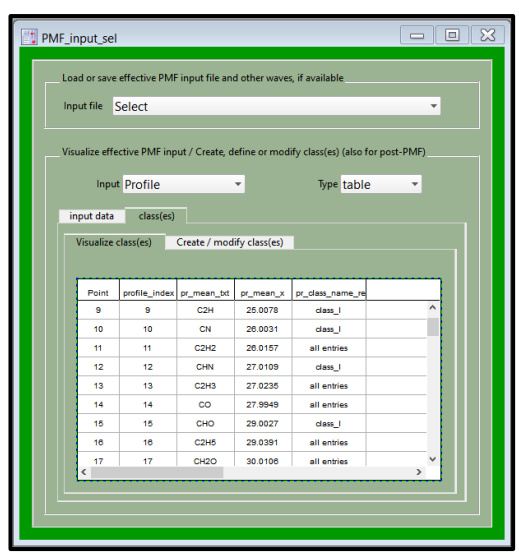


Figure 28 Inspection of the just defined class “class_I” as graph (left) and in table representation (right). The plot on the left colors the variables with the user-defined color. All other variables that currently still belong to the default class “all entries” are visualized in black.

Defining classes should occur before the blacklisting process as the class txt waves always refer to the entire PMF input. First blacklisting and then defining groups of variables or timepoints could cause some unexpected problems when visualizing the results later in the result panels.

5.4.2. Input data

Input data used for a PMF run can be modified using the blacklist function. This is advantageous, when points in time (spikes due to transient events, electronic noise or a PMF over only specific points is sought) or some variables (problematic and/or unnecessary variables, e.g. very noisy variables or variables with very little variability or a PMF over only specific variables is sought) shouldn’t be considered for a PMF run. The PMF input (green panel) shows the current PMF input, whereas blacklisting is performed on the adjacent subpanel (see Figure 29). The selection subpanel contains the same options as those already presented for the subpanel under 5.3.2 and the user is referred to this section for the overview of the options.

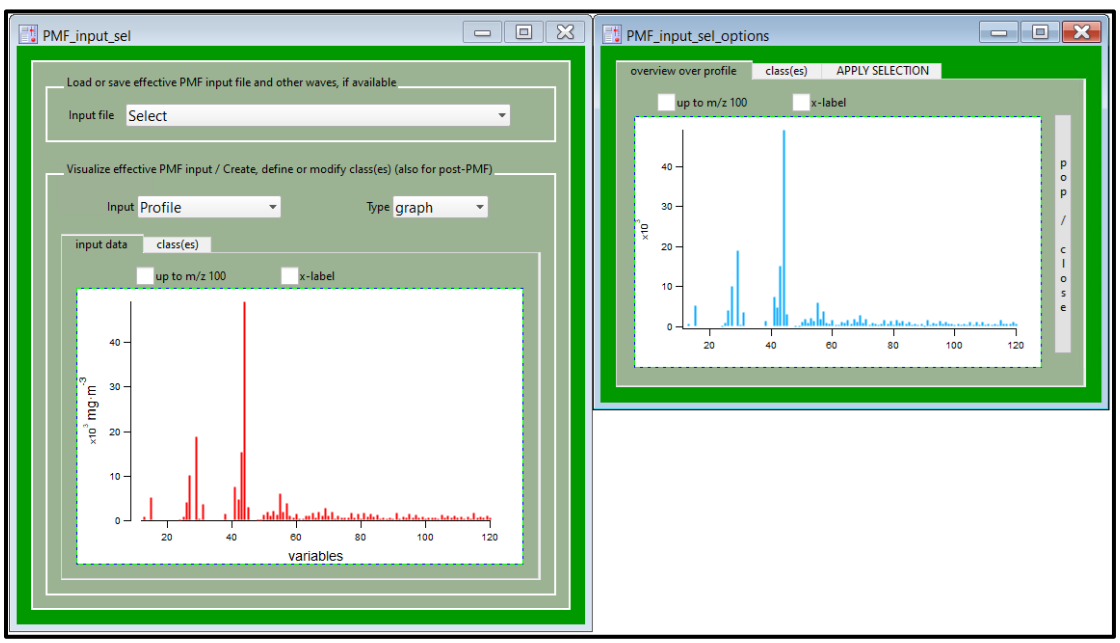


Figure 29 PMF input panel useful for controlling the user-specified PMF input and for tagging points in time or variables into classes.

The default case considers all points in time and all variables for the PMF input. Removing some points or variables requires the user to select options either using checkboxes, the marquee function or the cursor (for variables only) on the subpanel. Once the selection is completed (it is an additive process and must be performed for the profile and the time series separately) (see Figure 30 for an example over time), the user can blacklist the selection by pressing on the button

“BLACKLIST / REMOVE SELECTION…” on the third tab on the adjacent subpanel (see Figure 31). This will update the graphs and tables from the PMF input panel and the blacklisted data is supposed to disappear (see Figure 32).

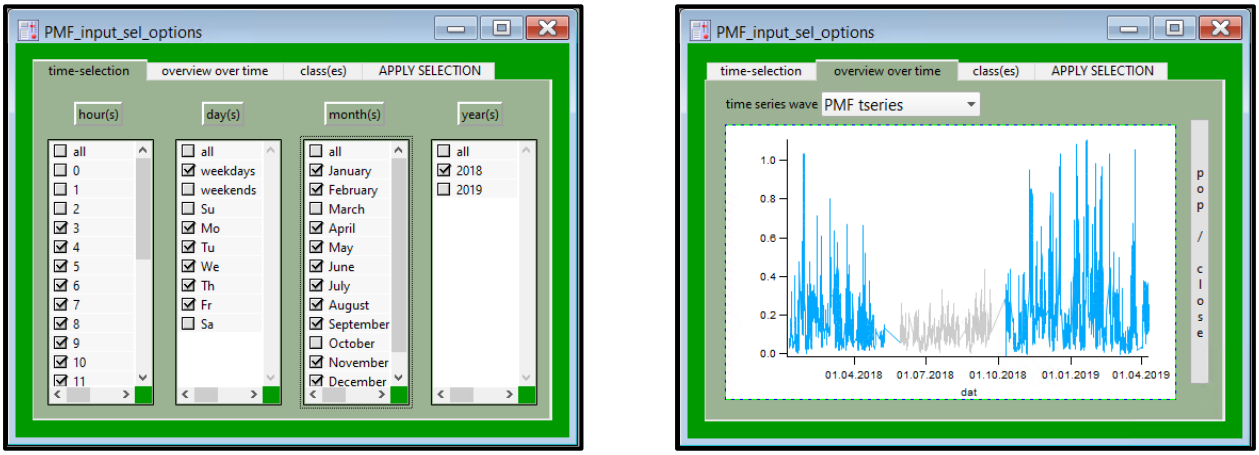


Figure 30 Left: example where points between 0 and 2 am during the weekends for the months March and October for 2019 are blacklisted / removed. Right: example where middle period is blacklisted / removed by marquee selection.

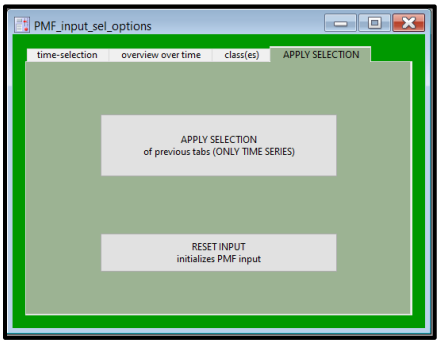


Figure 31 PMF input panel subpanel showing the third tab with the “BLACKLIST / REMOVE SELECTION…” and the “RESET INPUT…” buttons.

Figure 32 Left: default PMF input, i.e., all points over time and all variables are considered. Right: PMF input after being blacklisted by the user.

Blacklisting removes the selection, i.e., the user is supposed to unselect what should not be present later on in the PMF input. If some points have been erroneously removed, the only way to restore the PMF input is by clicking on "RESET INPUT" from the third tab. This button calls the option "Initialize PMF input & check for missing data" (Figure 16) and reinitializes the entire PMF input matrix (see 5.1).

User-specific PMF input data can be stored and loaded from IGOR txt files directly from SoFi under "Input file" (see Figure 29). Saving the PMF input will also store user-defined values for **F** or **G,** if already present and their corresponding *a*-values (see 6.3.1 and 6.3.3). This allows the user to store specific PMF input settings or initial values and constraints that can be loaded and tested, later, if desired.

5.4.3. Blacklist single points based on MDL

SoFi offers the option to blacklist single measurement points of selected variables based on a user-defined threshold. For this, the variable has to be selected and a threshold defined. By clicking on the button "Blacklist all data below threshold", SoFi will go through all measurements of the selected variable and all entries below the threshold will be set to missing data threshold value minus one. The missing data threshold value is defined in the orange panel (see 6.2.2). By default it is set to -99. This value, as well as the error matrix replacement value, should not be changed after blacklisting any value here otherwise the data won't be properly excluded.

Variables can also be "un-blacklisted" by clicking on restore all data for chosen variables and all data point will be reset to the original data and error matrix.

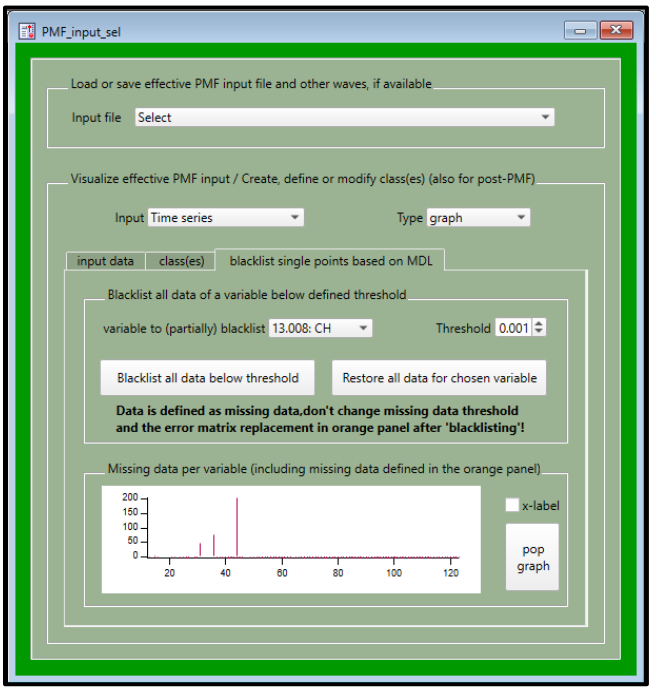


Figure 33 blacklist single points measurement points of a variable based on a user-defined threshold (e.g. detection limit).

The graph shows the number of data points that are being excluded from the PMF analysis. Either because they were blacklisted here based on a threshold value, or because there were empty cells that were treated in 6.2.2. So this graph not exclusively shows data below the threshold but also missing data in general (after being treated as missing data in the orange panel). The graph can be popped and also x-labels (variable names instead of numbers) are available.

6. SoFi – PMF run(s)

6.1. Define path, nb. of executables and name of HDF5 file

The “PMF run” tab is fully dedicated to set a specific PMF call using the PMF input and error matrices defined under the tabs “Import data” and “Pre-treatment” and to finally call ME-2 (see Figure 34).

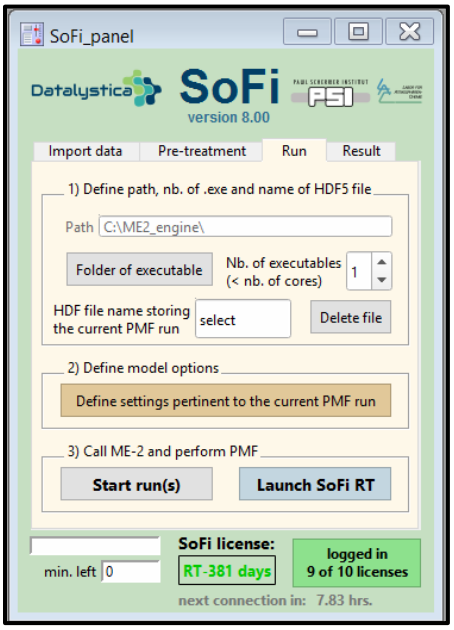


Figure 34 Main SoFi panel, PMF run tab.

6.1.1. Path

The user gives the location where the Me2.exe file resides by pressing the button “folder of executable”.

<u>The user need to have writing rights at the location where the ME2 engine folder is placed in, as hdf files will be generated and written into.</u>

6.1.1. nb. of executables

The user defines the number of ME-2 instances that will run in parallel. This number must be less than or equal to the number of cores in your PC.

Example: If you chose 4 cores, SoFi will assign to any idle ME-2 instance out of four instances a new task during the solving process.

<u>It could happen that some tasks are calculated faster (or slower) than the others. Therefore, the DOS windows are not necessarily all popped at the same time. Moreover, the run numbers in the</u>

log file / command history will not necessarily be in increasing order. Especially with rolling PMF, the initialization time for the ME2 engine might be longer than running PMF so that not all cores will be taken advantage of (e.g. with rolling PMF).

6.1.2. HDF5 file name

The user defines a proper name for the PMF run under "HDF5 file name storing the current PMF run". If this name already exits, SoFi prompts the user to try with another name.

!!! The "Delete file" button removes the specified HDF5 from the OS. Despite a last warning prompt, the user should pay attention not to delete important information. SoFi does not control hidden back-ups of the HDF5 files !!!

6.2. Define general model options

The orange button "Define settings pertinent to the current PMF run" (Figure 34) pops the panel governing all PMF settings (Figure 43).

6.2.1. Define number of factors

"starting factor"	defines the minimal number of factors (p) investigated by the model.
"ending factor"	defines the maximal number of factors (p) investigated by the model.
"nb. of PMF calls"	is the number of calls with the identical settings. SoFi will automatically adjust this value for fpeak analysis (see 6.3.2 for more details), for systematic a-value sensitivity analysis (see 6.3.3 for more details) and for C-value analysis (see 6.2.4.4)
Example:	"starting factor" = 1 and "ending factor"= 5 the model will return solutions with 1, 2, 3, 4, and 5 factors.

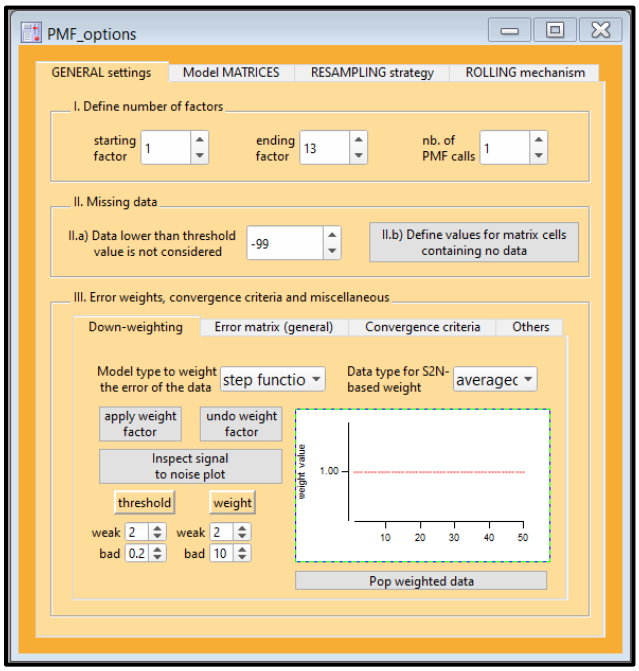


Figure 35 PMF settings, tab “General settings.

6.2.2. Missing data

The ME-2 solver needs clear instructions on what is a missing value. In a first step, during the initialization, completely empty rows and columns were removed, as described in 5.1. Since there might also be single measurements points that are missing/below detection limit, these must be defined here. <u>At this stage the PMF input must contain only one symbol for missing data, e.g. “NaN”, “inf”, “0” or “-999”, but not more than one.</u>

Clicking under “II.b) Define values for matrix cells containing no data” prompts the user to define the symbol for missing data together with an uncertainty value for its error cell (see Figure 36). Missing data that receives values that are smaller than the threshold value defined under A.II) (by default -99) are automatically excluded from the minimization process in ME-2. For safety reasons, their errors are also replaced by a very high value, e.g. 1000 times the highest error.

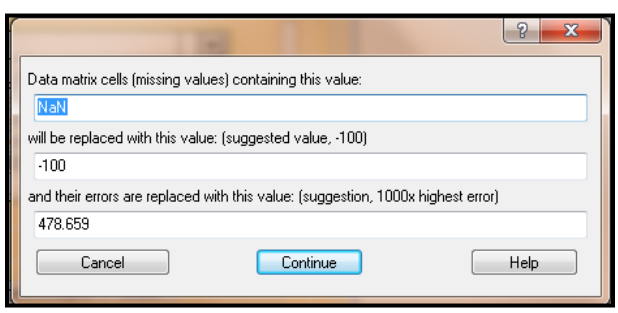


Figure 36 Prompting the value for missing data, the negative entry and its corresponding error value.

6.2.3. Down-weighting

Briefly, if the estimated measurement errors are uncertain, typically the case around detection limit and especially, if the systematic part of the error (background or blank) is not considered or poorly estimated, one could consider increasing the error / downweight the error. The weight strategy is based on the signal to noise ratio (see 3.3). Following options are available for weighing the errors (Figure 37):

<u>Model type</u>

"select"	the error matrix is untouched
"step function"	the two-step function proposed in Paatero and Hopke, (2003) is applied to the error matrix. The user defines the steps and their corresponding weights. By default, a variable is considered weak if the S/N is smaller than 2 and is downweighted by a factor 2, a variable is considered bad if the S/N is smaller than 0.2 and is downweighted by a factor 10.
"1/S2N function"	the error is continuously downweighted with the penalty function 1 / S2N for S2N lower than 1 and untouched otherwise. This idea was first discussed in Visser et al., (2015).

<u>Data type</u>

"averaged"	weighing the data is based on the averaged S/N ratio
"cell-wise"	weighing the data is performed on each cell separately.

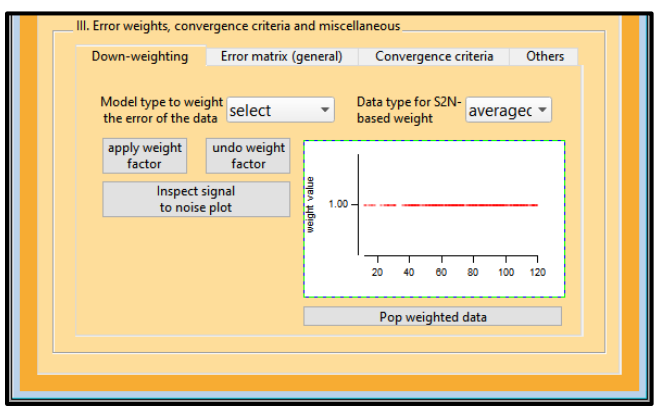


Figure 37 PMF settings, tab "General settings", down-weighting.

The weighted data is shown in the small graph. For more details, the S/N plot as well as the affected errors can be inspected using the buttons "inspect signal to noise plot" and "pop weighted data", respectively (Figure 38). <u>The weight can be applied and removed with the buttons "apply weight factor" and "undo weight factor", respectively.</u>

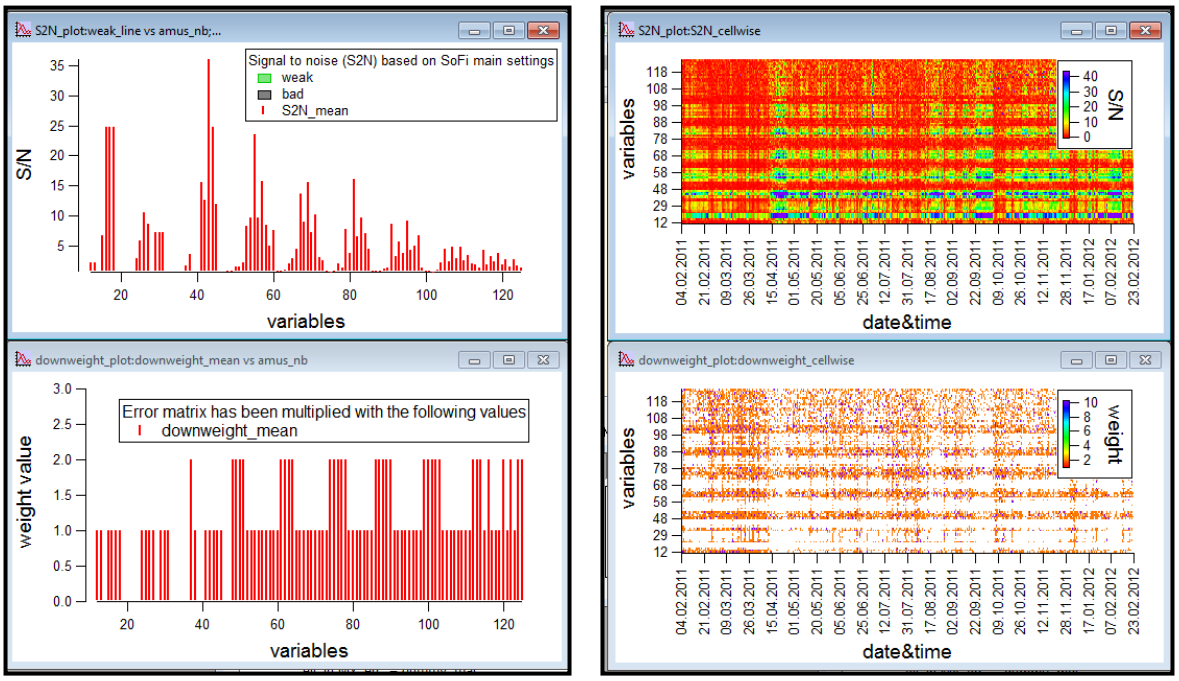


Figure 38 Signal to noise (top) and weighted errors (bottom) for the averaged data on the left and cell-wise on the right.

If the users is treating AMS / ACSM data and chooses the option "consider all variables" (not excluding the CO_2 related variables), the option for downweighing those variables is available (Figure 39). The user has the option to apply or remove the applied weight using the buttons "down-weight CO_2 related var." and "undo CO_2 weight", respectively.

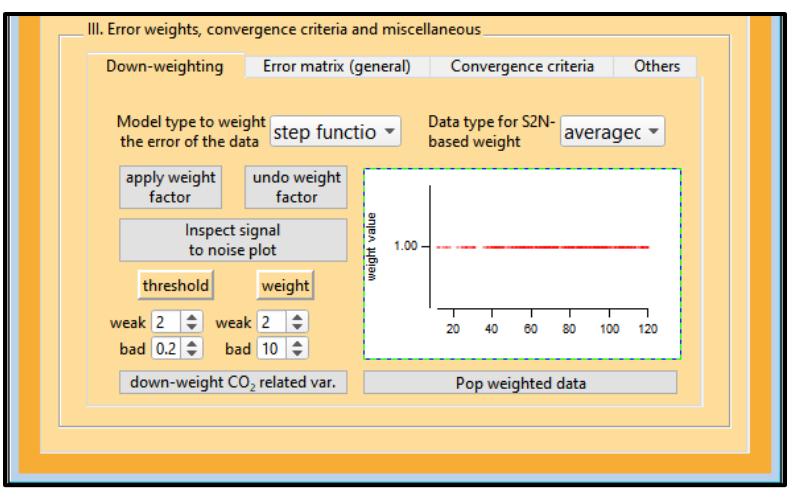


Figure 39 PMF settings, tab "General settings", down-weighting, showing the CO_2-related button.

<u>Many of the just mentioned free parameters have not been systematically tested and hence no recommendations can be given at this stage. There are plans for a study that will address the sensitivity of PMF results with respect to these settings.</u>

6.2.4. Error matrix (general)

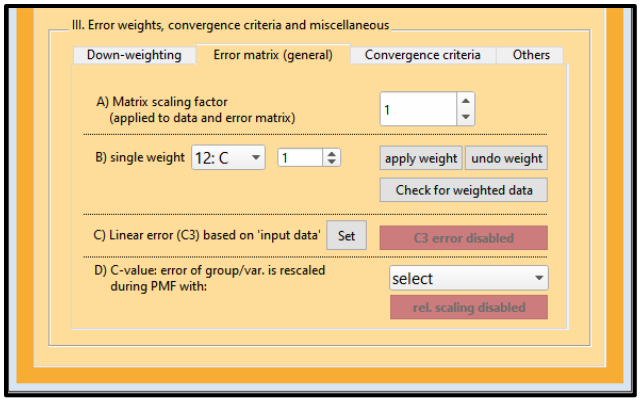


Figure 40 PMF settings, tab "General settings", error matrix (general).

6.2.4.1. Matrix scaling factor

If the smallest error that is passed on to the ME-2 solver is too small, the solver will not run. This can be critical for data sets with very low mass concentrations. In this case, a matrix scaling factor can be applied, meaning that the data and error matrix will be multiplied with this factor when sending the data to the solver. The input data in SoFi will not be affected. When importing the PMF run, the same matrix scaling factor is applied to the time series. So overall, the PMF result will be the same (as the scaling is applied to the error and data matrix and the result is normalized (profiles) and rescaled (time series), respectively).

6.2.4.2. Single weight

Single variables can be down-weighted using the drop-down menu for the variables and the weight value (Figure 40) and the applied weights always easily inspected (Figure 41). Apply single weights with prudence and based on logical reasoning, e.g., uncertainties of a variable are not well-characterized, PMF result is not effectively apportioning this variable, etc., as each weight must be justified in the end. Ideally, a sensitivity analysis of these single weights is also performed to assess the impact on the PMF result.

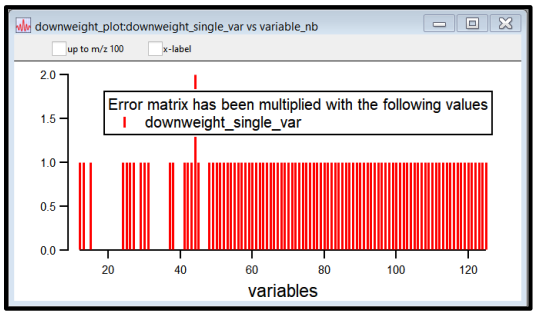


Figure 41 Overview over the user-defined weight values. Here the error of *m/z* 44 was multiplied by a factor of 2.

6.2.4.3. **Linear Error**

In SoFi, the linear error expression (C3 expression) present in ME-2 can be added and enabled for the PMF calculation (Figure 40).

The current error matrix will be updated accordingly:

$$error\ matrix[i][j] = error\ matrix[i][j] + C3 \cdot data\ matrix[i][j] \tag{17}$$

Clicking on "A) Linear error (C3) based on 'input data'" allows the user to define the C3 value. Once the C3 variable in IGOR is different than 0 (default), the C3 error button turns green and the C3 value is considered.

6.2.4.4. **Relative scaling (C-value approach)**

Relative scaling is useful in case the analyst has reasons to believe that the error of certain variables needs to be increased / decreased compared to others. Several causes could legitimate this approach. Two cases where this might be relevant are briefly discussed.

"various relative errors"	data from various instruments has different relative errors, e.g. one instrument 0.1% and the other 10% errors. PMF will fit the data from the first instrument more, as its Q contribution is much larger.
"different amount of var."	one instrument has 100 variables, the other 10. Variables from both instruments have a proper error estimation, e.g. 5% error. However, PMF will fit the data from the first instrument more, as its Q contribution is much larger. This becomes even more relevant, if co-variation between variables of the same instrument is high.

Rescaling the error of certain variables is achieved using the definition of classes in SoFi. The variables that should be rescaled must be part of one class (see 5.4.1). The following options are available:

"select"	The error matrix is untouched
"User-specific C-values"	The user defines C-values that should be tested on the adjacent table popping after having chosen this option (see Figure 42). The C-value is used to divide the current error, as shown in the following formula:

(18)

$$\sigma_{new} = \sigma_{old} \cdot \frac{1}{C_{value}}$$

where σ represents the error to be rescaled. Default C-value is one, i.e., the new error stays untouched.

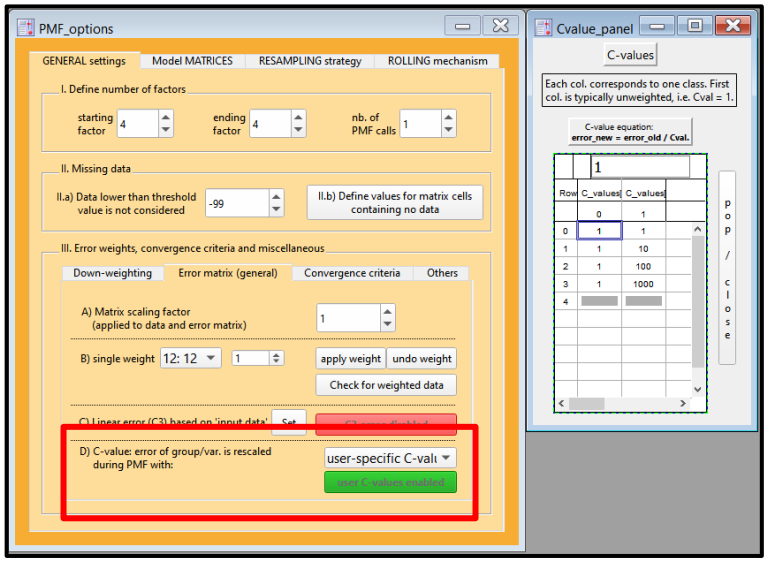


Figure 42 "user-specific C-values" option with the corresponding C-values in the table on the right.

A C-value can be specified for every defined class. If for example three classes have been defined, the adjacent table will pop three columns to be filled. The number of entries also defines the number of PMF runs. In the current example (see Figure 42) four runs are made, where the errors of the first class are left untouched, whereas the errors of the second class are systematically divided first by 1, 10, 100 and finally 1000, i.e., the second class receives 1, 10, 100 and 1000 times more weight during these runs, respectively.

See also **Fehler! Verweisquelle konnte nicht gefunden werden.** (example of combined criterion) for monitoring C-value runs using the criteria panel.

For further details on the manual technique of the C-value the user is referred to the paper discussing the weighing strategy as C-value between AMS and gas-phase data (Slowik et al., 2010; Crippa et al., 2013).

6.2.5. Convergence criteria

The parameters controlling the goodness of fit during the PMF run can be configured directly from the table on the first tab (Figure 43). Model optimization is performed in three sequential stages; each stage corresponds to a row in the table. Final convergence is achieved only if the parameters of the third stage (last row) are fulfilled. These parameters can be modified by the user, and this is necessary to achieve convergence for e.g. large datasets, strongly non-zero fpeak values, etc.

There are two convergence criteria. The first criterion inspects the stability of Q between steps of the iteration process. If the difference in Q between the last and the actual iterative step is smaller than “deltaQ”, e.g. <0.1, for the number of consecutive steps provided in the column “cons_steps”, the algorithm converges. The other convergence criterion is “gg2_steps”, where the norm of the squared Jacobian matrix (matrix containing all partial derivatives of Q with respect to the model variables) is compared to the threshold value provided, e.g. of 0.1. The “max_steps” column is simply the number of maximal iterative steps allowed at a given stage. If this maximal number of steps is reached during stages 1 or 2, the model moves on to the next stage and continues to search for a convergent solution. If this however occurs during the final stage, the model returns a non-convergent solution. Non-convergent solutions remain empty after having been imported and this is also reported in the log file. The user should easily recognize non-convergent solutions and can thereafter rerun the model using less restrictive convergence criteria, e.g. increasing by 10 or 100 the values of “deltaQ” and/or “gg2_steps” of the three stages (they should be increased/decreased together).

Note: Most often there is a problem in the error matrix or with the PMF settings and therefore convergence is not achieved. So before changing the convergence criteria, the user should double check the error matrix and the PMF settings.

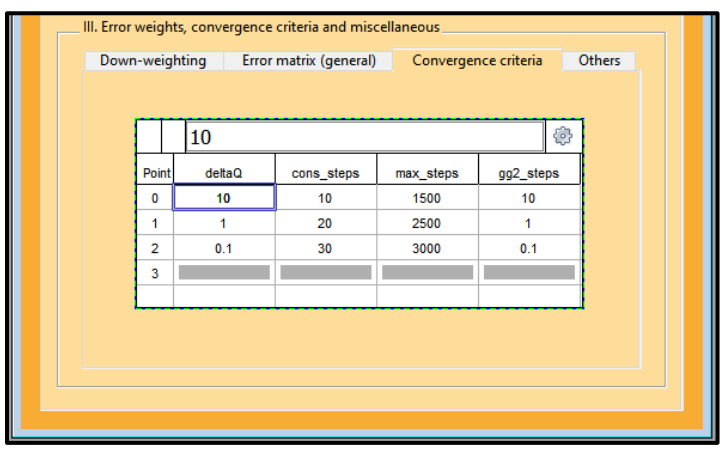


Figure 43 PMF settings, tab “General settings”, convergence critieria.

6.2.6. Pre-PMF evaluation

A quick SVD analysis can be performed before running PMF. SoFi will show the singular value vs the number of factor as well as the slope of the singular value change. This analysis can give some insights on the matrix rank of the PMF input, which could help to support the determination of the ideal number of PMF factors.

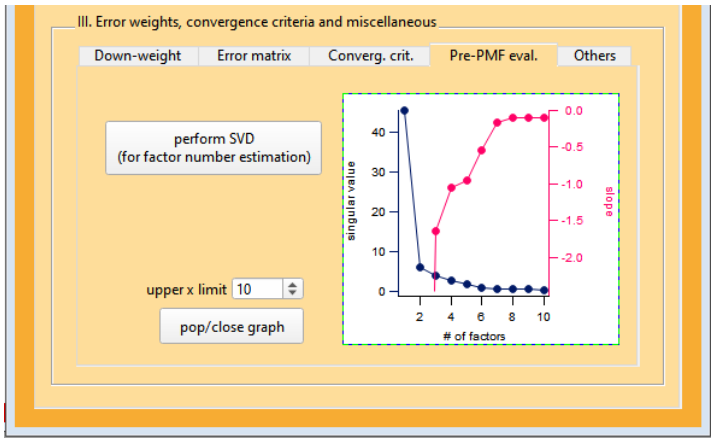


Figure 44 PMF settings, tab “General settings”, Pre-PMF evaluation.

6.2.7. Other various options

6.2.7.1. Robust model threshold (A)

PMF is run in the robust mode to safeguard the PMF run from drifting away due to outliers (see 2). The user can enable or disable the robust mode and define the threshold for the robust criterion in section A) (Figure 45).

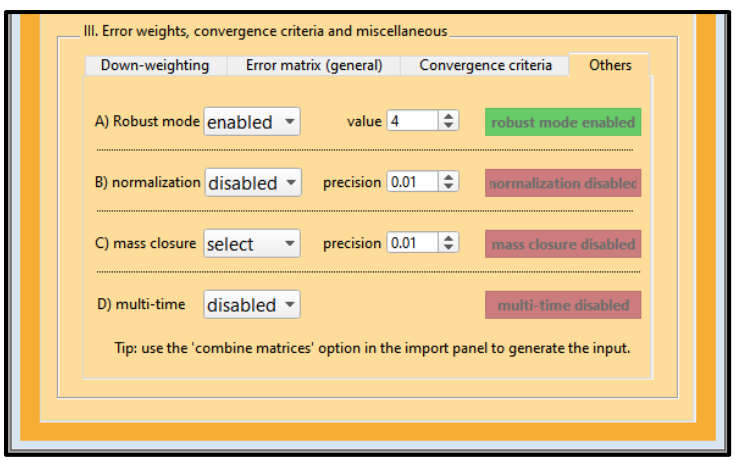


Figure 45 PMF settings, tab “General settings”, others.

6.2.7.2. Model normalization (B)

The user can select whether the sum of each factor profile is normalized to one during the iteration (enabled) or factor profiles are left unnormalized (default case, disabled).

6.2.7.1. Mass closure (C)

There is the option to fit the sum of all factor contributions to a certain mass, e.g. PM_1, $PM_{2.5}$, PM_{10} that can be chosen with the pop-up menu. The mass wave must be present under

"root:SoFi:External_data:Reference_ts:raw" (consult section 5.2.2). Selecting a possible time wave changes the red button "mass close disabled" to the green button "mass closure enabled". This adds mass balance equations to Q^{aux} and those are simultaneously fitted with all other equations defined by the regular PMF algorithm Q^m (Figure 45).

This option should only be used, when the mass difference is very small. The user is not supposed to regularly fit "invisible" mass. It is also recommended to run the model without fitting the missing mass and to compare the mass differences.

6.2.7.1. **Multi-time (D)**

The multi-time feature offers the possibility to run the PMF algorithm on a dataset containing two and more classes of data that differ in temporal resolution (Ogulei et al., 2005). The goal is to utilize the native resolution of each of the classes of instruments without being forced to average beforehand but rather consider and taking care of the different resolutions when performing PMF. The PMF equation that is solved instead of (1) is the following:

$$x_{sj} = \frac{1}{(t_{s2} - t_{s1} + 1)} \sum_{k=1}^{p} \left(f_{jk} \sum_{i=t_{s1}}^{i=t_{s2}} g_{ik} \right) + e_{sj} \tag{19}$$

where x_{sj} represents a single point of the model matrix **X** for the instrument sample *s* and the variable *j*, t_{s2} and t_{s1} are the start and end time of the sample *s*, respectively, f_{jk} is the mass fraction of variable *j* from factor *k*, g_{ik} is the mass contribution of factor *k* for a given time-point *i* for the variable *j* and e_{sj} is the non-modelled part of the residual matrix **E** for sample *s* and variable *j*.

The sample with the highest temporal resolution represents the main equations in **X**. For the main equations t_{s2} and t_{s1} are identical and the inner sum disappears, as well as the normalization fraction at the beginning. For the other samples, the point x_{sj} is represented as the average over all points from the main equations present during the sampling period, i.e., between t_{s2} and t_{s1}, as stated in equation (19).

In case the measurement for the classes with lower time resolution contains some gaps, then it is extremely important to pass the information of the start and end time for these classes. In order to pass this information, the user MUST pass the PMF input time series as a 2D matrix, where the first column contains the start time and the second column the end time. Note that the start time for the main/default class is not used, hence for this class the time series input wave can (in case it is easier for the user) contain two times the same information.

In order to perform multi-time, the user is supposed to prepare and read the input and error matrices in SoFi as described with the following schematic:

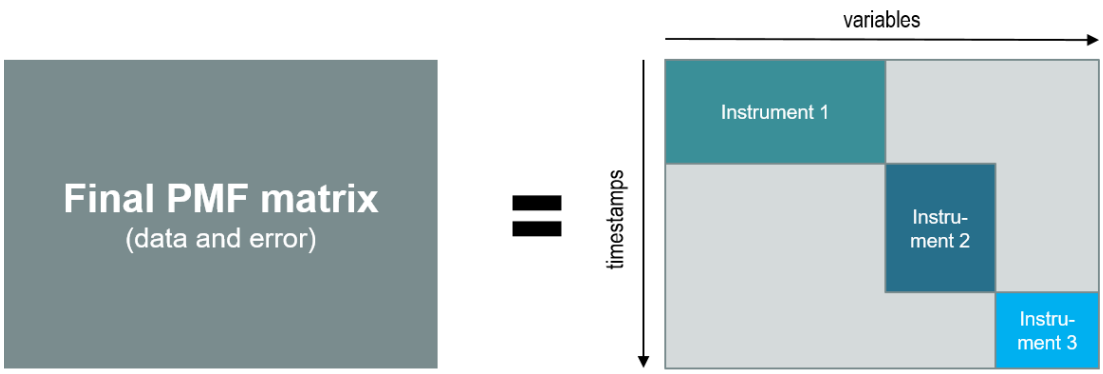


Figure 46 Input data and error matrices for multi-time

where sample 1 has the highest time-resolution and sample 2, 3, etc. have always lower time-resolution. The empty/white parts in the final matrix (right Figure) are all "nans". The time series of all samples are passed in their native time-resolutions. This means that the time series wave will move back and forth, as the samples 2, 3, etc. are supposed to cover the same time-range as sample 1. SoFi will then find the proper index values and pass them to ME-2 for the additional averaging for the equations x_{sj} for the samples 2, 3, 4, etc. given in the example in Figure 46. The points of the higher classes are expected to be within the main time series, i.e., that of the first class. If that is not the case, SoFi will automatically not consider these averaged points (gap-safe). Moreover, the multi-time code is also safe when used in combination with the bootstrap resampling technique (see 6.4.1).

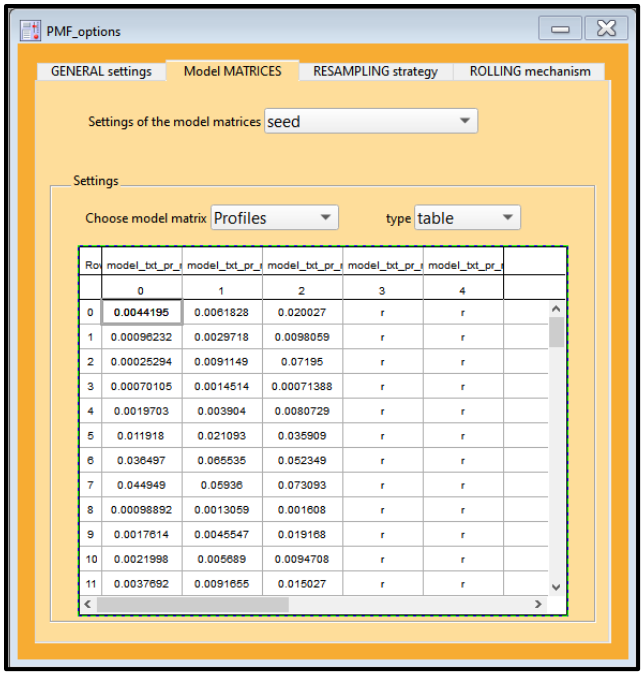
To prepare the diagonal matrix, the user can use the "combine matrices" option in the import panel (4.1.4). <u>The user is supposed to define the non-diagonal elements in the data and error matrix as missing values (please consult 6.2.2). In addition, the user must define a class for each instrument sample over time and over the variables (consult 5.4.1), the multi-time equations are not properly passed otherwise.</u>

The resulting adjustment factors η_j described before are reported as wave "adjustment factor" in the corresponding result folder: "root:SoFi:Results:HDF filename:Solutions:run_X", for further manual inspection. Note that adjustment factors close to unity indicate a good agreement between the different measurements.

6.3. Define options for the model matrices

6.3.1. Seed

The seed option populates the model matrices **G** and **F**. Random initialization typically explores the amount of local minima, where the final Q-value is still high, compared to the global minimum or regions in the Q-space with almost identical minimal Q-value (rotational ambiguity, see 2.3). Increasing the number of PMF runs with completely random values for **G** or **F** augments the changes to find these regions.



Row	0	1	2	3	4
0	0.0044195	0.0061828	0.020027	r	r
1	0.00096232	0.0029718	0.0098059	r	r
2	0.00025294	0.0091149	0.07195	r	r
3	0.00070105	0.0014514	0.00071388	r	r
4	0.0019703	0.003904	0.0080729	r	r
5	0.011918	0.021093	0.035909	r	r
6	0.036497	0.065535	0.052349	r	r
7	0.044949	0.05936	0.073093	r	r
8	0.00098892	0.0013059	0.001608	r	r
9	0.0017614	0.0045547	0.019168	r	r
10	0.0021998	0.005689	0.0094708	r	r
11	0.0037692	0.0091655	0.015027	r	r

Figure 47 PMF settings, tab “Model MATRICES”, seed.

The table in the tab “Model MATRICES” for the selection “seed” shows the overview of the chosen model entries for the matrix **F** “Profiles” or **G** “time series” (Figure 47). The default value is “r” for random. The user governs the various possibilities with the subpanel that appears after pressing on “Define model entries”.

6.3.1.1. Settings

“Factor” For which factor position a seed value should be set - all factors or a single factor position, e.g. factor 1, can be chosen

“Initial model values” Option on how to set the seed value, can be one of the following options:

- <u>random (r):</u> always completely random entries (default)

- pseudo random (pr): randomly initialized at the beginning of a PMF call, and the same random value is used through all other runs belonging to the same PMF call
- zero(0): sets the seed value for the entire factor/selected variable to zero (works without external reference data)
- from vector (fv): solution matrix **F** / **G** is populated using external information from the folder "root:SoFi:Reference_ts:raw or root:SoFi: Reference_pr". Consult 5.2 for preparing external data directly in SoFi.

"Variables" Whether the seed value for all variables (all entries), all variables of external vector or single variables should be set

"Normalize factor" vector used for setting seed value can be normalized (set sum factor below, by default 1) or not.

"Write into model matrix" Update the model matrix, table in the main PMF settings window will update accordingly.

"reset (…) matrix" reset (to random) either entire matrix or the current factor position

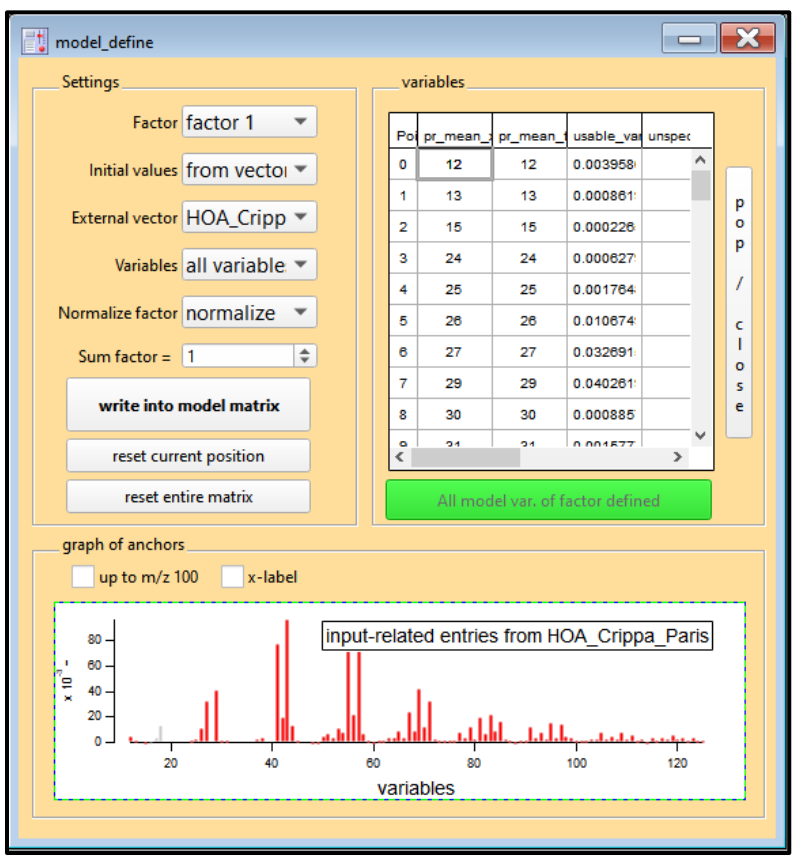


Figure 48 Panel for initial model values from vector. HOA_Paris profile normalized to 1 is going to be written on factor position 1 for the profile matrix F.

6.3.1.2. **Options for "from vector (fv)"**

When choosing "from vector (fv)" the panel "model_define" (Figure 48) 4 additional options are available. For each external vector one can either choose all variables of the vector or a specific variable to populate the model matrix. The chosen variable(s) of is (are) shown in the left table "usable variables from external vector". The entries of current model column which are still undefined are displayed in right table "unspecified model variables". The tables get updated each time some information is passed to the model matrix by clicking the button (write into model matrix". Writing information to the model matrix is additive. If the vector of the matrix **F** / **G** contains some undefined entries, the button "Some var. of factor still unspecified" remains yellow. As soon as all entries are defined, this button turns to green "All model var. of factor defined".

The user has the option to normalize the entry to the value provided under "Sum factor var.". This option is linked with the information that is written and not to the final vector present in the model matrix.

Example

The user wants to normalize the sum of a profile vector (matrix **F)** to one and the information of this vector is stored in two independent external vectors, **A** and **B**. Then the following calculation should be commuted:

$$w \cdot (sum(\mathbf{A}) + sum(\mathbf{B})) = 1 \tag{20}$$

w representing the weight value.

$$nor.value\,(\mathbf{A}) = w \cdot sum(\mathbf{A}) \tag{21}$$

$$nor.value\,(\mathbf{B}) = w \cdot sum(\mathbf{B}) \tag{22}$$

with $nor.value\,(\mathbf{A})$ and $nor.value\,(\mathbf{B})$ being the values entered under the voice "Normalize factor var." on the panel "model_define" (Figure 48).

Writing the information of one or more external vectors to the model matrix **G** or **F** from the "seed" option does not mean that this information is constrained. This information simply represents the starting condition for the PMF run(s). Constraining this information or data that is derived upon this information is performed using the other two techniques in: "a-value" or "pulling equations", see 6.3.3 and 6.3.4 for more details.

6.3.2. fpeak

The following fpeak options are available:

"disabled"	fpeak rotations are disabled.
"global fpeak"	enables rotations that affect the entire model matrices.
"individual fpeak"	enables rotations that affect only chosen contributions, profiles.

For these fpeak options the following types of analyses are supported:

"exact fpeak"	exact fpeak value is used.
"sensitivity fpeak"	fpeak values are systematically tested between an initial and final fpeak using a user-defined fpeak resolution.
"random fpeak"	fpeak values are randomly tested between an initial and final fpeak with a user-defined fpeak resolution.

6.3.2.1. **Global fpeak**

The global and the individual fpeak are those described earlier under 2.4. The user must enter the initial and final fpeak and the fpeak step interval. The range of fpeak values should be selected in a way that allows to fully investigate the range of solutions that are environmentally interpretable and at the same time are of mathematically similar quality (i.e., similar Q value) (Figure 49).

<u>Example</u>

One might explore the range of fpeak that leads to a 10% increase of the normalized scaled residuals (Q/Q_{exp}). A number of past studies have simply investigated a predefined fpeak range, e.g. fpeak values between -1 and 1, regardless of its effect on Q. This approach should be avoided, since the fpeak rotation is a function of the model and therefore, Q/Q_{exp} will vary based on the input data and the model result.

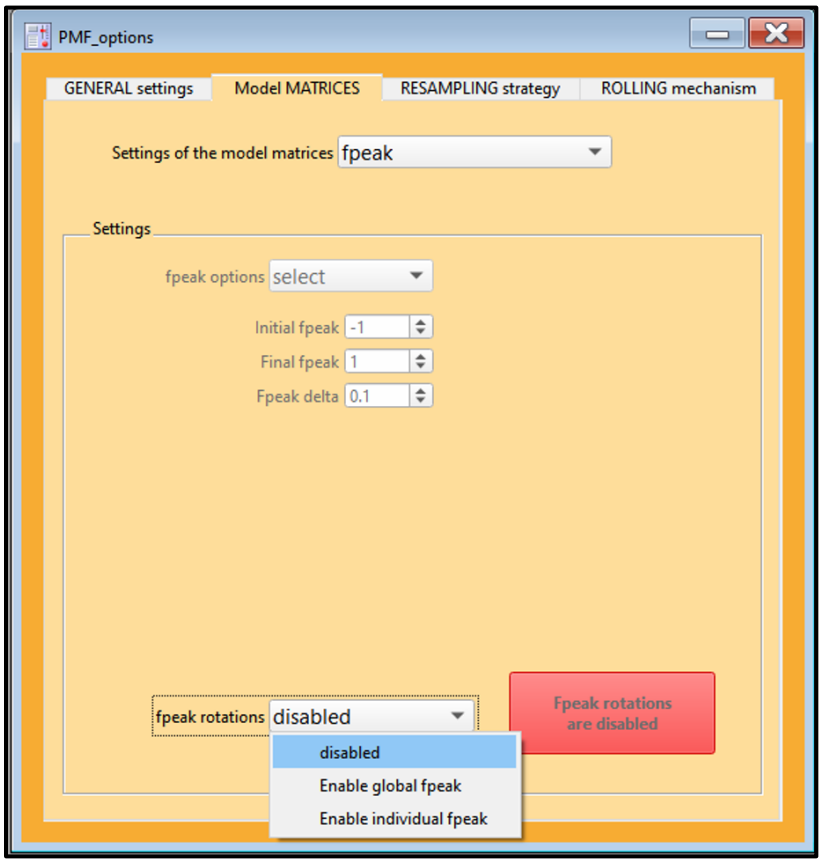


Figure 49 PMF settings, tab “Model MATRICES”, fpeak.

Selecting one of the two options, “global fpeak” or “individual fpeak”, enables the fpeak rotations for the PMF run(s).

6.3.2.2. **Individual fpeak**

The user defines the initial, final fpeak and the fpeak step interval as for the global fpeak. In the cells of the matrix (Figure 50) passing “NaN” tells SoFi to pass the current fpeak during the PMF run(s). Conceptually, it can be regarded as the scalar with which time series of one factor is added or subtracted (based on the sign of fpeak) from another factor time series. The notation follows the rules of linear algebra.

Example

If for example the entry **T** [0,1] = -5, this implies that the time series of factor 1 is subtracted by magnitude -5 from the time series 2. Vice versa the factor profile of factor 2 is added to the factor profile of factor 1 by the same magnitude with opposite sign, 5.

However, a zero entry in **T** does not prevent addition/subtraction from occurring, but rather indicates that this rotation, should it occur, is neither favored nor penalized by the model.

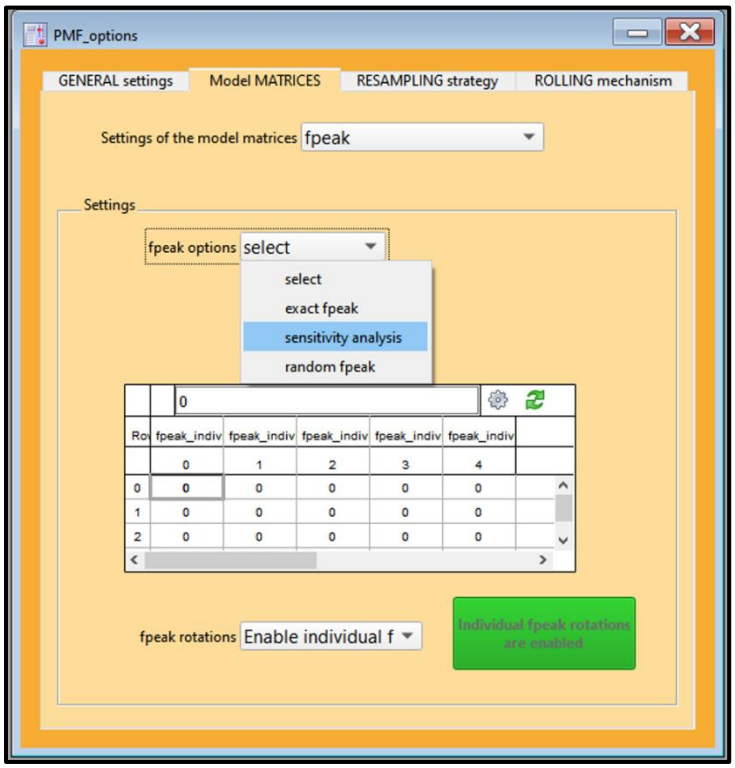


Figure 50 PMF settings, tab “Model MATRICES”, individual fpeak enabled with the example for T[0,1] as “NaN” for the fpeak value.

6.3.3. *a*-value constraints

a-values can only be selected AFTER the seed values have been defined.

The *a*-value option allows constraining one and more entries of the profile matrix **F** or the time series matrix **G** (see 2.4.4). The constrained information can be monitored on the table / graphs of the panel “PMF_options” (Figure 51). Adding or changing the constraints is done using the adjacent subpanel “constraint define” (Figure 52).

To enable the *a*-value constraints, use the dropdown menu below the table. The *a*-values are only used by the ME-2 solver when the button “A-value constraints are enabled” turns green (Figure 50) and seed values are set first.

a-value option combined with displacement for the seed

In case the a-value strategy fails in providing sufficient variation within the user-defined a-value range, for example the case when always the same result value is reached, independent of the a-value range, then, the user has the possibility to displace the seed value in combination with the a-value settings. The difference compared to classical a value approach is that within the displacement option, the seed value is perturbed based on the settings of the a-value, but then the current/new seed value is locked during the PMF iteration. This allows to explore values slightly above or below given anchor values that wouldn’t be accessible using classical a-value

exploration otherwise. Anchor values can be perturbed systematically, for example using sensitivity analysis (supported up to 5 independent dimensions) or using random exploration (as typically the case in EPA PMF). The reader is referred to Paatero et al. (2014) for more details about this technique.

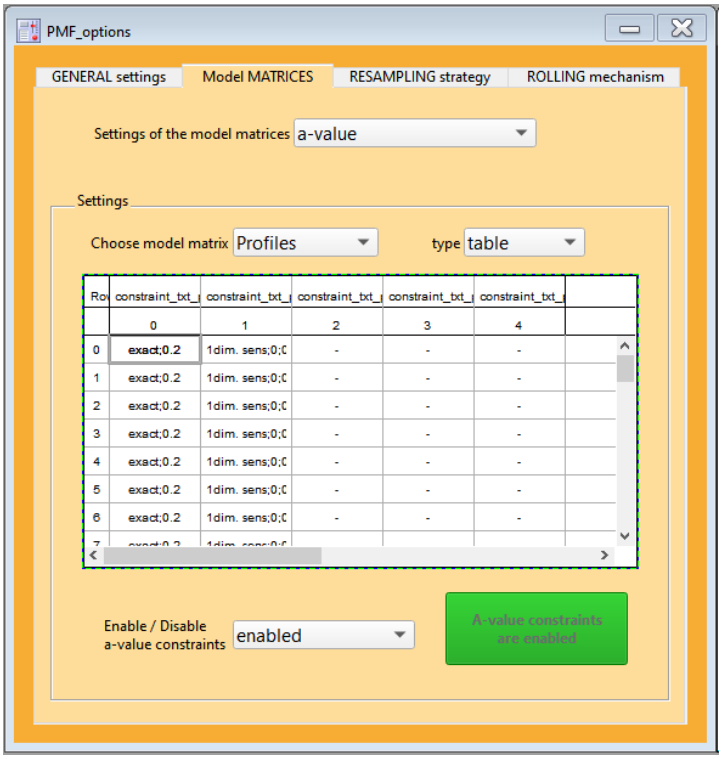


Figure 51 PMF settings, tab "Model MATRICES", *a*-value.

6.3.3.1. **Main settings**

"Factor"	all factors or a single factor position, e.g. factor 1
"Type of constraint"	"exact *a*-value" "1st dim. sensitivity" "2nd - 5th dim. sensitivity" "random" "pass limits directly to ME-2"
"Variables"	all model variables, all variables of external variables, all entries, single variables, e.g. *m/z* 44 The user has the option to choose which base of variables is used to pass the information of the constraint. <u>Constraining variables is additive, as for defining the variables</u> (see 6.3.1).

Some more details on the options are the following.

6.3.3.2. **Exact a-value (exact)**

An exact *α*-value, e.g. 0.1 is used to constrain the defined information (profile / time series)

6.3.3.1. **1st dim. sensitivity (1dim. sens)**

The *α*-value is systematically changed from "initial a-value", using the step interval given by "*α*-delta" up to "final *α*-value". This allows studying the PMF outcome as a function of the *α*-value in one dimension.

If you erroneously define two different settings for sensitivity analysis in one dimension, SoFi will only use the settings that are encountered first.

Example

First factor has a final *α*-value of 1 and second factor a final *α*-value of 0.5 and both factors have the constraint: one dimensional sensitivity analysis. For the one-dimensional sensitivity analysis the final *α*-value passed to ME-2 will be 1, since the first factor is internally evaluated before the second one.

6.3.3.2. **2nd dim. sensitivity (2dim. sens) to 5th dim. sensitivity (5dim. sens)**

This option should be used if the *α*-value combinations should be tested individually. The *α*-value is systematically changed from "initial a-value", using the step interval given by "*α*-delta" up to "final *α*-value". This allows studying the PMF outcome as a function of the *α*-value in a 2nd up to 5th dimension.

If only a sensitivity analysis in one dimension should be performed, the first dimension must be chosen (see above). Using higher dimensions and keeping previous dimensions empty or non-defined generates an error message in SoFi.

Example

The first factor has a final *α*-value of 1 and a *α*-delta of 0.2 and performs a one-dimensional sensitivity analysis. The second factor has a final *α*-value of 0.5 and a *α*-delta of 0.1 and is linked with the two-dimensional sensitivity analysis. All possible *α*-value combination will be run, SoFi will perform

$$\left(\frac{1}{0.2}+1\right)\cdot\left(\frac{0.5}{0.1}+1\right)=36\,runs \tag{23}$$

6.3.3.3. **Random a-value (random)**

Random *α*-values are passed from a uniform distribution between zero and "Final *α*-value" with a precision given by "*α*-delta".

6.3.3.4. **Pass limits directly to ME-2 (limits)**

If for a specific source, e.g. a traffic source the range of the profile is known, this range (limits) could be directly passed to ME-2 using this option. Two vectors representing the lower and the higher limits of the profile / time series are provided and ME-2 optimizes the profile during the PMF run within these limits. Note that a seed vector must be defined too. Use the lower or higher vector as starting condition when defining the corresponding seed information.

6.3.3.5. **All model variables**

This option uses as base the variables defined in the model matrix **F** / **G**. This leads to the same *α*-value constraint throughout all variables.

6.3.3.6. **All model variables of external vector**

This option uses as base the variables of an external vector. This is useful, when various *α*-values should be used within a constrained vector (coming from various external profiles), e.g. if some variables of the constrained vector are more accurate than others or the profile that should be constrained does not contain all ions that are present in the PMF input and the missing ions should be constrained with a loose *α*-value too. In such a case the less accurate ones could receive a higher *α*-value.

Similar to the “model_define” panel (Figure 48) the “constraint_define” shows the variables that can be constrained (left table “variables that can be constrained”) as well as the ones that are still unconstrained (right table “still unconstrained variables). The latter table gets updated each time some information is passed to the constraint matrix. If the vector of the matrix **F** / **G** contains some unconstrained entries, the button “Some var. of factor still unconstrained” is yellow. As soon as all entries are constrained, this buttons turns to green “All model var. of factor constrained”.

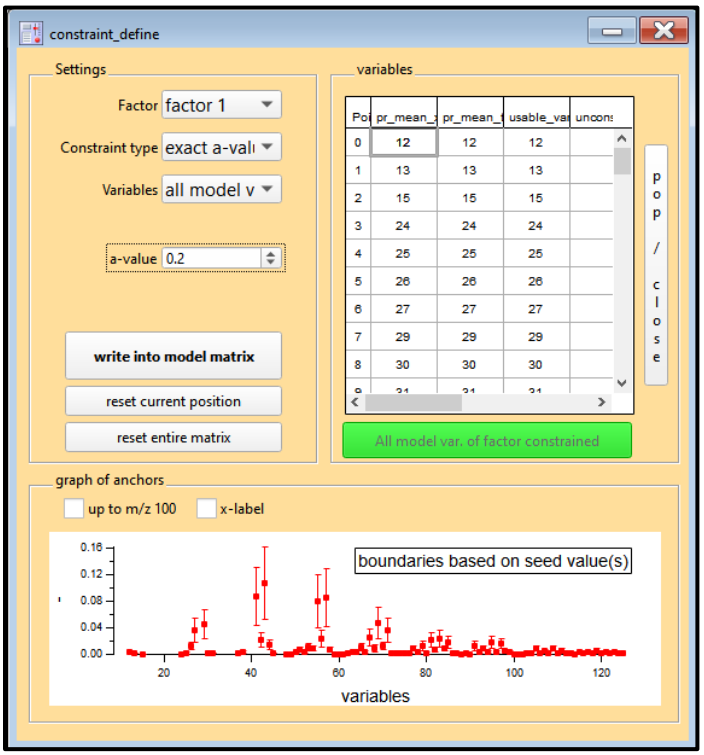


Figure 52 The HOA_Paris profile defined on factor 1 of the matrix F earlier (see Figure 48) is going to be constrained with a fixed *a*-value of 0.2.

The button “write into model matrix” writes the chosen settings and the user can monitor the changes on the main panel (Figure 51). If an entry was erroneously defined, the button “remove last settings” can be used to remove the last writing. The button “reset entire matrix” resets the current constraint matrix to default, i.e., empty entries (“-“) for the entire matrix (Figure 52). Every time the user wants to ”write into model matrix”, SoFi checks, whether undefined information is going to be constrained and constraints only existing information, i.e. seed values need to be set beforehand.

It is recommended to constrain an entire profile / time series. If a wave is only partially constrained, its unconstrained part could grow during the PMF run and become dominant upon normalization in SoFi.

6.3.4. Displacement

The displacement approach in SoFi allows the user to introduce random and/or systematic perturbations of selected anchor elements across up to five independent dimensions. Initially, the anchor values are modified according to the chosen a-value constraints. Once the displacement is applied, however, these anchors are kept fixed (“frozen”), ensuring that the model explores variations around a defined configuration rather than continuously re-adjusting the constraints. Displacement helps probe how sensitive the solution is to small, controlled shifts in key anchor positions. Because anchors are frozen after displacement, there is a risk of over-

constraining the model, additionally, only a limited portion of the rotational space is explored, and relevant alternative solutions may still be missed if the chosen displacements are not representative.

6.3.5. Pulling equations

Pulling is another method used in ME-2 to introduce *a priori* information into the factorization. In addition to fitting the measured data, the model can be guided towards predefined values or relationships between variables in the profile matrix **F** or the time series matrix **G** (see 2.4.5).

This is achieved by adding additional equations that penalize deviations from the specified targets during the optimization. Pulling therefore allows the user to incorporate external knowledge while still allowing the model to adjust the solution according to the measured data and is therefore regarded as a soft constraining technique.

Pulling equations can be written manually or with the help of SoFi. The ME-2 solver is very sensitive to the form of the pulling equations and its limits, so the equations should be defined carefully. To enable the pulling equations, use the dropdown menu below the table. **The pulling equations are only used by the ME-2 solver when the button "Pulling equations are enabled" turns green** (Figure 53)**.**

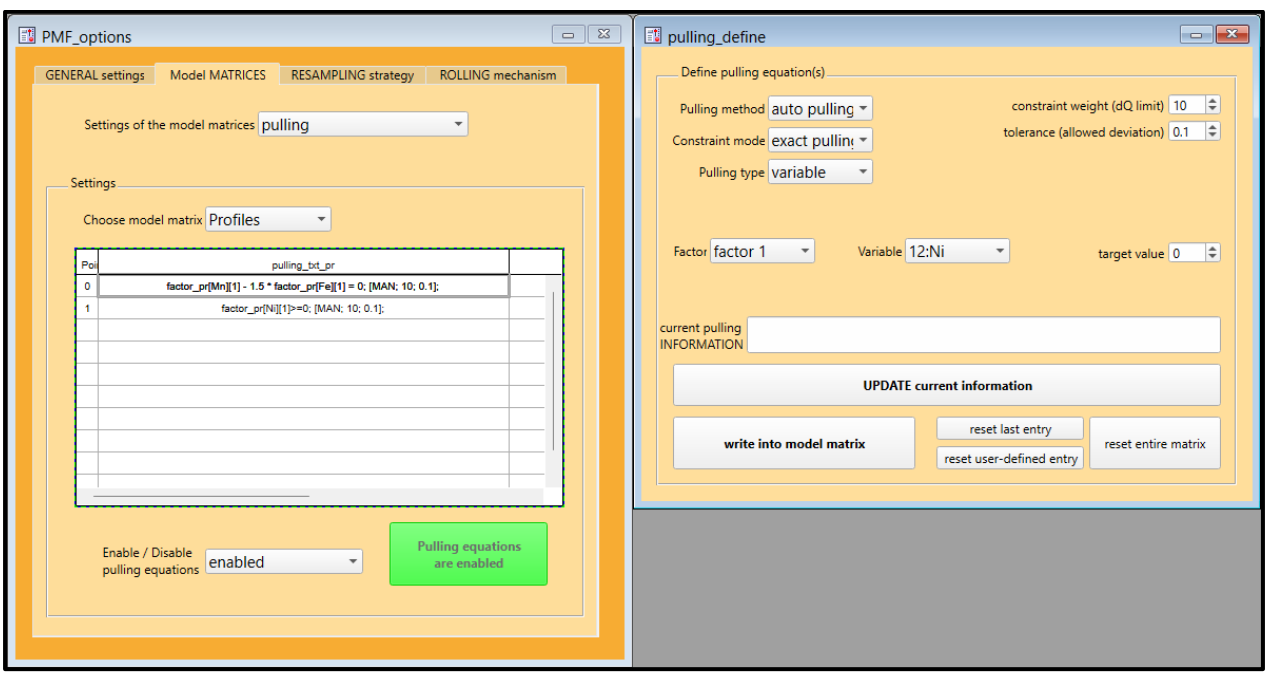


Figure 53 PMF settings, tab "model MATRICES", pulling equations. The table shows the defined pulling equation, the newly popped panel next to it should be used to define the pulling equations.

SoFi offers a wide range of options for different pulling equations. They are described in the following. Not all combinations of mathematical options are compatible with each other. Additionally, some information has to be added manually.

6.3.5.1. **Pulling method**

SoFi (and the ME-2 solver) distinguish between two types of pulling: **automatic pulling** and **manual pulling**. In automatic pulling, the user specifies a target value together with a constraint strength and a tolerance. The strength determines how strongly deviations from the target influence the optimization, while the tolerance defines the acceptable range around the target value. This approach provides a flexible way to guide the solution without strictly enforcing the constraint.

In manual pulling, the constraint is treated as an additional equation with a defined uncertainty. The user specifies the target value and an expected deviation, and the model minimizes the difference between the calculated value and the target during the optimization. Manual pulling is therefore typically used when reliable prior information is available and a more explicit constraint is desired.

6.3.5.2. **Pulling type and following pulling equation**

The first part of the pulling information is dedicated to the actual pulling equation and the target value. Depending on the type of what should be pulled, the pulling equation can look slightly different, but the main format is:

$$\text{model variables to be pulled} = [\text{pulling value}];[\text{pulling strength}];[\text{pulling uncertainty}] \quad (24)$$

For more user-friendliness, SoFi offers templates for pulling single variables, sums, differences and ratios. Especially for the ratios, the user should make use of them as the equation syntax of the ME-2 solver is quite rigid, and the ratio has to be written as a difference equal to zero, division in the pulling equation is not allowed/recognized, the ME-2 solver only accepts a sum or difference of terms, where terms are either single terms or products (Figure 54).

For manual pulling, only equal equations are allowed, for auto pulling, the solver also accepts unequal pulling equations (>= and <=), as the ME-2 solver has access to different so-called error models that allow one-sided pulling, where as for manual pulling only symmetrical pulling is allowed.

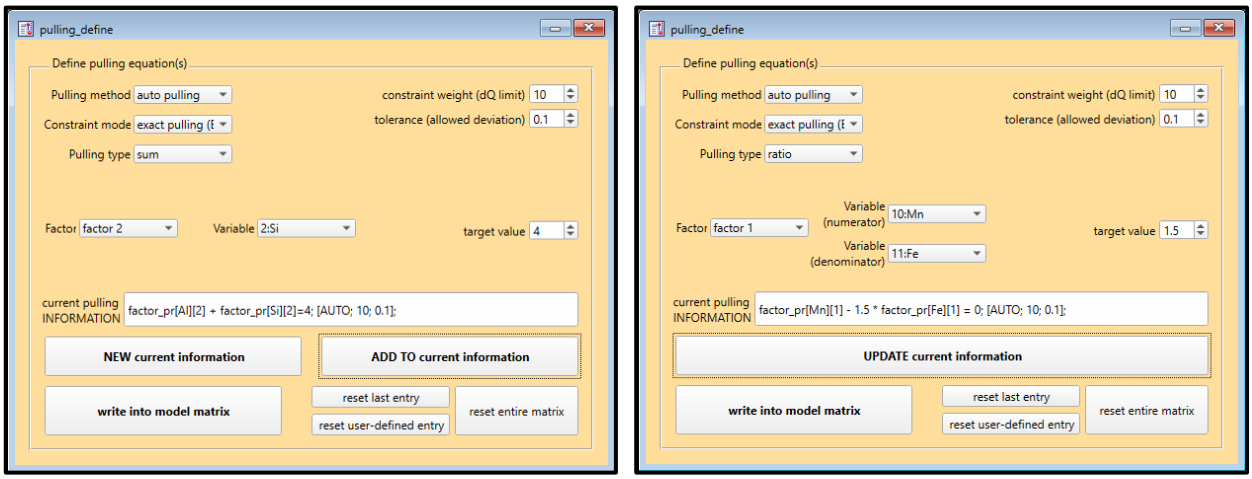


Figure 54 For sum or difference of terms, the “New current information” starts a new equation, other terms can be added via the “add to current information” (left). Ratios have to be written as a difference of two terms that is equal to zero, where one term includes the multiplication with the target value.

6.3.5.3. **Pulling strength**

The influence of a pulling constraint on the solution is controlled by its associated parameters, so the second part of the pulling information is dedicated to these pulling details, i.e. the pulling method (AUTO or MAN), the weight/strength and the tolerance/uncertainty.

The constraint weight/strength determines how strongly the model attempts to satisfy the specified constraint. Increasing this value increases the contribution of the constraint to the objective function (Q), thereby forcing the solution closer to the target value. Low values result in weak constraints with little influence on the solution, whereas very high values may dominate the optimization and reduce the influence of the measured data.

For automatic pulling, the tolerance defines the acceptable deviation from the target value. Smaller tolerances enforce the constraint more tightly, while larger tolerances allow greater flexibility. The tolerance should typically be chosen in accordance with the expected variability or uncertainty of the target value. For manual pulling, the uncertainty plays a similar role. It represents the expected error of the constraint and determines how strongly deviations are penalized. Smaller values correspond to stricter constraints, while larger values reduce the influence of the constraint.

6.3.5.4. **Constraint mode**

Similar to the a-value approach, SoFi offers exact, random and automatic sensitivity analysis of several dimensions. This can be applied to the target value and/or the constraint weight/strength and or the constraint tolerance/uncertainty. For any sensitivity analysis or random pulling value, the notation is the following:

$$[\text{“type”}, \text{“initial value”}, \text{“final value”}, \text{“resolution”}] \tag{25}$$

These three parameters, i.e, target, weight and tolerance are treated as independent parameters. Therefore, the user can perform independent sensitivity analysis for these three parameters, if desired. This results in three independent sensitivity analysis dimensions, i.e., one is 1D, the second is 2D and the third one is 3D.

6.3.5.4.1. *Exact pulling (EXT)*

An exact target value is used to pull the current information (profile / time series) during the iteration towards this anchor value. No square brackets need to be defined, for easy readability the exact value is just passed to the pulling information.

6.3.5.4.2. *1st dimension sensitivity analysis (1D)*

The user-defined information (target or weight or tolerance) is systematically changed from an initial value, using user-defined step intervals to a final value. This allows studying the PMF outcome as a function of the pulling information in one dimension.

For the target value, the user can make use of menus for the minimum and maximum value as well as the step size. For the weight/strength and the tolerance/uncertainty, the user has the option to add a template into the pulling information and has to change the information there manually (Figure 55).

<u>If you erroneously define two different settings for sensitivity analysis in one dimension, SoFi will only use the settings that are encountered first</u>.

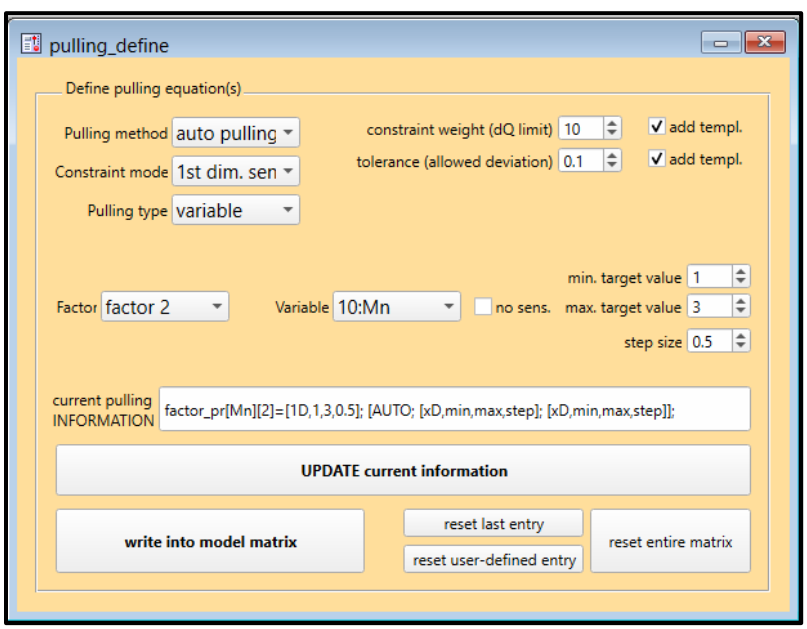


Figure 55 Define pulling equations. Sensitivity analysis for the target value can be chosen through entering the desired values for minimum, maximum and step size. For the weight and the tolerance a template will be inserted and the user has to manually enter the values in the "current pulling information" field.

6.3.5.4.3. *2nd dim. sensitivity (2D) to 5th dim. sensitivity (5D)*

The user-defined information (target or weight or tolerance) is systematically changed from an initial value, using user-defined step intervals to a final value. This allows studying the PMF outcome as a function of a pulling information in a 2nd up to 5th dimension.

If only a sensitivity analysis in one dimension should be performed, the first dimension must be chosen. Using higher dimensions and keeping previous dimensions empty or non-defined generates an error message in SoFi.

Example

The first equation has a min. target value of 0 and a final target value of 1 and a step size of 0.2 and performs a one-dimensional sensitivity analysis. The second equation has a minimum and maximum weight of 10 and 50, respectively, and step size of 10 and is linked with the two-dimensional sensitivity analysis. SoFi will perform

$$\left(\frac{1}{0.2}+1\right)\cdot\left(\frac{40}{10}+1\right)=30\,runs \tag{26}$$

6.3.5.4.4. *Random pulling (RND)*

Random pulling values are passed from a uniform distribution between initial value and final value with resolution of delta-value.

6.3.5.5. **Write pulling information**

Once all the pulling parameters are defined, the pulling information needs to be written into the model matrix, via the button “write into the model matrix”. The user also has the option to remove either the last or a specific pulling information from said matrix as well as to reset the entire matrix.

Don’t forget to enable the pulling equations before running PMF.

6.3.1. Factor specifications

Each factor receives the default name “factor [position]”. The user can change these names under “factor specs” (Figure 56) and give proper source names instead. The factor names in the analyze PMF plots are based on these names and it is also important for averaging PMF solutions with a different number of factors.

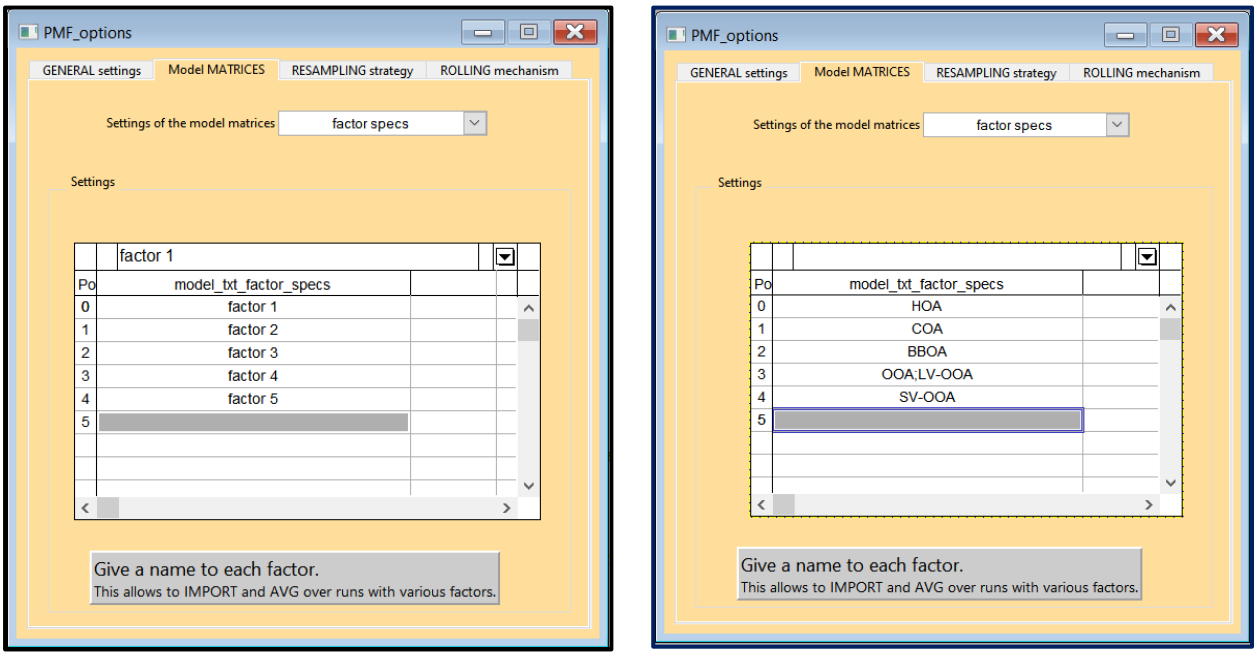


Figure 56 PMF settings, tab “Model MATRICES”, factor specs (left) default names, (right) user-defined names.

When averaging PMF results in SoFi (see 7.5.1), factor possessing the same name will be averaged together. This allows the user to import various HDF files with various sources. The averaged result will then be based on all factor names defined by the user. As a consequence, the final factor specs list of the averaged solution might be larger than the factor specs list of a single HDF file.

If an HDF file contains multi-factor solutions and the various factor solutions at the same position shouldn’t necessary be averaged together, then the user has the possibility to pass various names for a factor position. Factor names should not include “+” in the name, otherwise SoFi will later falsely mistake it as two factors that should be added.

Example

Averaging three different HDF file together, where only HOA and COA is a common factor between all of them.

Table 4 factor specifications of three HDF files and what that means for the averaged solution.

HDF 1	HDF2	HDF3	**avg. solution in SoFi**
HOA	HOA	HOA	HOA form all files
COA	COA	COA	COA from all files

BBOA	SV-OOA	BBOA	BBOA from HDF1 and 2
OOA	LV-OOA	SV-OOA	OOA from HDF 1
		LV-OOA	SV-OOA from HDF 2 and 3
			LV-OOA from HDF2 and 3

Example

An HDF file with 4 and 5 factor PMF runs, where the 4th position in 4-factor runs represents something different then the 4th position in 5-factor runs. When averaging over 4- and 5-factor solutions, SoFi will use dedicated factor positions for OOA and LV-OOA. This leads to a six-factor solution for the averaged solution, since OOA and SV-OOA are separated from each other.

Table 5 factor specifications for one HDF file containing 4 and 5 factor runs and the averaged solution.

HDF	**avg. solution in SoFi**
HOA	HOA
COA	COA
BBOA	BBOA
OOA;LV-OOA	OOA
SV-OOA	SV-OOA
	LV-OOA

If this was not done prior to running PMF, so the information on the user-defined factor names is not stored in the HDF file, the user can also manually change the factor names before averaging the solutions. The factor specs list can be found here:

root:SoFi:Results:"NAME_OF_HDF_FILE":Parameter:model_txt_factor_specs

6.3.2. Consideration of the priority of the rotational tools

The explorative techniques for the rotational ambiguity are additive. The fpeak rotations can be enabled besides other techniques, such as the a-value or pulling equations or C-value approach, etc. The fpeak rotations have least priority, i.e., rotations are not allowed, if the resulting values are for example outside the limits given by the a-value constraints.

6.4. Estimation of the statistical uncertainty

The final PMF result / solution **must** also contain a quantitative measure of the statistical uncertainty. For ambient aerosol data the highest contributors to the statistical uncertainty are mainly random changes of the emission sources. Past studies ((Paatero et al., 2014) and (Ulbrich et al., 2009)) typically employed the resampling strategy "bootstrap" (Davison and Hinkley, 1997) to address and quantify this type of error for AMS data.

6.4.1. Bootstrap resampling strategy

he bootstrap method generates a new set of input matrices for analysis by randomly resampling from the original dataset (Efron, 1979). This resampling strategy involves duplicating some data points while omitting others, which effectively perturbs the input data (see Figure 57). Given that a sufficient number of resamples has been performed, the variation within the identified factors across all bootstrapped runs allows to estimate the statistical uncertainty.

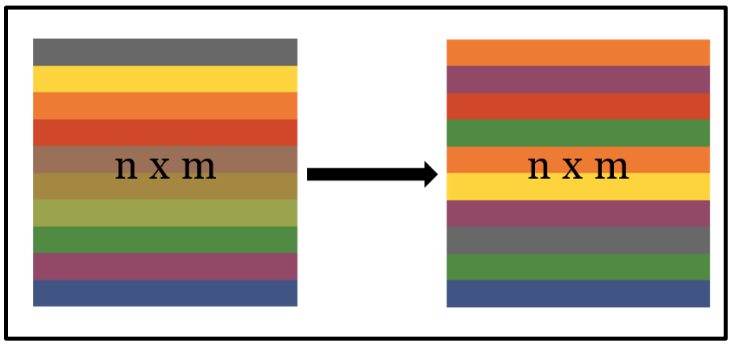


Figure 57 Schematic of the default PMF input matrix (left) and a resampled one (right side). The colors represent the resampled rows. Note that the dark green color appears twice, whereas the olive-green present in the default PMF matrix is absent.

Main Settings

"unblocked"	individual time points are randomly sampled
"using defined classes"	all points belonging to the same class are grouped (blocked) together. Check 5.4.1 for how to create classes. Classes can be created on external text waves, indices or keywords.

In the **unblocked bootstrap**, individual data points are randomly sampled with replacement from the original dataset to create new, resampled datasets. Each new dataset contains the same number of data points as the original, but certain data points may appear multiple times, while others may not appear at all. A crucial point when applying the bootstrap strategy on the PMF model is the presence of sufficient variability for retrieving the expected sources. If for example, the data covers a special period such as a fireworks event, it might happen that after resampling, the points reflecting this specific period are eventually missing and little variability is available to

separate the specific source. The **blocked bootstrap** addresses the limitations of the standard bootstrap for time-dependent dependent data. Instead of resampling individual data points, blocks of data are resampled together. These blocks can be defined by the user (5.4.1). These blocks preserve the natural dependencies within the data. The choice of block size should align with the natural variability in the data. For example, in traffic data, daily blocks or blocks of rush hour vs non-rush hour might capture diurnal patterns, while smaller blocks would be inappropriate.

Bootstrap repeats play a crucial role in ensuring that the uncertainty estimates are robust and reliable. A higher number of repeats typically improves the accuracy of the uncertainty assessment for the identified factors. This is because each repeat introduces a different sample of the data, helping to capture a broader range of variability that might exist in the dataset. Insufficient bootstrap repeats can lead to misleading uncertainty estimates, as the resampled datasets may not capture all the relevant patterns or variabilities, particularly in complex or noisy datasets. The number of bootstrap repeats is controlled by the number of PMF calls under "GENERAL settings". For standard PMF analysis (static PMF input), it is generally recommended to perform at least 500 - 1000 bootstrap repeats, but in the end choosing the appropriate number of repeats is often a balance between computational feasibility and the need for accurate uncertainty estimation. In rolling PMF, the number of bootstrap repeats per window can be lower than in standard PMF, due to the smaller size of the data window that is rolled over the dataset which also lead to additional repeats for each day (Chen et al., 2022).

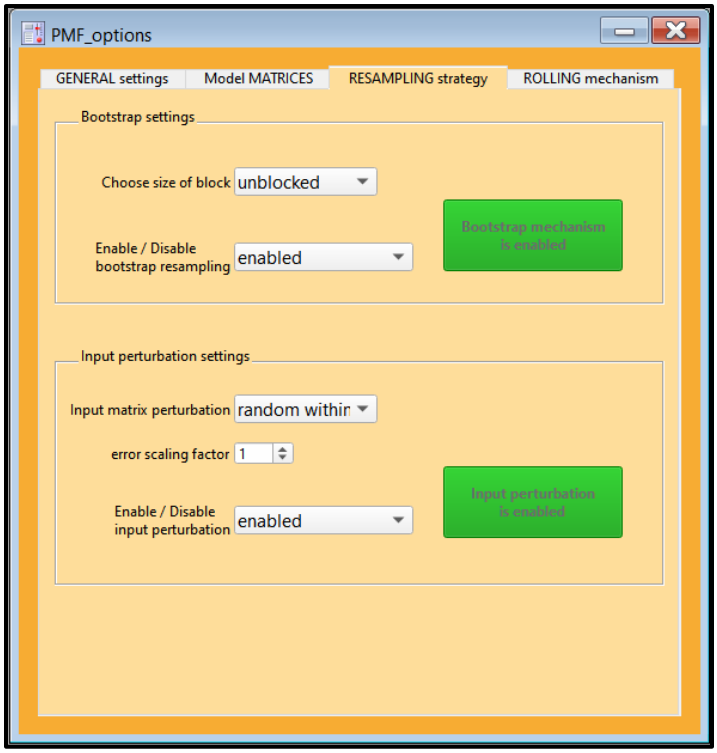


Figure 58 PMF settings, tab "RESAMPLING strategy"

6.4.2. Input perturbation

The input perturbation will randomly perturb the input matrix with a value within ± x * the error matrix. The scaling factor x can be defined by the user, by default it is set to 1 so that each value is randomly varied within ± its error. This technique is used to assess how sensitive the PMF solution is to variations in the input data and to estimate the robustness and uncertainty of the factorization results. By perturbing the input data multiple times, a range of PMF solutions can be generated. The variability across these solutions helps to estimate statistical uncertainties in the factor profiles and contributions. This provides a more realistic view of the robustness of the identified sources, taking into account the inherent uncertainties in the input data.

6.5. Rolling mechanism

The main limitation of PMF is the modeling of static factor profiles that stand in contradiction with a dynamic system, where source profiles are supposed to vary over time. To account for a possible temporal variation of the factor profiles a rolling approach was suggested (Parworth et al. (2015), Canonaco et al. (2021)) for ambient AMS/ACSM data.

Unlike traditional PMF, where the entire dataset is analyzed at once, rolling PMF applies the PMF model to small, overlapping windows of data that move across the timeline. This approach allows for tracking how factor contributions evolve over time (see Figure 59). To maximize the effectiveness of rolling PMF, it is essential to follow some best practices and guidelines:

- The PMF window must be smaller compared to the entire dataset, e.g. several days/weeks if the total PMF input comprises several years. More importantly, the window length should be chosen as such that the assumption of static factor profiles is legitimate.
- To avoid sharp transitions between consecutive windows, it's important to overlap the windows which ensures that the factors identified in adjacent windows evolve smoothly over time rather than abruptly changing from one window to the next.

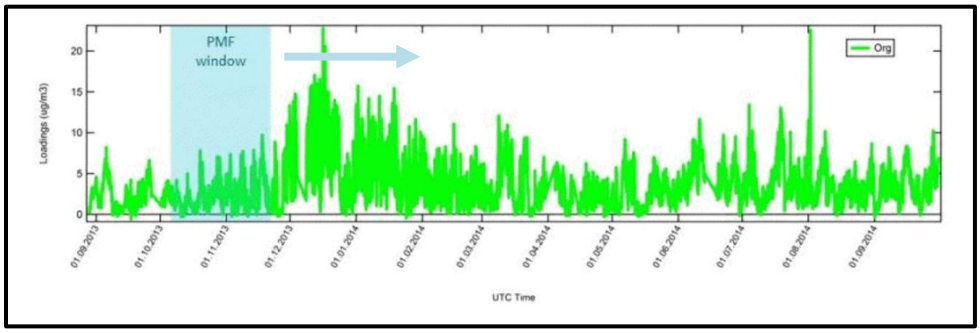


Figure 59 Schematic of a rolling PMF window.

Main Settings

“window width”	size of the PMF window expressed in days
“window shift”	shift of the PMF window expressed in days
“nb. of repeats”	estimate of the nb. of repeats due to the overlap of windows

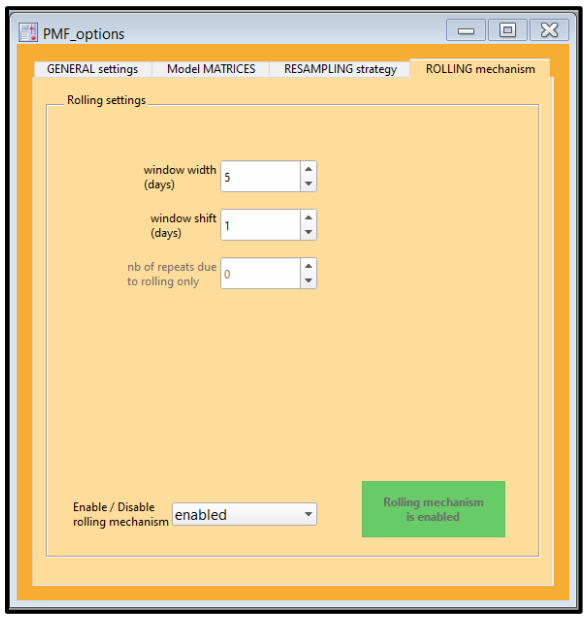


Figure 60 PMF settings, tab “ROLLING mechanism”

Ideally the rolling PMF approach is combined with the bootstrap resampling strategy to estimate the uncertainties in rolling PMF. Because each window contains fewer data points than the full dataset, fewer bootstrap runs are typically needed. Furthermore, if any factors are constrained, the rolling PMF approach should also be combined with the random α-value strategy. For more details please refer to Canonaco et al. (2021) and Chen et al. (2022) and analysis therein.

6.6. Call ME-2 and perform PMF

SoFi calls ME-2 and passes the PMF input with the current instruction file together with the defined and constrained information, if present, to ME-2.

Before the ME-2 solver is called, SoFi will give a short warning on how many PMF calls will be performed and whether the user wishes to continue. The progress can be monitored with the Progress bar “ME-2 call” at the end of the main panel (Figure 34). The progress bar “SoFi” reacts on calculations performed in IGOR for SoFi only.

7. SoFi – PMF result(s)

7.1. Import result files (HDF file(s))

7.1.1. Define path of ME2.exe file

The user gives the location where the ME2.exe file resides by pressing the button “folder of executable” (Figure 61). If the PMF run(s) were performed from the same experiment, this path is already defined. <u>However, the “PMF result” tab is completely independent from the previous tabs and PMF results can be consulted even if they have been generated from another IGOR experiment or from another PC</u>. The only restriction is given by the existence and the structure of the ME-2 folder under the OS. If the user wants to inspect some PMF results that was run on another PC, then the HDF5 file containing the PMF results must be stored in “ME2_Results” of the “ME2_engine” folder: …**\ME2_Engine\ME2_Results**.

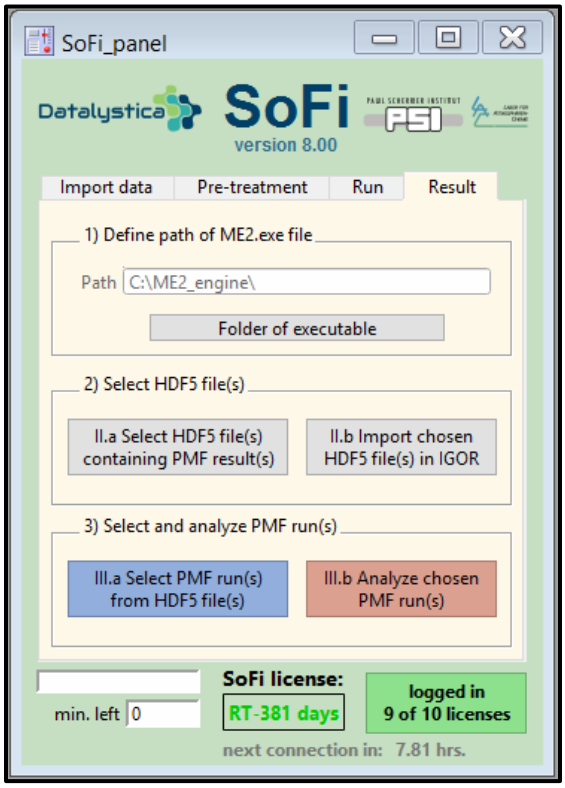


Figure 61 Main SoFi panel, PMF result tab.

7.1.2. Select HDF5 file(s) containing the PMF run(s)

The current SoFi version supports analysis of multiple PMF runs of a single PMF call (one HDF5 file) but also among various PMF calls (several HDF5 files).

The button "II.a Select HDF5 file(s) containing the PMF result(s)" opens a subpanel where the user decides which of the HDF5 files will be considered (Figure 62). The list contains all HDF files under the previously chosen solver location. This window must be closed by pressing on the "II.a […]" button again.

Multiple PMF runs can be chosen by pressing the "SHIFT" key selects adjacent PMF runs or by pressing the "CTRL" key to add non-adjacent PMF runs. Unselected PMF runs will not present in IGOR or if previously present, they will be removed after having asked the user with a prompt message. This is necessary to keep the IGOR experiment small to avoid possible "out of memory" errors. Removing the PMF runs from IGOR will obviously not affect their storage place on the HDF file(s).

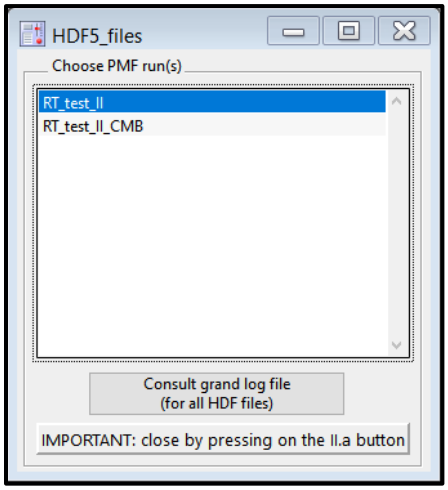


Figure 62 Subpanel showing the list of HDF5 files to be chosen for the PMF analysis.

The option "Consult grand log file (for all HDF files) will pop a list of the main parameters, e.g., nb. of factors, nb. of timepoints, types of constraints, etc. in each file listed. This allows the user to keep the overview over the conducted PMF runs without being forced to import each of them for visual inspection.

7.1.3. Import the chosen HDF5 file(s) into IGOR

Pressing on the button "II.b Import the chosen HDF5 file(s) into IGOR" will import the main information of the HDF5 file(s) to the corresponding folders. With this, only a subselection of the results, necessary for the plots in 7.2.2 is imported, not the full HDF file.

7.2. Manual selection of PMF run(s) form a HDF file

On the blue panel that pops after pressing “III.a Select PMF run(s) from HDF5 file(s)” (Figure 61), the user browses through the HDF5 files and selects the run(s) that should be kept for the subsequent PMF analysis (Figure 63).

7.2.1. Settings

“Choice base on”	Manual choice (default case), automated selection (see 7.3)
“Select HDF5 file”	List of imported HDF5 files to choose the PMF runs from
“avg. PMF run(s)”	Whether PMF runs should be averaged. For more information, see 7.5. This option is useful for the finalization and should only be used in combination with the criteria-based selection. Default is “no average”.
“x-axis type”	User defines the *x*-label for the graph (see Figure 63) from the list: run nb., nb. of factors, fpeak value, *α*-value, pulling value, C-value

7.2.2. Choose PMF runs(s)

As a small overview, the user has four different types of graphs to decide which PMF run(s) can be interesting for further inverstigation: Q/Q_{exp}, absolute values, relative values, explained variation plot. Whereas the Q/Q_{exp} graph shows the Q/Q_{exp} for each run, the other plots show a bit more details on how the factors are distributed: in absolute values, in relative values and also taking into account the (un)explained variation.

For selecting one or several PMF runs, the user best says in the Q/Q_{exp} tab, where an overview graph of the Q/Q_{exp} of each PMF run in the current (chosen) HDF file is shown. For selecting a PMF run the user drags a marquee around the bar and confirms the selection by clicking on “SoFi marquee” > “Selection options” > “SELect current” in the context menu of the right mouse button. The choice “UNSELect current” and “UNSELect all” resets the option for the current HDF file or for all HDF files, respectively. The gray bars (only x- value) representing PMF runs that should be further investigated must be in the marquee!

Chosen PMF runs are highlighted in the Q/Q_{exp} graph in blue. The marquee selection is additive and can be performed on any of the graphs present on the panel “Select_PMFruns” as well as on the popped graphs. But only in the Q/Q_{exp} graph the chosen vs non-chosen runs are distinguishable by color.

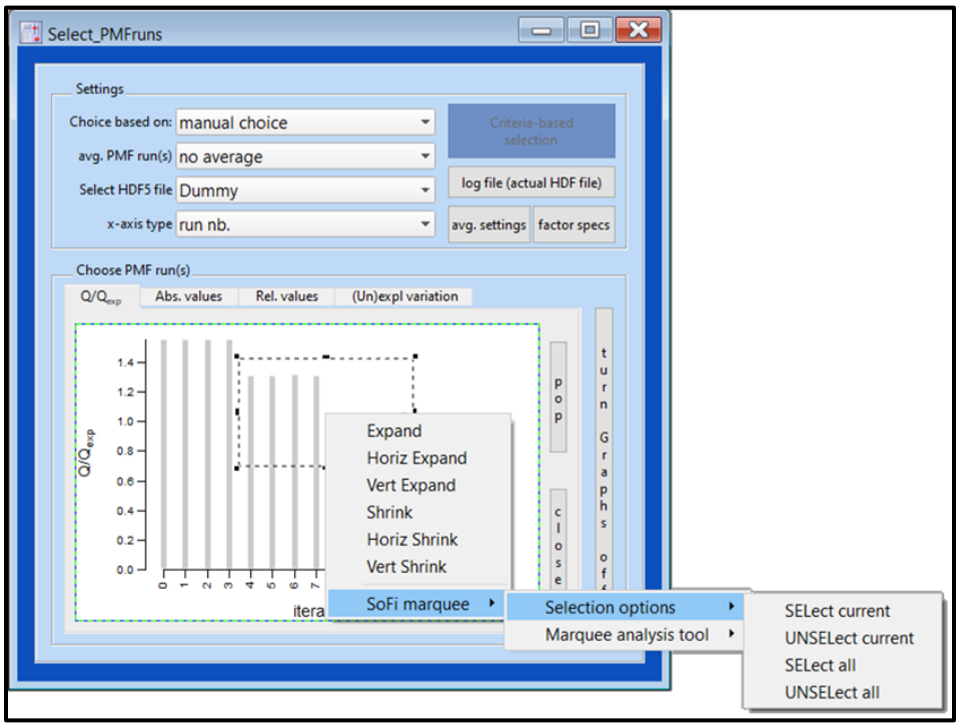


Figure 63 Selecting a set of solutions from the HDF5 file “Dummy” with the marquee.

7.3. Criteria-based selection

When a PMF analysis involves several hundred or even thousands of individual runs, as for instance during **bootstrapping** (see 6.4) or a **rolling window analysis** (see 6.5), it quickly becomes impractical to review each solution manually. To simplify this process, SoFi provides a **criteria-based approach**, that condenses the information from all runs into a few representative metrics, referred to as *criteria*.

Instead of examining every single PMF run, the user can define specific criteria that act as proxies for the overall model behavior. These may include, e.g., a correlation coefficient, a ratio between selected variables, or the sum of certain parameters. Advanced statistical tests for more objective selection of the proper PMF runs are also accessible. Moreover, 1D array, e.g., a particular factor profile or diurnal cycle and 2D arrays in form of scatter plot with user-defined x, y and color can be explored as a function of each PMF run (see below for more details).

Each run is then assigned a set of scores for the chosen criteria, which can be visualized and compared directly within SoFi. To access the criteria-based selection the user must select the option “automated selection” in the drop-down menu “Choice based on:” (Figure 64).

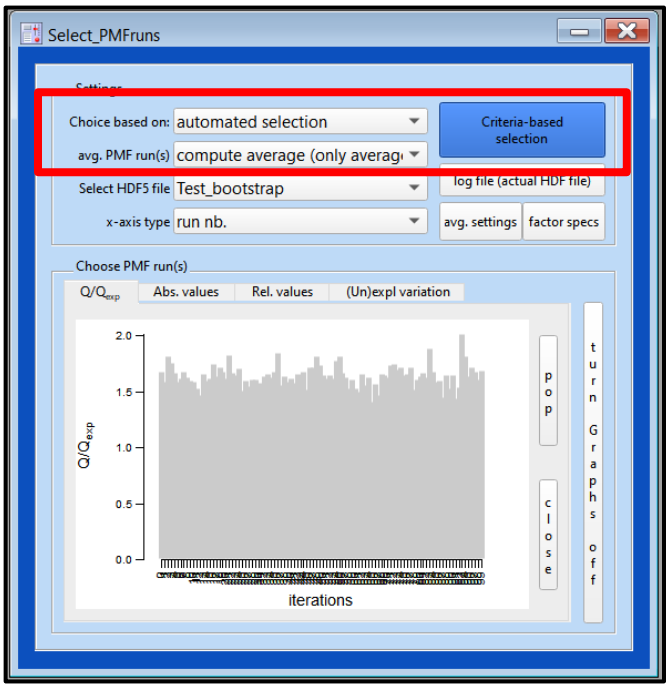


Figure 64 Select_PMFruns panel showing the criteria-based selection button.

Clicking on the button “Criteria-based selection” launches the subpanel “PMF_criteria” (see Figure 65), in which the criteria will be defined, the runs inspected, and environmentally reasonable solutions chosen.

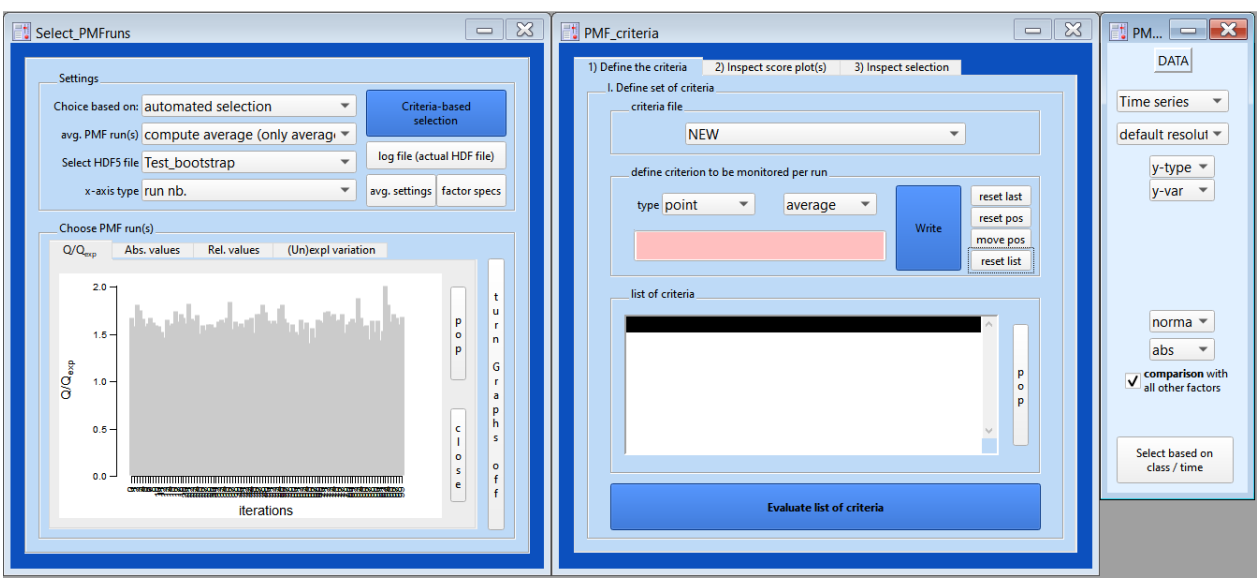


Figure 65 “Select_PMFruns” with its subpanel “PMF_criteria”

7.3.1. Type of criterion

The first tab in the "PMF_criteria" governs the definition and overview of the user-defined criteria (see Figure 66). Point criteria can be used for sorting.

"point"	single point is monitored for each PMF run
-average	single value (the final wave that could be a sum, product, ratio, etc. of other waves) is averaged.
-sum	single value is the sum of the final wave.
-Multilinear regr.	multilinear regression on PMF results. Up to ten independent variables can be defined
-R Pearson	R-Pearson value useful for linear relationships
-R Spearman	also rank correlation, useful for any type of relationship.
-Kendall tau	rank correlation coefficient that analyzes concordant and discordant pairs
-R uncentr	cosine similarity, use for profile information
"image"	1D array is monitored for each PMF run.
"movie"	2D array with user-defined colors, i.e., a colored scatter plot is monitored for each PMF run.

The various types of criteria are color-coded to allow to user to easily distinguish between them in the overview list "list of criteria" (Figure 66).

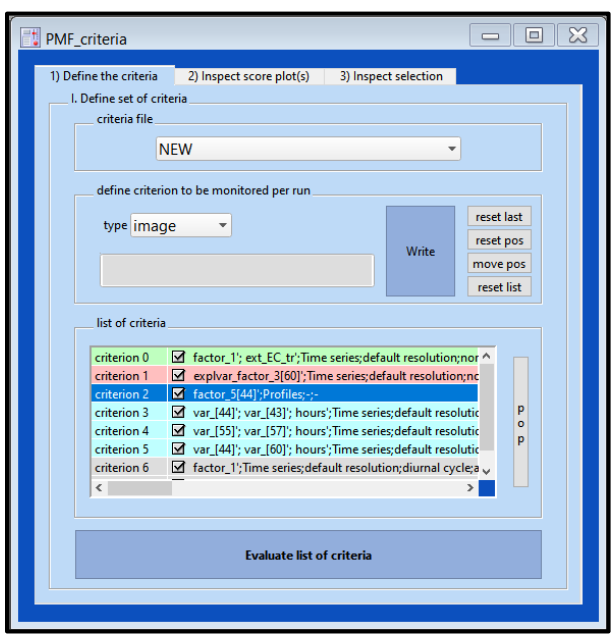


Figure 66 Different colors for the different types of criteria help to easily distinguish them.

7.3.2. Criteria list

Several different criteria can easily be defined in SoFi and the risk of losing track of which criterion was defined for which factor is high. Therefore, a small overview table, including the color-code of type of criteria, showing the defined criteria is always visible from the first tab. A more detailed batch table containing all information to a criterion, like exact time point chosen, type of data (abs or fraction) or resolution, can be consulted from the second tab (see Figure 78).

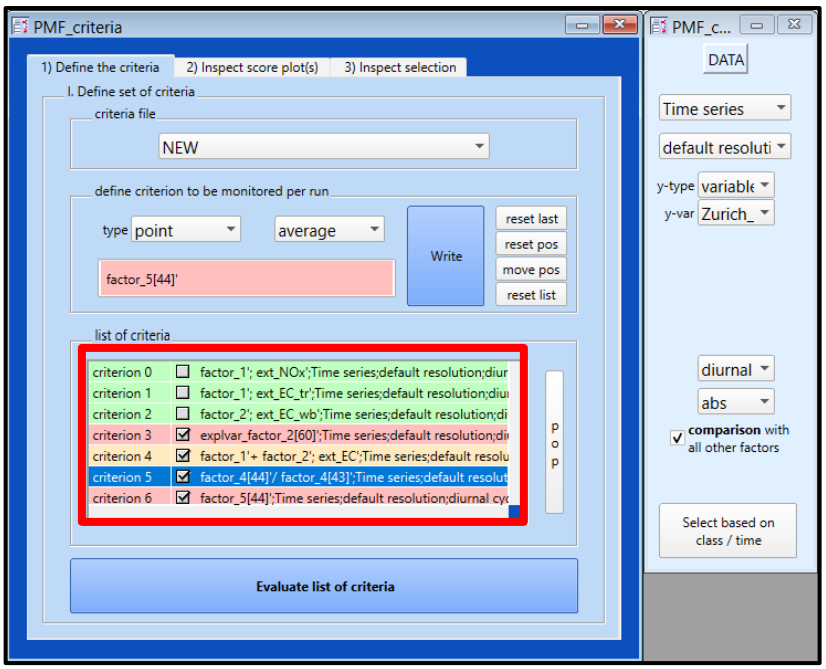


Figure 67 Several criteria visualized in the criteria list with the proper color-code. Criteria in bold are those that contain a temporal sub-selection. Highlighted criteria are used for repositioning unconstrained factors.

Criteria that must be evaluated are checked. If a criterion has already been evaluated in a previous round, there is no need to re-evaluate it again (besides when it is used for sorting). This saves a lot of computational time and makes the criterion analysis faster. Criteria that should be used for repositioning (sorting) of unconstrained factors must be selected and are highlighted in the criteria list (see below under “The criteria approach on unconstrained factors” and “active versus passive criteria” for more details (see Figure 67)). Furthermore, constraints always overpower sorting criteria, meaning that constraint factors cannot be sorted but are fixed in the position that they were constrained in. Criteria that contain a sub-selection from the panel “Select based on class / time”, are visualized in **bold** in the criteria list.

7.3.3. Active (sorting) versus passive (monitoring) criteria

In case many PMF runs with several unconstrained factors are run and the solutions are not inspected manually, the problem arises that the positions of the factors will most likely vary.

Table 6 Example with 4 factors and 2 PMF runs

PMF run 1				
position	1	2	3	4
factor	HOA	COA	BBOA	OOA
PMF run 2				
position	1	2	3	4
factor	HOA	BBOA	COA	OOA

Swapping occurred for position 2 and 3 between the PMF runs 1 and 2. The factors need to be repositioned based on certain conditions / criteria, to guarantee that position 1 carries always HOA, position 2 COA, position 3 BBOA and position 4 OOA.

The problem of repositioning factors on available positions is the classic "sampling without replacement" problem that scales with the factorial of the number of positions, e.g. for 6 unconstrained positions there are 6! or 6*5*4*3*2 = 720 possibilities. SoFi computes all combinations and evaluates all defined criteria. The scores are then internally weighted (z-score followed by gaussian rank transformation), so that they all have similar weight and range. SoFi Pro employs the Hungarian algorithm to optimally sort factors. While the Hungarian algorithm is highly effective, its computational cost increases rapidly with the number of unconstrained factors. E.g., for 10 unconstrained factors there are already 3'628'800 combinations per run, so constraining known and very stable factors (e.g. fixed elemental ratios) reduces the dimensionality of the sorting problem and allows SoFi Pro to focus the computational effort on the remaining unconstrained components. The combination with the highest sum of scores is returned as chosen combination for ordering the factors. This guarantees that the unconstrained factors are score-sorted in the best possible way.

A criterion used for repositioning an unconstrained factor is a so-called active criterion. For each factor, only one active (sorting) criteria can be defined, unless it is a constrained factor, then the constraint overpowers the sorting. A passive criterion, will simply give the score for the current position, without affecting its sorting. Passive criteria are typically used for constrained factors, where the factor position is already fixed and known in advance, or as further criteria for factors that are being sorted.

The user governs whether a criterion is active / passive directly from the criteria list by selecting the active ones with the mouse by clicking on “criterion x”. Active criteria are highlighted (see Figure 67). All “point” criteria besides the multilinear regression can be used for sorting/as an active criteria.

7.3.4. Criterion syntax

The “DATA” subpanel (see) provides all options required to define specific criteria. The following list is shown for a temporal criterion (ts). Options for criterion on profiles are characterized by (pr). Each data type is internally distinguished by a characteristic prefix or suffix (shown in bold brackets in the list below). These identifiers are essential for SoFi to correctly recognize and evaluate the respective criterion. It is therefore important that users do **NOT** modify these prefixes or suffixes. When criteria are entered manually, the appropriate prefix or suffix must always be included; otherwise, SoFi will not be able to process the criterion properly. Note that the features from this subpanel are typically additive. This means that existing information in the criterion line is not overwritten but new information is rather appended to the preexisting one.

“Time series, Profiles”	to define the type of data (contributions (ts) or profiles (pr))
“default, hourly, daily res.”	to define the temporal resolution
“y (x)-type”	solution (ts, pr), variable of solution (ts, …**[var.]**), external (ts, pr, **ext_**), variable (ts, **var_**), res. (ts, **RES_**), abs. res. (ts, **ABSRES_**), scaled res. (ts, **RESW_**), abs. scaled res. (ts, **ABSRESW_**), Q res. (ts, **QRES_**), explained variation (ts, **explvar_**), hours (ts, …**[hour]**)
“y (x)-var”	varies depending on the “y-type” selection
“normal, cycles”	to define whether normal, daily, weekly, monthly or yearly cycles should be used
“abs, fraction”	to define whether absolute data or fraction, i.e., data normalized to the sum should be used
“comparison with all other”	additional p-values from significance tests and information about highest other score (in different factor) can be calculated
“select based on class/time”	option to evaluate criteria based on class or time. Note, the computational time of the criteria will increase when a specific class or time has been chosen. Check 5.3.1 or 5.3.2 for more information about this time/class selection.

There are only few rules to the criterion line:

1) **The pre and suffixes described earlier in "The subpanel DATA" MUST be strictly followed**
2) **EACH term containing a wave specification, e.g. factor 1 or RESW[total] as well as at the end of each criterion line MUST end with a single quotation mark '.**
3) **EACH term MUST start with either a space (the case for the first term) or some mathematical symbols, e.g. "()", "+", "-", "*".**

This is essential, as the evaluation of each term has the following architecture:

"any symbol"-"term"-"any symbol"-"quotation mark" (termination)

If those rules are not followed, SoFi will fail when treating the individual strings and this causes a bad evaluation of the criterion or unspecific IGOR error messages. Therefore, make sure you know the criteria *syntax* before changing and defining new manual criteria from the criterion line.

In addition to this subpanel, the user can directly change, add or remove terms from the criterion line and perform simple mathematical operations that are also typically supported from the IGOR command line.

<u>Mathematical operations currently supported from the criterion line</u>

Table 7 mathematical operations for criteria definitions.

mathematical operation	**examples**
addition, subtraction, multiplication, division	wave1' + wave2', wave1'-wave2', wave1'*wave2', wave1'/wave2'
exponentiation	exp(wave1'), wave1'^2, 10^wave1'
square root	sqrt(wave1')'
min, max	min(wave1', wave2')', max(wave1', wave2')'
histogram(x min, x max, nb. of bins, wave)	histogram(-5, 5, 1000, wave1')'
rules with brackets	(wave1'+wave2')*(wave3'+wave4')/wave5'
constant values	wave1'+wave2'+5', wave1'+wave2'+5'
Some combined examples	(wave1'+wave2')/2+sqrt(wave3'*wave4'^2)'
	min(histogram(-5,5,100, wave1'), histogram(-5,-5,100, wave1'))'

In these examples wave1, wave2, wave3, wave4 and wave5 can be anything from the options presented under presented under “The subpanel DATA”.

For the expressions min(wave1, wav2) and max(wave1, wave2) the comparison is performed on a point-by-point basis for the two waves. The result wave has the dimension of wave1 or wave2 (they must have the same dimension) and contains for each point *i* the smallest value, i.e., min(wave1[i], wave2[i]) is performed for all points *i*.

histogram(x min, x max, nb. of bins, wave1) creates the histogram for the specified wave1 with the parameters, x min, x max and nb. of bins. Selecting the option “fraction” normalizes the sum of the histogram to 1. Note that the histogram option can be applied only once per criterion. Moreover, histogram is interesting when comparing the residuals of the classes on a bin-by-bin basis in combination with the option min(), see with the examples a few pages below.

A combined criterion

The user has the possibility to combine two and more criteria into a new one. This is important when for example one criterion is not enough to capture the desired expression. This is typically the case, when, e.g. various time or class settings (subpanel ‘Select based on class / time’) should be combined.

The combined criterion has the following syntax: criterion X’ + criterion Y’

where X and Y stand for the positions of the two criteria. This example shows the sum, but any mathematical combination as well as combinations between normal and GRAND criteria are in theory possible.

7.3.5. Examples on how to set up the list of criteria

Some examples that illustrate how to set up a criterion follow here.

7.3.5.1. Correlation between the contribution of factor 1 (hypothetical traffic) and NO_x (see Figure 68)

1) Under "type" select "point" and "R-Pearson" or "R-Pearson^2". Note that the box turns green as this is the color-code for the Pearson option.
2) In the subpanel "DATA" select the option "Time series", "solution" for "y-type" and then the first entry, i.e., factor_1. In addition, select "external" for "x-type" and NO_x data.
3) At this stage, the criterion line shows **factor_1';ext_NOx'.**
4) Clicking on "Write" will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.
5) The successfully written criterion is reported under "list of criteria".

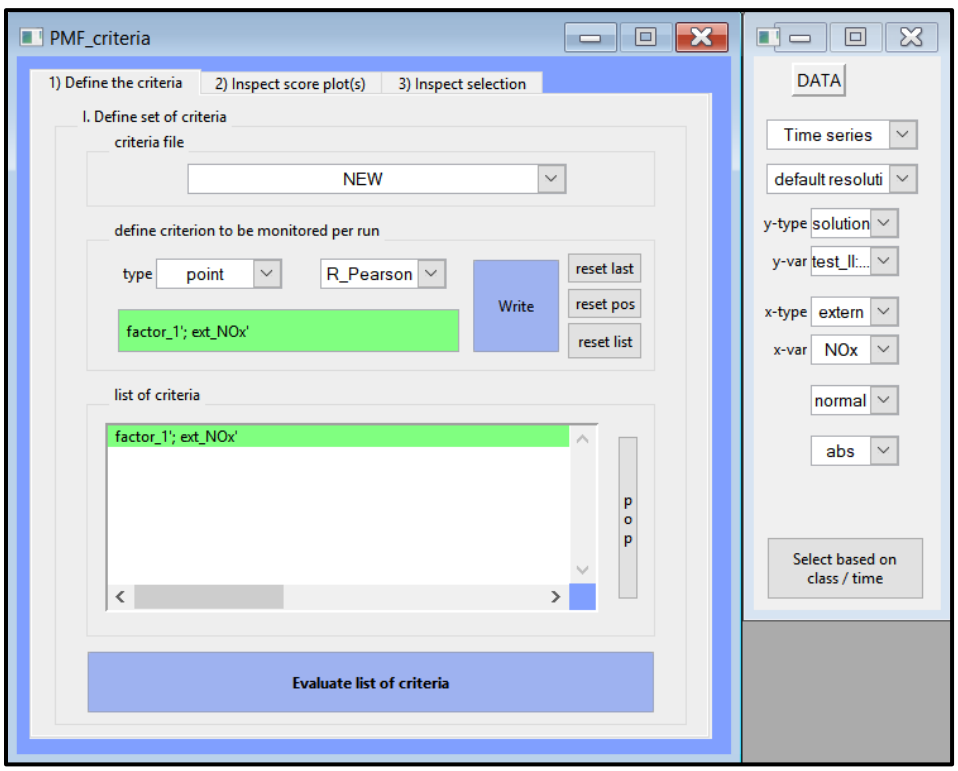


Figure 68 "PMF_criteria_define" subpanel with the example of R_pearson between factor1 and NOx data

7.3.5.2. **Ratio between certain hours of the day for factor 2 (see Figure 69)**

1) Under “type” select “point” and “average”. Note that the box turns red as this is the color-code for the average option.

2) In the subpanel “DATA” select the option “Time series”, “hours” for “y-type” and then select “factor_2[hr11].

3) At this stage, the criterion line shows **factor_2[hr12]’.**

   The user can directly add the remaining information, i.e., **+ factor_1[hr12]')/2)/((factor_1[hr6]'+ factor_1[hr7]'+ factor_1[hr8]')/3)’**

   This expression will evaluate the ratio between the peak at noon (11 am and 12 pm) compared to the average value of 6, 7 and 8 am.

4) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

5) The successfully written criterion is reported under “list of criteria”.

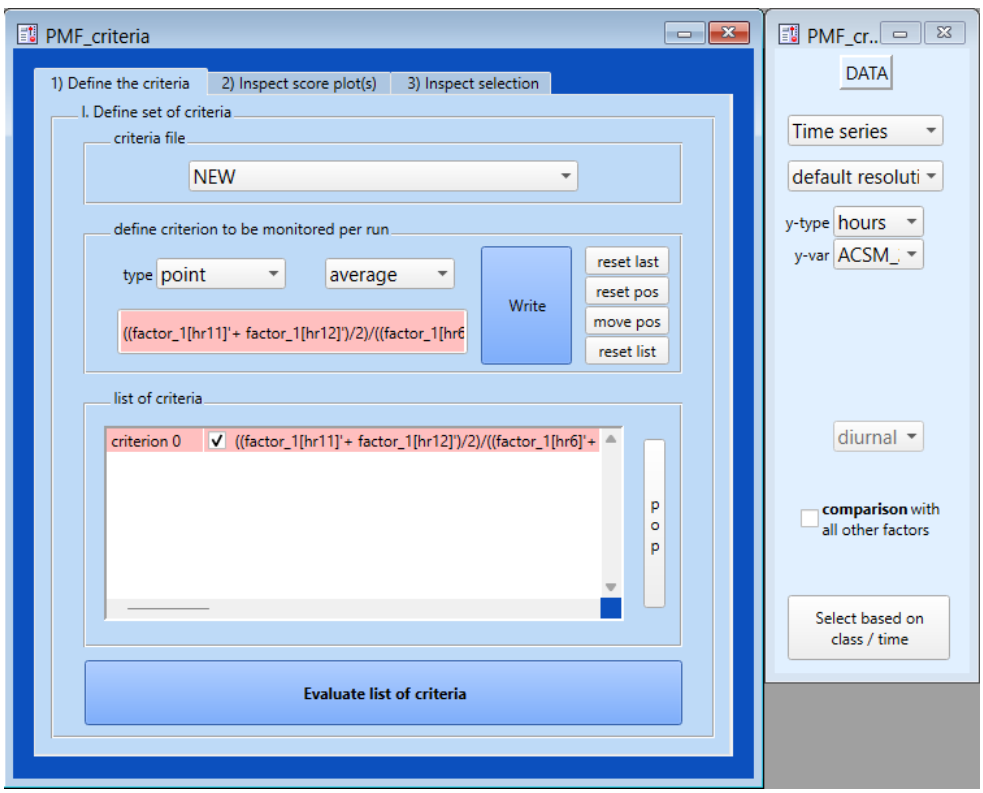


Figure 69 “PMF_criteria_define” subpanel with the example of the ratio between the lunch peak to some morning hours for factor 2.

7.3.5.3. **Monitoring the mass fraction of m/z 60 (f60) of the third factor (see Figure 70)**

1) Under “type” select “point” and “average”. Note that the box turns red as this is the color-code for the average option.

2) In the subpanel “DATA” select the option “Time series”, “solution of var.” for “y-type” and then select “factor_3 60:60. In addition, select also “fraction”.

3) At this stage, the criterion line shows **factor_3[60]’.**

   This expression will evaluate the mass fraction of m/z 60 of the third factor.

4) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

5) The successfully written criterion is reported under “list of criteria”.

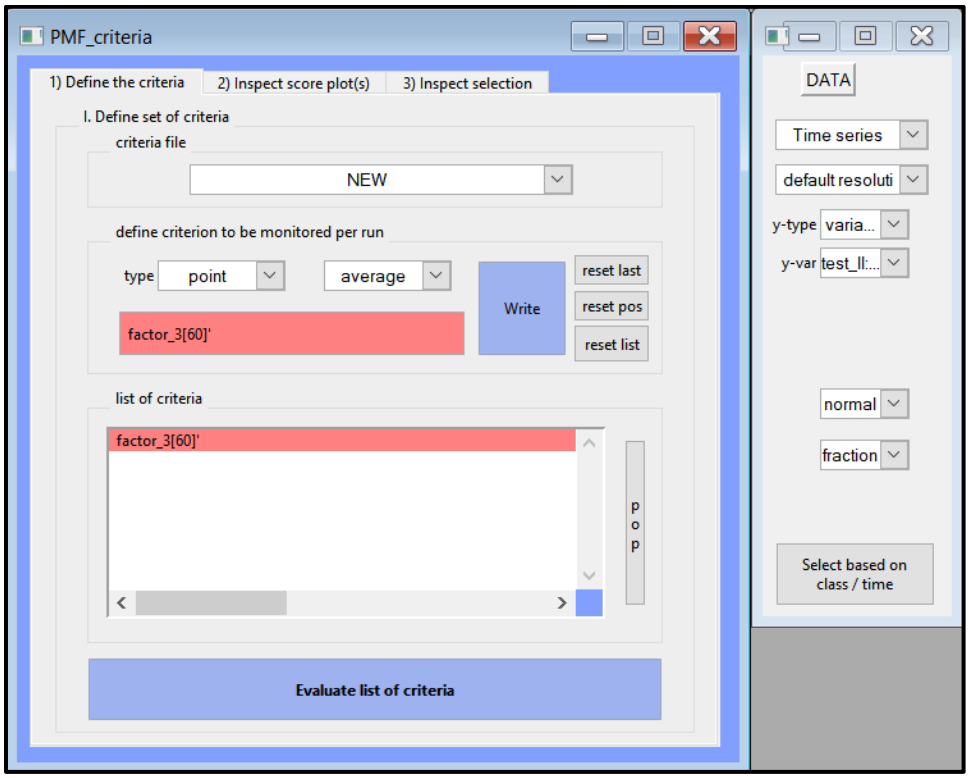


Figure 70 “PMF_criteria_define” subpanel with the example of *f*60 for the third factor.

7.3.5.4. **Monitoring the correlation value R upon a multilinear regression (see Figure 71)**

The linear regression is between the time series of factor 1 (hypothetical traffic) and factor 3 (hypothetical biomass burning) to black carbon EC, assuming only traffic and biomass burning to contribute to black carbon.

1) Under "type" select "point" and "Multilinear regression". Note that the box turns orange as this is the color-code for the Multilinear regression option.

2) In the subpanel "DATA" select the option "Time series", "solution" for "y-type" and then select "factor_1". In addition, select "external" for "x-type" and then "EC".

3) At this stage, the criterion line shows **factor_1'; ext_EC'.** The user can directly add the remaining information, i.e., **+ factor_3'** to the first part of the string. The final expression reads: **factor_1'+ factor_3'; ext_EC'**

   This expression will perform a multilinear regression and find the optimal values for the slopes a and b in: EC = a·factor_1 + b·factor_3. EC is the dependent variable and factor_1 and factor_3 are the two independent variables. <u>The current option allows the user to add up to 5 independent variables and always only one dependent variable. For sorting, this criterion is only useful to a limited extent.</u> For sorting, all factor combinations should be tested. However, SoFi can only iterate over the last-mentioned factor in the combination, meaning that with the current criterion, factor_1 remains fixed while combinations with other factors are evaluated. Hence we recommend using alternative criteria for sorting.

4) Clicking on "Write" will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

5) The successfully written criterion is reported under "list of criteria".

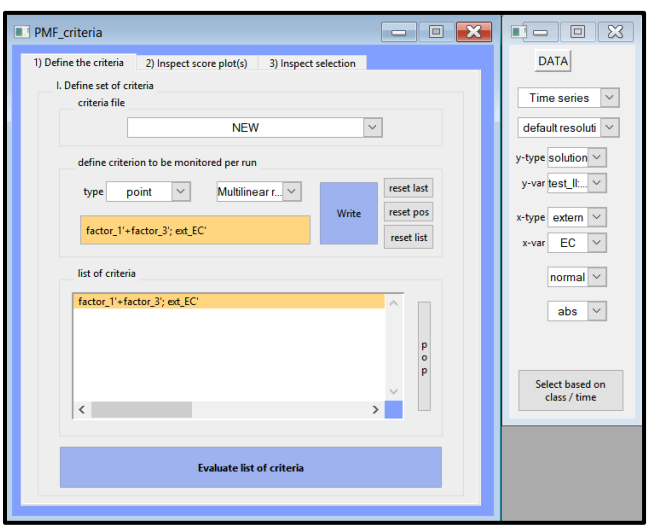


Figure 71 "PMF_criteria_define" subpanel with the example of the multilinear regression between EC and factor_1 and factor_3

7.3.5.5. Monitoring the ratio between the fraction of *m/z* 57 and *m/z* 55 for the second factor (see Figure 72)

1) Under “type” select “point” and “average”. Note that the box turns red as this is the color-code for the average option.

2) In the subpanel “DATA” select the option “Profiles”, “solution” for “y-type” and then select “factor_2 57:57.

3) At this stage, the criterion line shows **factor_2[57]’.**

   The user can directly add the remaining information, i.e., **/factor_2[55]’.**

   This expression will evaluate the ratio between *m/z* 57 and *m/z* 55 for the second factor.

4) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

5) The successfully written criterion is reported under “list of criteria”.

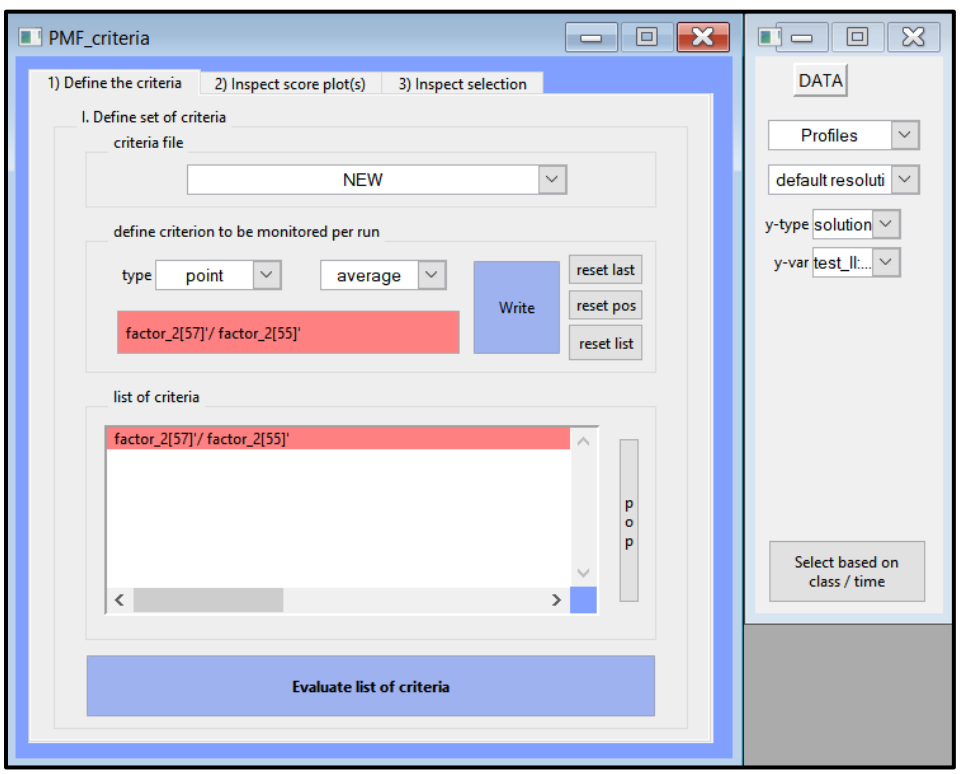


Figure 72 “PMF_criteria_define” subpanel with the example of the ratio of *m/z* 57 to *m/z* 55 for the second factor.

7.3.5.6. **Correlation on the diurnal cycle between the contribution of factor 1 (hypothetical traffic) and NO_x for the weekdays only (see Figure 73)**

1) Under “type” select “point” and “R-Pearson” or “R-Pearson^2”. Note that the box turns green as this is the color-code for the Pearson option.
2) In the subpanel “DATA” select the option “Time series”, “solution” for “y-type” and then the first entry, i.e., factor_1. In addition, select “external” for “x-type” and NO_x data.
3) At this stage, the criterion line shows **factor_1’;ext_NOx’.**
4) Select diurnal cycle and weekdays from the “Select based on class / time” panel
5) Clicking on “Write” will write this criterion and pass the selected information that is used for the evaluation.
6) The successfully written criterion is reported under “list of criteria”. Note that the criterion is bold, as a temporal sub-selection from the panel “Select based on class / time” is applied.

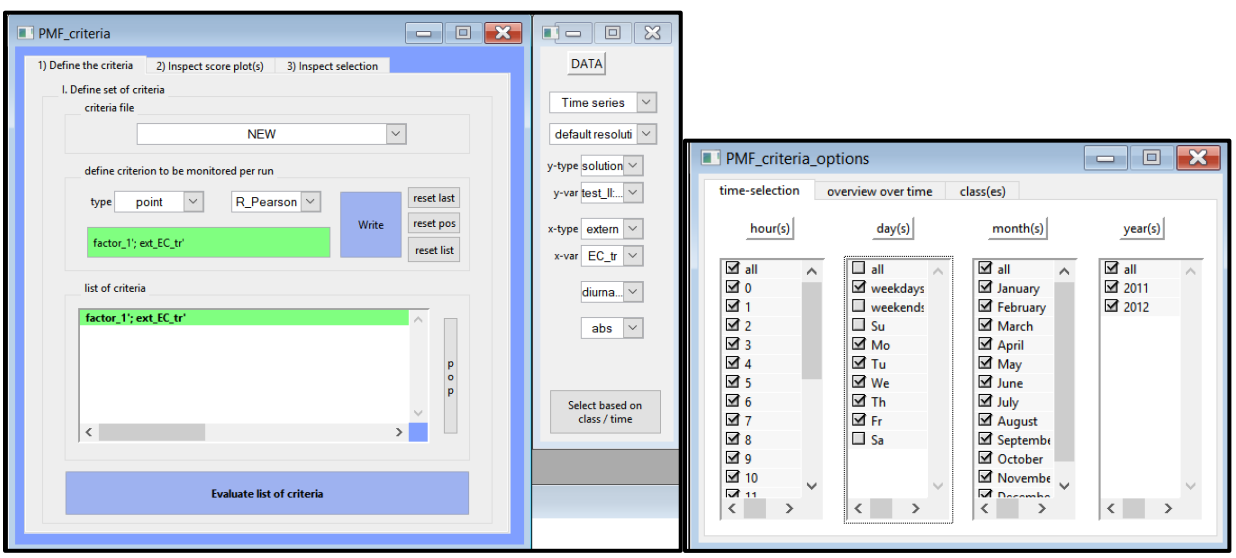


Figure 73 “PMF_criteria_define” subpanel with the example of R_pearson between factor1 and NOx data for a specific temporal sub-selection.

7.3.5.7. Monitoring the ratio between the mass fraction of *m/z* 44 for the fourth factor to the scaled residual of m/z 44 (see Figure 74)

1) Under “type” select “point” and “average”. Note that the box turns red as this is the color-code for the average option.

2) In the subpanel “DATA” select the option “Time series”, “solution of var.” for “y-type” and then select “factor_4 44:44.

3) At this stage, the criterion line shows **factor_4[44]’.** The user can directly add the remaining information, i.e., **/RESW_[44]'.**

   This expression will evaluate the ratio between the mass fraction of *m/z* 44 of the fourth factor that is supposed to highly contribute to *m/z* 44 and the scaled residual of *m/z* 44.

4) Clicking on “Write” will write this criterion and pass the information that all data (no temporal sub-selection was chosen) is used for the evaluation.

5) The successfully written criterion is reported under “list of criteria”.

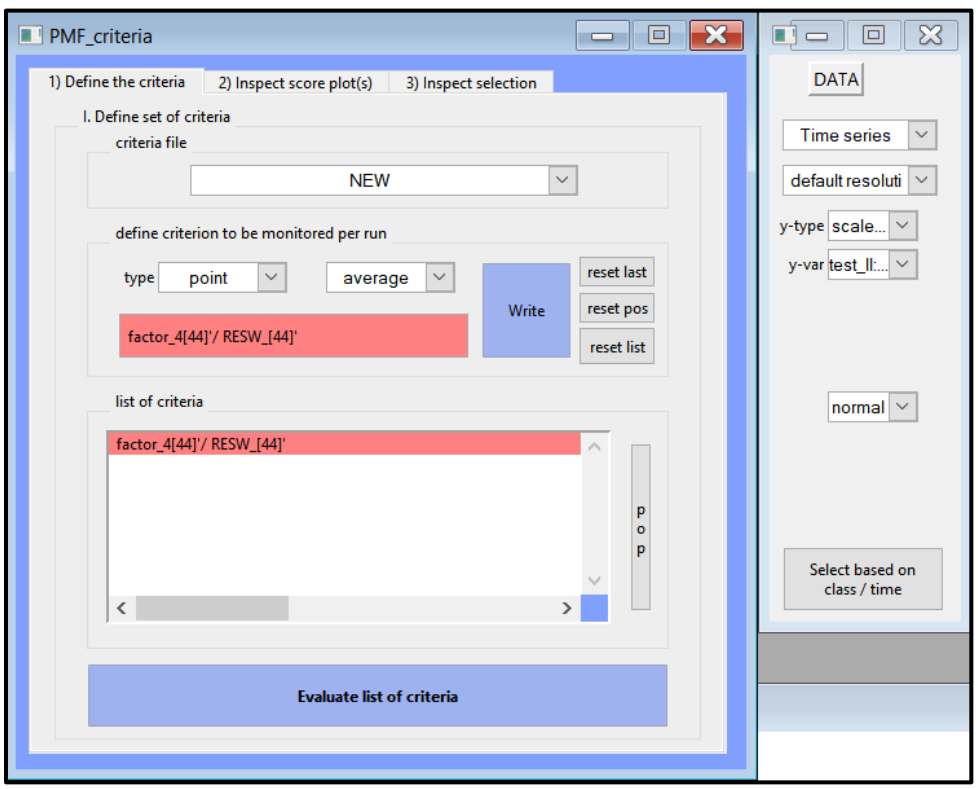


Figure 74 “PMF_criteria_define” subpanel with the example of the mass fraction of *m/z* 44 of the fourth factor to the scaled residual of *m/z* 44.

7.3.5.8. **Monitoring C-value runs with a combined criterion**

When data from two different instruments, e.g. AMS and PTR-MS, are combined, a relative rescaling is performed using the C-value approach (see 6.2.4.4). Ideally, the scaled residuals for each instrument should be comparable (same mean value and spread), i.e., their histograms should have maximal overlap. This can be monitored using the following scheme of criteria as example:

1) Under "type" select "point" and "average". In the subpanel "DATA" select the option "Time series", "scaled res." for "y-type, " and then select ""total_matrix". At this stage, the criterion reads RESW[total_matrix]'.

2) Add the histogram options (min, max and number of bins) to the criterion line, so that it reads e.g. **histogram(-5,5,1000,RESW[total_matrix]')'.** Select "fraction" and select **class 1 (e.g. AMS)** from the "select based on class/time" option. Then write the criterion into the list.

3) Repeat, but this time, select **class 2 (e.g. PTR-MS)**, from the "select based on class/time" option.

4) Now create the criterion where SoFi actually compares the histograms. Under "type" select "point" and "sum". Check that **both classes** are checked.

5) Manually write **min(criterion 0',criterion 1')'** and add the criterion to the list.

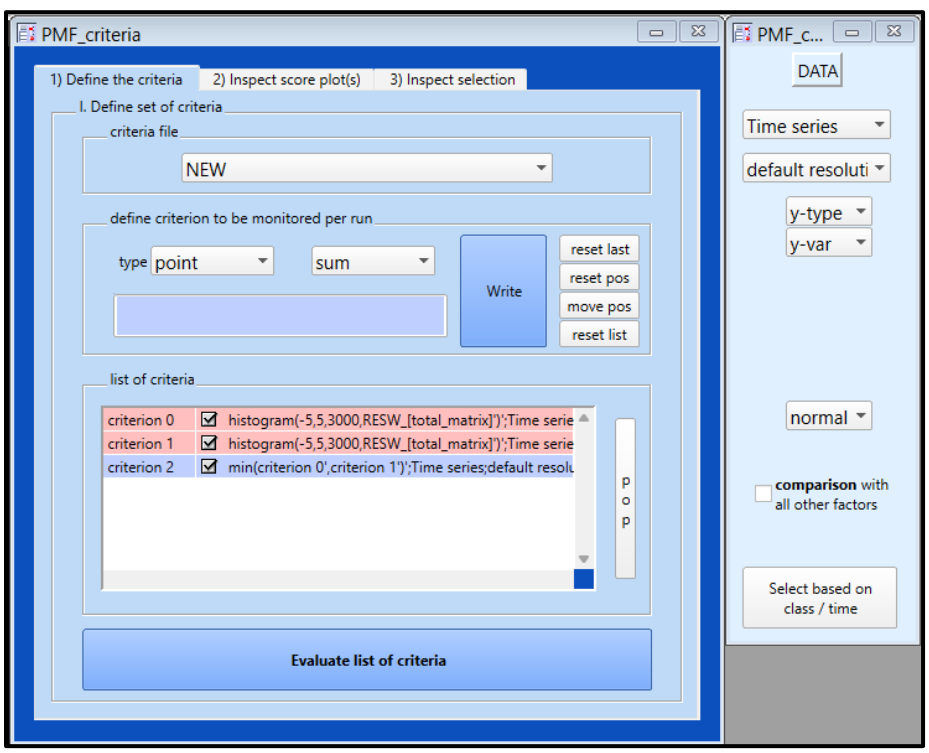


Figure 75 "PMF_criteria_define" subpanel with the example the histogram approach for C-value PMF.

7.3.6. Evaluate the criteria

Pressing on the button “Evaluate list of criteria” allows SoFi to loop over all PMF runs belonging to the current HDF file and to evaluate the user-defined list of criteria (see Figure 76). SoFi will (re-)evaluate all criteria that have a checked checkbox. To save time when updating only one criterion, the user can uncheck all criteria that don’t need to be reevaluated.

In addition, the dropdown menu under “criteria file” controls saving/loading options of the list of criteria.

“NEW”	creates a new criteria list. If the current list is not empty, it will be deleted.
“LOAD list of criteria”	A criteria list stored as an IGOR txt file can be imported
“LOAD list of criteria and evaluation”	A criteria list including its scores (makes only sense if the criteria file to be loaded belongs to the same HDF file) stored as an IGOR txt file can be imported
“SAVE list of criteria”	A criteria list will be saved in an IGOR txt file
“SAVE list of criteria and evaluation”	A criteria list including its scores will be saved as an IGOR txt file

If a criteria list is loaded, it should be from the same experiment, as indices are loaded as well. If those do not match, the evaluated scores will be NaN.

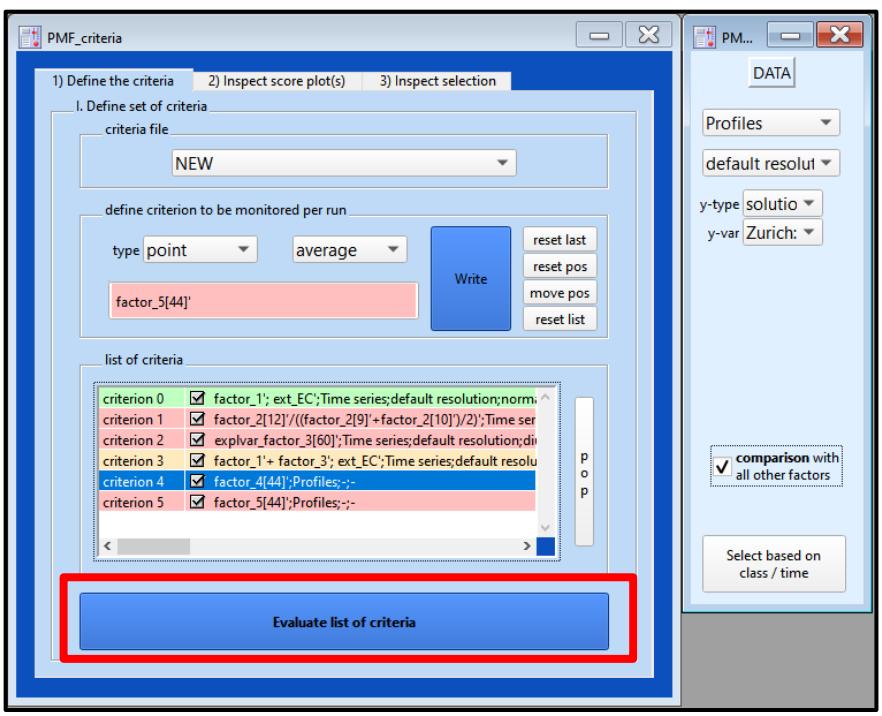
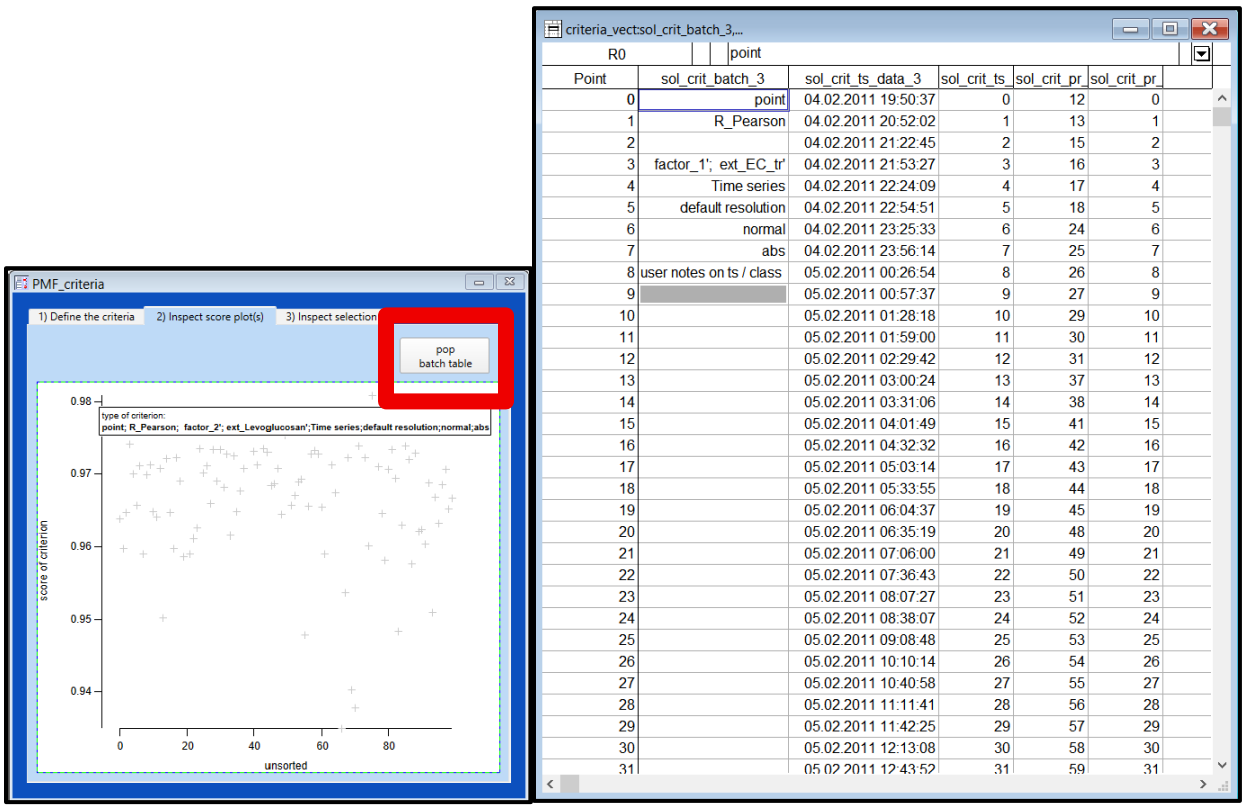


Figure 76 “PMF_criteria” showing the list of possible user-defined criteria. Only checked criteria will be evaluated. Bold criteria are taking into account class/time, highlighted criteria are used for sorting.

7.3.7. Inspect criteria score

SoFi reports the scores for each criterion and PMF run in the 2nd tab “2) Select PMF run(s)” of the “PMF_criteria” panel (see Figure 78). The default case reports the scores for all PMF runs as is, i.e., the PMF runs are unsorted. Using the adjacent option subpanels allows the user to control various aspects of the scores to better judge which PMF runs are worth being further investigated as a whole.

The button “pop batch table” will open a table which contains all the information on what was selected in the previous tab for the selected criterion (see Figure 77). Note that the first 8 entries in the batch file are reserved and the batch file is not removed, as long as the experiment stays imported into IGOR and the criteria are not removed. Therefore, the user can use the cell of the 8th position to enter user-specific notes, if required.



Point	sol_crit_batch_3	sol_crit_ts_data_3	sol_crit_ts_	sol_crit_pr_	sol_crit_pr_
0	point	04.02.2011 19:50:37	0	12	0
1	R_Pearson	04.02.2011 20:52:02	1	13	1
2		04.02.2011 21:22:45	2	15	2
3	factor_1'; ext_EC_tr'	04.02.2011 21:53:27	3	16	3
4	Time series	04.02.2011 22:24:09	4	17	4
5	default resolution	04.02.2011 22:54:51	5	18	5
6	normal	04.02.2011 23:25:33	6	24	6
7	abs	04.02.2011 23:56:14	7	25	7
8	user notes on ts / class	05.02.2011 00:26:54	8	26	8
9		05.02.2011 00:57:37	9	27	9
10		05.02.2011 01:28:18	10	29	10
11		05.02.2011 01:59:00	11	30	11
12		05.02.2011 02:29:42	12	31	12
13		05.02.2011 03:00:24	13	37	13
14		05.02.2011 03:31:06	14	38	14
15		05.02.2011 04:01:49	15	41	15
16		05.02.2011 04:32:32	16	42	16
17		05.02.2011 05:03:14	17	43	17
18		05.02.2011 05:33:55	18	44	18
19		05.02.2011 06:04:37	19	45	19
20		05.02.2011 06:35:19	20	48	20
21		05.02.2011 07:06:00	21	49	21
22		05.02.2011 07:36:43	22	50	22
23		05.02.2011 08:07:27	23	51	23
24		05.02.2011 08:38:07	24	52	24
25		05.02.2011 09:08:48	25	53	25
26		05.02.2011 10:10:14	26	54	26
27		05.02.2011 10:40:58	27	55	27
28		05.02.2011 11:11:41	28	56	28
29		05.02.2011 11:42:25	29	57	29
30		05.02.2011 12:13:08	30	58	30
31		05.02.2011 12:43:52	31	59	31

Figure 77 Batch table for criterion number 3. The first column contains the batch file with information on the criterion. The remaining columns represent the effective time series, its time series index, the effective variables and its index, respectively.

<u>The subpanel “DATA”</u>

“criteria”	defines which criterion should be investigated
“factors”	If PMF call contains runs with different number of factors

"sorting"	scores of criteria can be "unsorted", sorted in "ascending order" or "descending order" (see below for more details)
"other highest"	plot the score of either the 2nd highest (if score is highest) or total highest (in case score is not highest compare to all others) classified score (see below for more details)
"p-values"	plot p-values based on a T-test, and also p-values for the correlation coefficients, if available. This information can be directly used to exclude correlation coefficient that are not statistically significant.
"… MARQUEE selection…"	use exact limits instead of the marquee selection in the plot
"Find PMF runs…"	SoFi finds the PMF runs that fulfill the user selection on all criteria. This option is crucial when selecting on sorted scores.
"nb. of common runs:"	The number of common PMF runs is visualized here
"Apply selection"	The runs in common are directly passed to the marquee selection (see Figure 63) ready to be further analyzed in SoFi

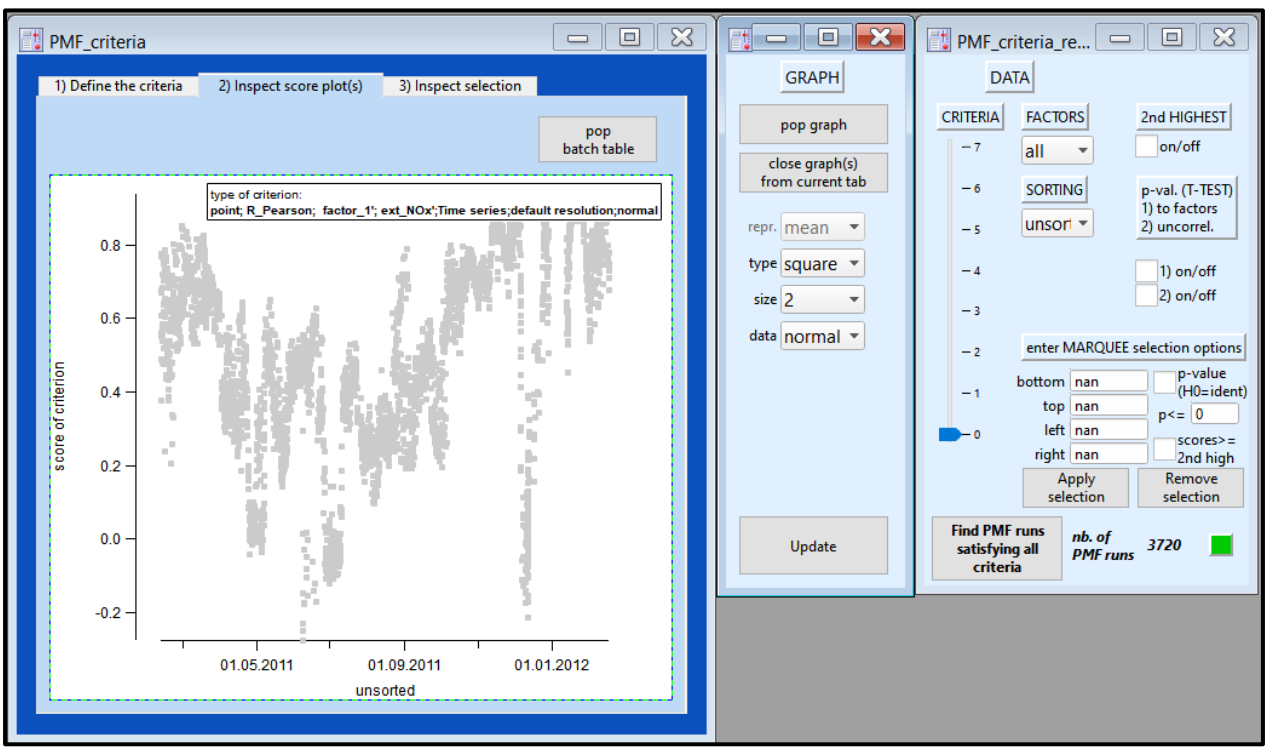


Figure 78 Criterion *f*43 for the profile of factor 5 (gray line). The adjacent subpanel governs all options related to the selection of a PMF run.

The task of the user is to select, using the marquee tool, scores (PMF runs) that should be investigated as full PMF result. To better visualize the range and density of the scores, the score plot can be sorted either in ascending or descending order. When e.g. high scores are expected, the user can more easily select the proper scores in a sorted plot.

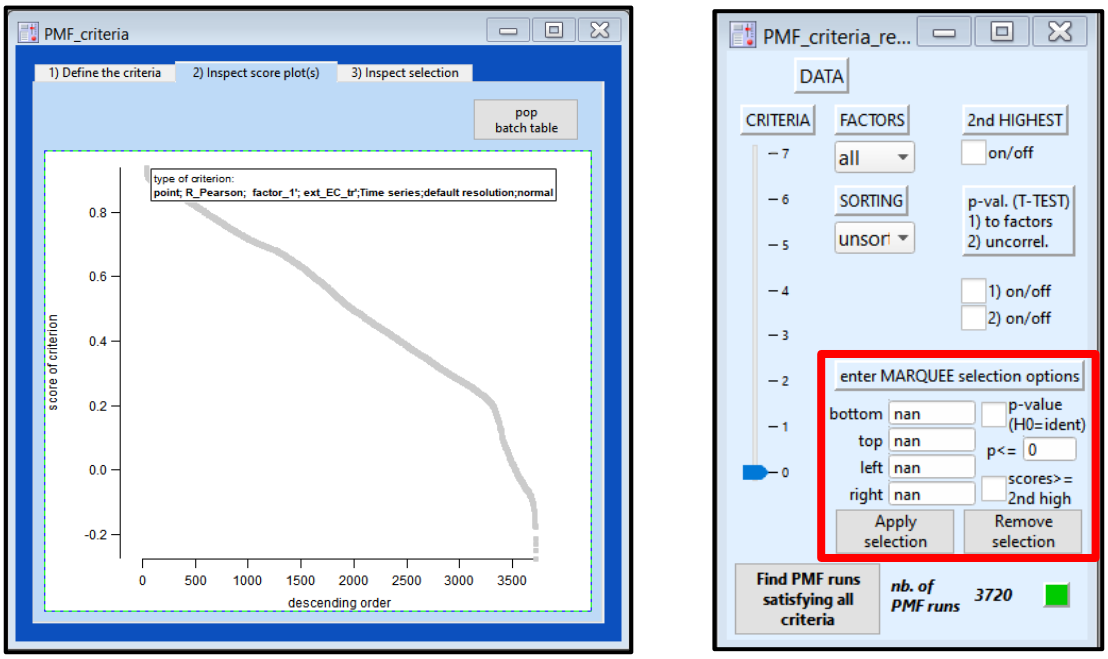


Figure 79 Criterion *f*43 for the profile of factor 5 (gray line) including the slope of the score plot (red curve). Drawing the marquee and select the points around a region of interest highlights them in blue.

The crucial question is where to stop with the selection. This can either be based on absolute values known from *a priori* thresholds or derived from other statistical analyses (see stability of criteria 7.3.9 or BS 9.2). Another option is to rely on statistical measurements. For this, all point criteria have the option of using p-values of a T-test. For average type criteria, the p-values are derived from an ANOVA test across all factors, for correlations it is multiple-testing-corrected p-values (p values to factors). For correlations, the p-value of the correlation (p value uncorrel.) are also available and for the multilinear regression the MLR coefficients are also available. The p-values that should be plotted can be selected via the checkbox. And instead of using a marquee, the p-value can be used for the selection of the good runs. For this, the user needs to define the maximum value below which the runs should be chosen (p-vale (H0=ident)).

Another possibility is to only select solutions for which the highest score compared to all other factors can be achieved. For this, the option “other highest” can be used. If selected, SoFi plots the next lower (if the score of this run is already highest) or the highest (if the score of this run is not the highest). This can also be used for automated selection (scores >= other high).

Furthermore, the shape of the score plot, when sorted can also be investigated. As first metric, strong decreases of the score plot suggest a high degree of mixing within the factors (see Figure 79). Note that the user can select PMF runs to be further investigated using the marquee tool (just presented) or by directly passing the limits on the subpanel (see red rectangle on Figure 79). ‘Left’ and ‘right’ refer to the boundaries on the x-axis. These could be either the index values, e.g. 0 and 100 for left and right, respectively. In case a rolling PMF run is performed, then the user can also pass the time information, e.g. 21/04/2019 and 21/06/2019 for left and right, respectively.

In addition, the user can also filter out score points based on the p-values from significance tests as well as the 2nd classified scores values (see 7.3.9 for more details).

Some impressions on the score plots for various criteria are shown in the next few plots.

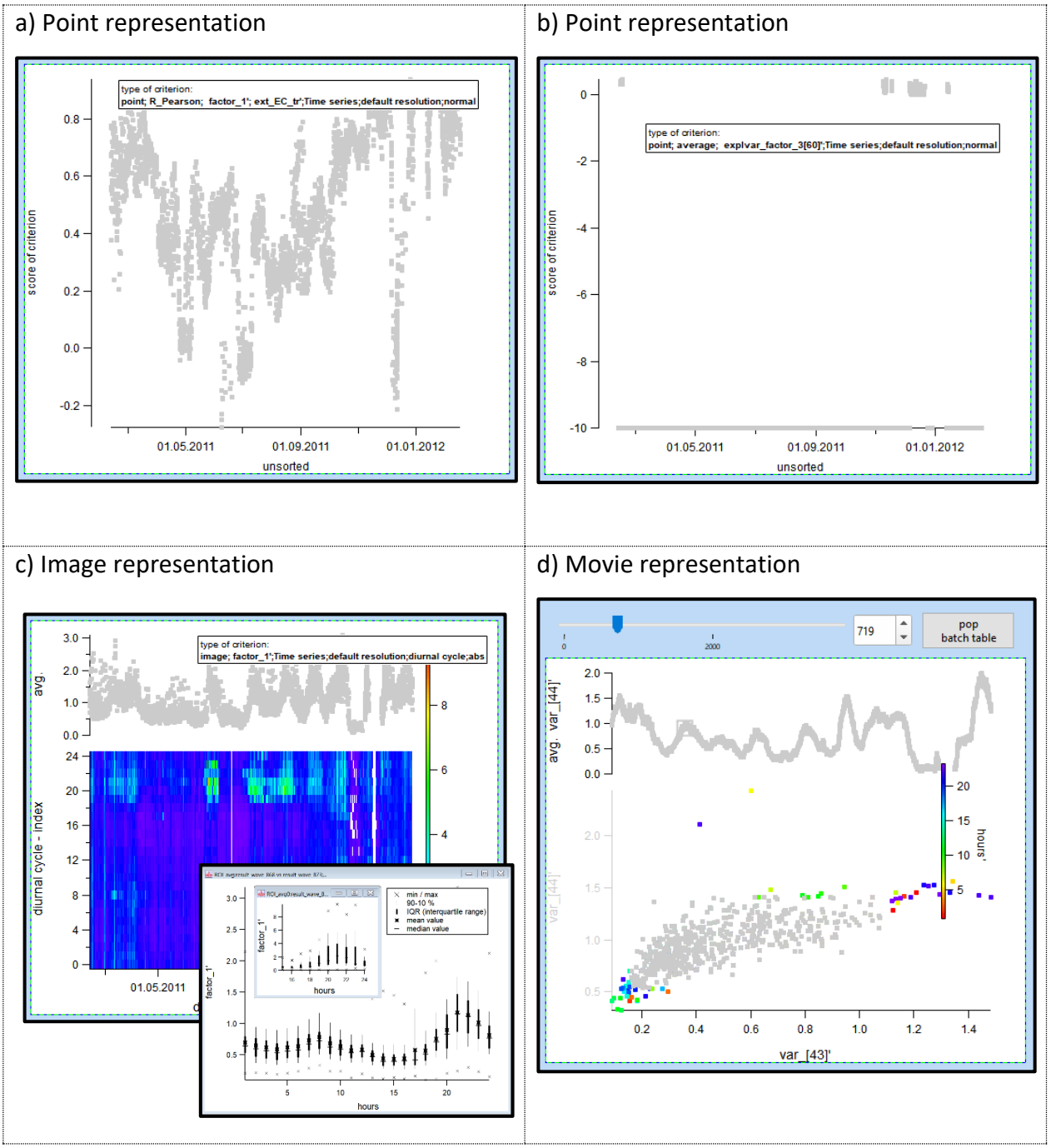


Figure 80 a) R-Pearson between factor 1 and EC_tr. b) Explained variation of m/z 60 in factor 3. Note that for missing values the default is -9.99. c) Diurnal cycle of factor 1 over the PMF runs. The user can marquee regions of interest and pop the corresponding statistics. d) Scatter plot of m/z 44 vs m/z 43 for PMF run 719. The slider or variable menu can be used to roll over the various PMF runs.

Once the regions in the score plots for all criteria are chosen the runs fulfilling all selections are identified and highlighted for further analysis “Find PMF runs fulfilling selection for all criteria” and “Apply selection” (see Figure 81). After pressing on “Apply selection”, the user is also asked whether a quick scan over the selected PMF runs should be performed to estimate the Q-statistics and the amount of missing days, a relevant quantity for rolling PMF (see 6.5) runs (see Figure 82).

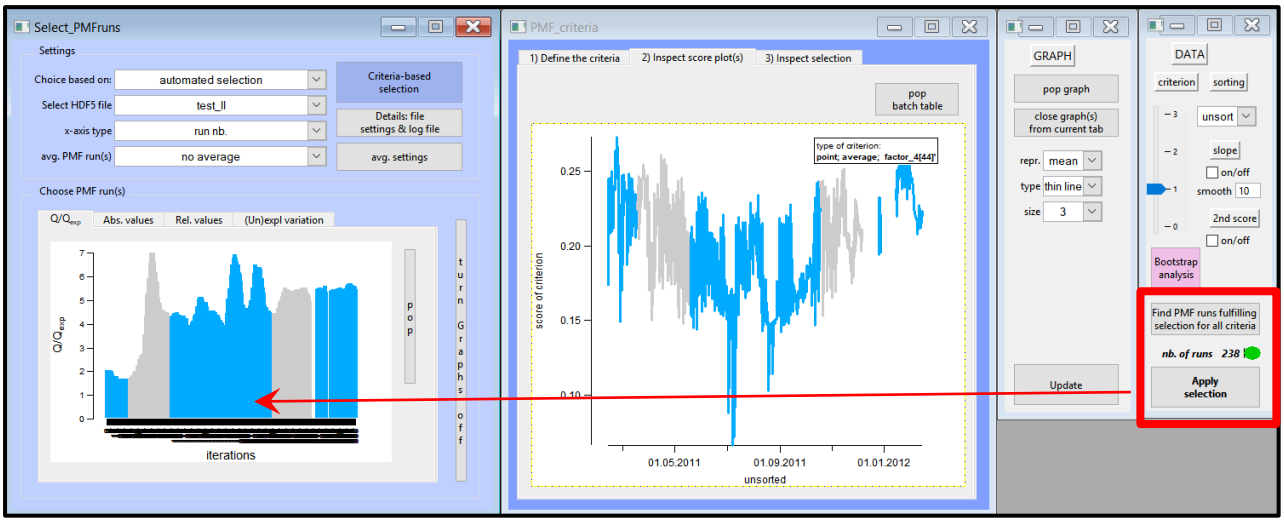


Figure 81 PMF runs fulfilling the selection for all criteria are visualized and selected for further analysis.

7.3.8. Inspect selected PMF runs

The selection (see Figure 81) together with the effectively chosen PMF runs can be visually inspected from the third tab “3) Inspect selection”. The user can compare the score or the score histogram of all PMF runs (gray histogram), with the user-selection (blue histogram) and the effectively chosen PMF runs (green histogram) (see Figure 82). This visualizes which score range is effectively chosen, as the chosen region of PMF runs from the 2nd tab (see Figure 79) sets only the boundaries. However, the effective score limits may be even smaller.

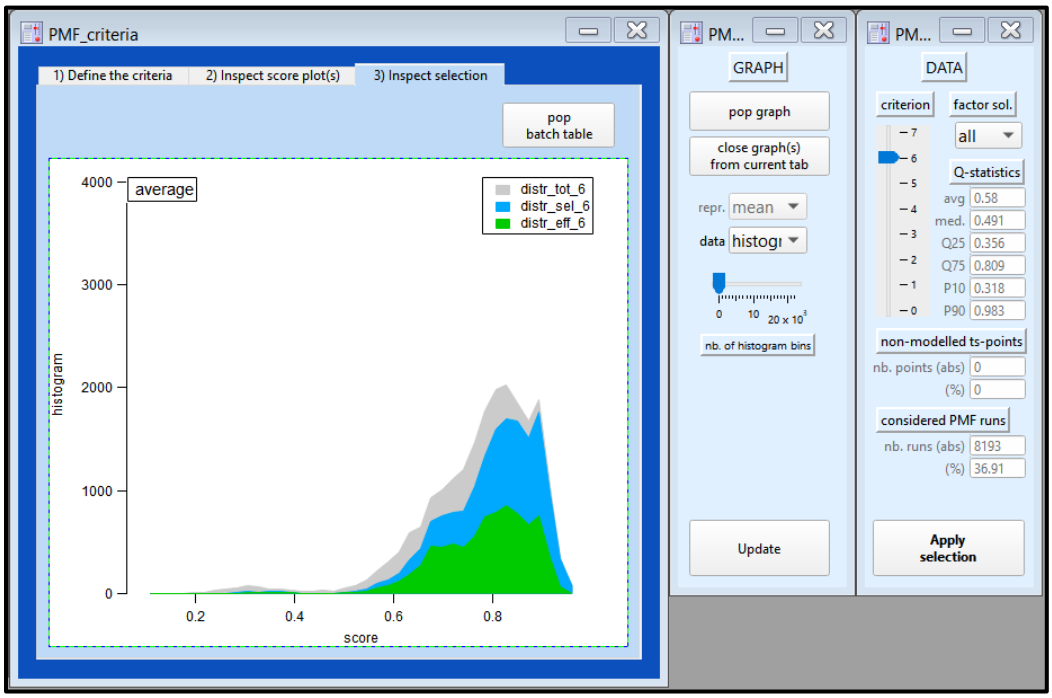


Figure 82 All PMF runs (gray histogram), user-selected PMF runs (blue histogram) and the effective PMF runs (green histogram).

Available options on the adjacent subpanels

"data"	select between normal (score plot), histogram (normal) and histogram(pdf)
"sorting"	in case of normal: scores of criteria can be "unsorted", sorted in "ascending order" or "descending order"
"nb. of histogram bins"	in case of histograms: slider for the number of bins for the histograms
"criterion"	defines which criterion should be investigated
"Q-statistics"	Q statistics on the effective PMF runs after a quick scan over the PMF runs
"non-midelled ts-points"	statistics on the missing time-points
"considered PMF runs"	statistics on the considered PMF runs

7.3.9. Inspect criteria stability

The degree of stability for the score-choice in SoFi can be judged by comparing the highest score of another factor, independent from whether a factor was constrained or not. If the score of another factor is higher than for the current factor (where that score was supposed to be highest), this solution could be regarded as mixed and discarded. In addition, to the score values, its p-

values from significance tests can be inspected too (only when checkbox “comparison with all other factors” is checked, see Figure 66). This option will perform a T-test for comparing two regular waves, e.g. the explained variation or a relative lunch peak or a Chi-Test for a correlation coefficient. If the p value is low, the scores tend to be statistically significant (see Figure 83). Note that for factors with higher score values higher than for that under consideration, an arbitrary p-value of one is passed pointing towards mixed/bad PMF solutions.

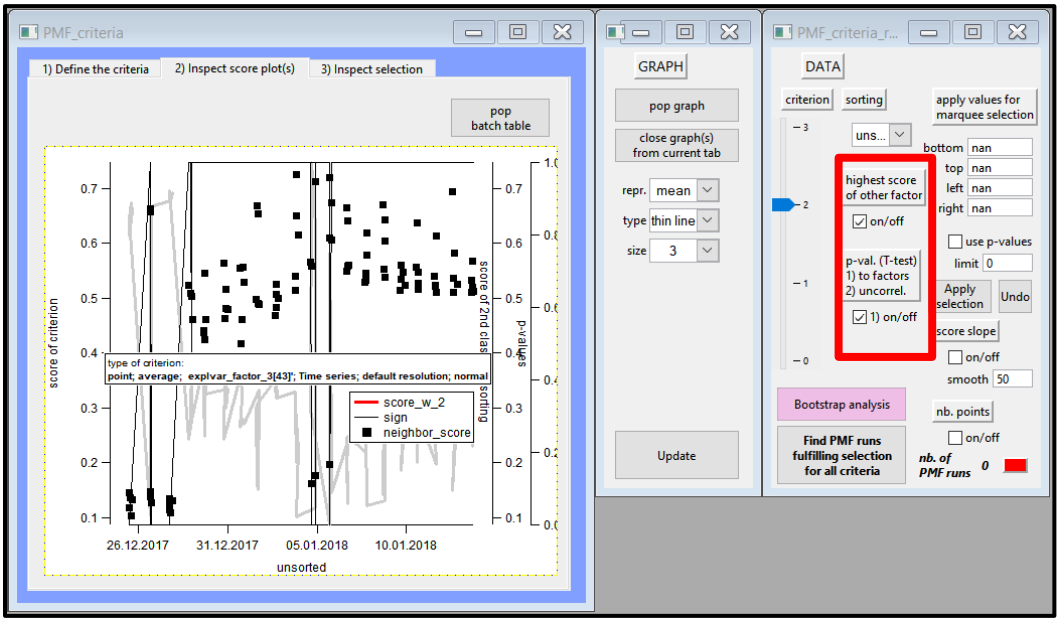


Figure 83 Criterion score explained variation of m/z 60 for factor 3 (gray line) including the score of highest other factor (black) and the p-values from a T-test (black line). Mainly only the gray points at the beginning of the plot are statistically significant and should be considered. For all the other points, the score of the other factors is mostly higher and these PMF runs should be discarded, as they are mixed.

7.3.10. Different number of factors in one HDF file

This option is relevant when inspecting multi-factor solutions with the criteria panel, e.g. PMF runs within one HDF file with e.g. five and six factors. When defining a criterion for the sixth factor, the score plot for this criterion will be empty half of the time, namely for all five-factor runs. This will cause all five-factor PMF runs to be missing in the final solution as they have a non-defined value (NaN) and they would be all discarded when finding the PMF runs that fulfill all criteria selections. Therefore, SoFi passes a zero to these non-defined values allowing the user to select these runs (zero values in Figure 84).

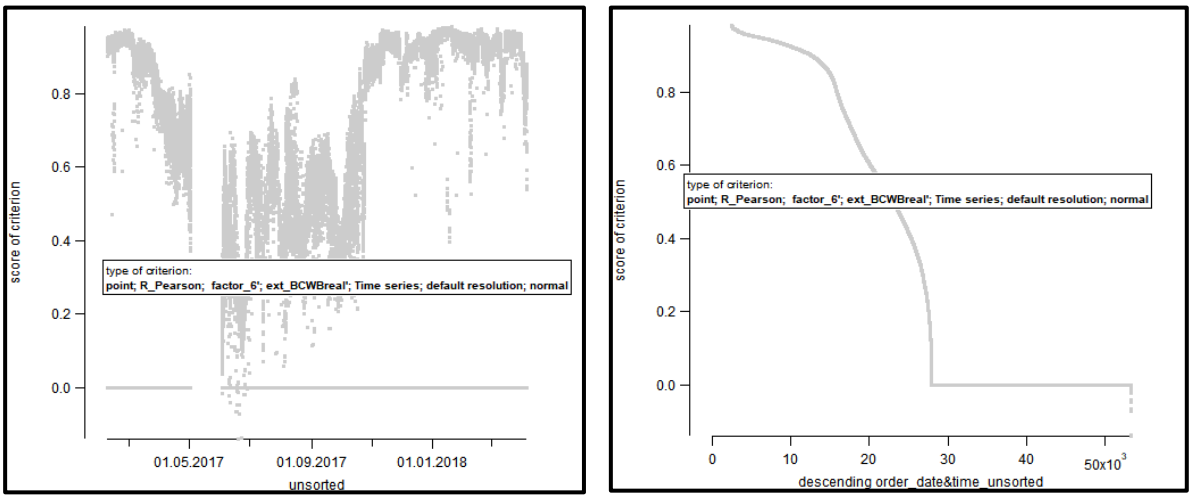
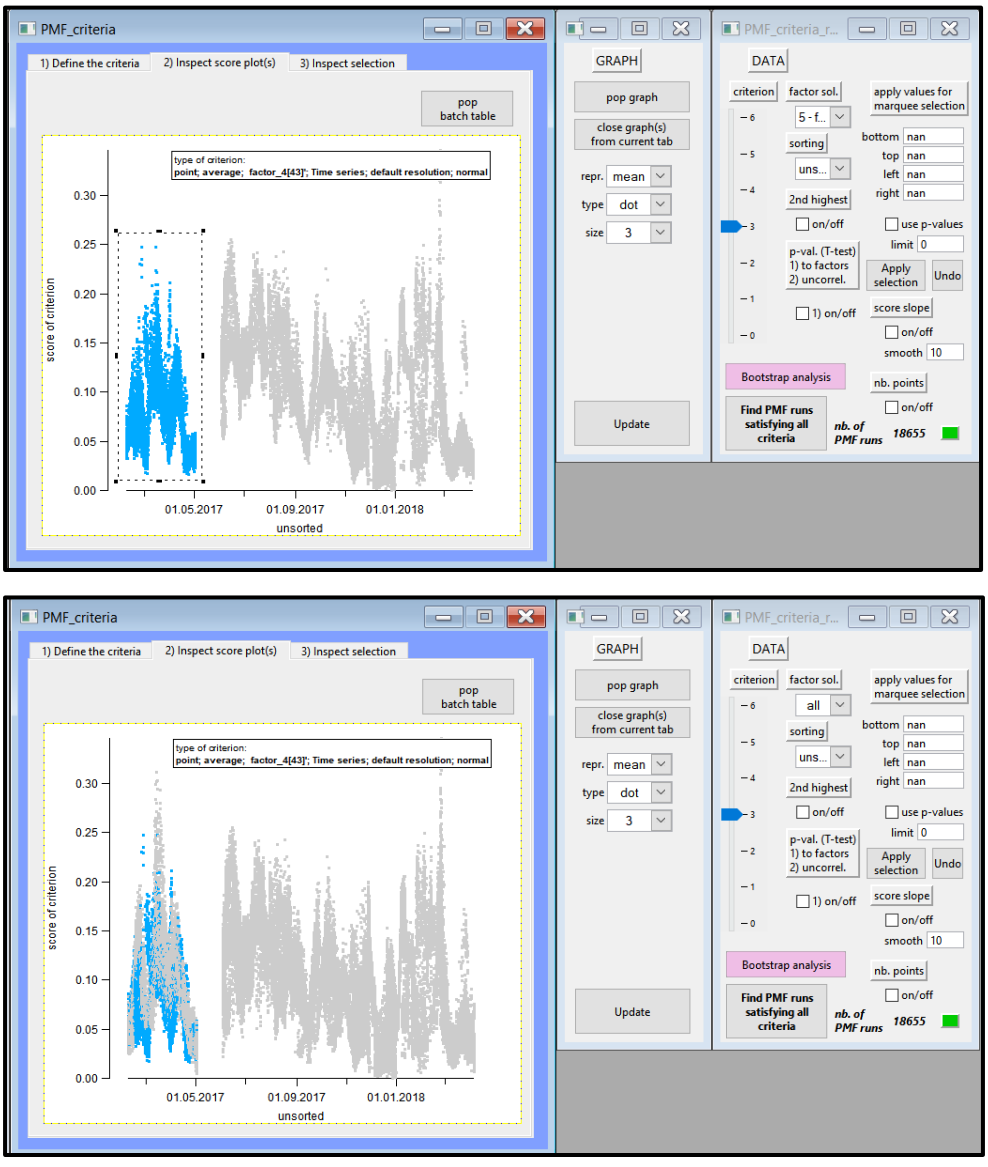


Figure 84 Score plot for the sixth factor in a rolling PMF with five- and six-factor solutions. The score curve is unsorted (left) and sorted in descending order (right). Note half of the score points have a-value of zero. These points represent PMF runs with five factors and therefore no score was evaluated for the sixth factor, as it is inexistent.

Moreover, inspecting specific factor-runs from a multi-factor solution, e.g. the correlation coefficient between a factor and its tracer for 5-factor solutions only out of all PMF runs, is difficult on the entire score plot. The user can specify which factor-solution should be extracted (red rectangle in Figure 85 a)). The resulting score plot is visualized for these runs allowing the user to make selections based on score values for a specific factor-solution, e.g. score values for 5-factor PMF runs only. Finally, all user-selections are stored and the score plot over all PMF runs reports all selections. In the discussed example the blue points in Figure 85 b) are the selected 5-factor PMF runs. They are alternated by the non-selected gray points that represent the 6-factor PMF runs.

SoFi prevents averaging PMF runs with different number of factors. Whenever the user selects with the marquee or from the button “Apply selection” (Figure 79) some PMF runs, then PMF runs with unequal number of factors within the same PMF window or in case of rolling for the same rolling window (6.5), are automatically disabled. (In case a selection of PMF runs in an HDF file with multi-factor solutions leads to overlap period, SoFi will only consider the factor solutions with the highest occurrence.

When averaging PMF runs from various HDF files, then the factors with the same name specification (see 6.3.1) are grouped and averaged together.



a)

b)

Figure 85 a) score plot for 5-factor PMF runs only. The user has selected all PMF-runs within the marquee (blue points). b) score plot for all factor-solutions. The previously selected 5-factor PMF runs are the blue points, whereas for the same period marked in a) the gray points are the non-selected 6-factor PMF runs. All other gray points laying outside of the marquee in a) are 5- and 6-factor PMF runs, respectively.

7.4. File settings and logfile

The user can pop the main settings of the PMF call using the button “log file (actual HDF file)” (Figure 86). Clicking on the “Consult log file” button opens a separate window, where the history of all PMF runs (ME-2 read-out) belonging to the current PMF call (Figure 87) is shown (same as typically printed in the Igor history).

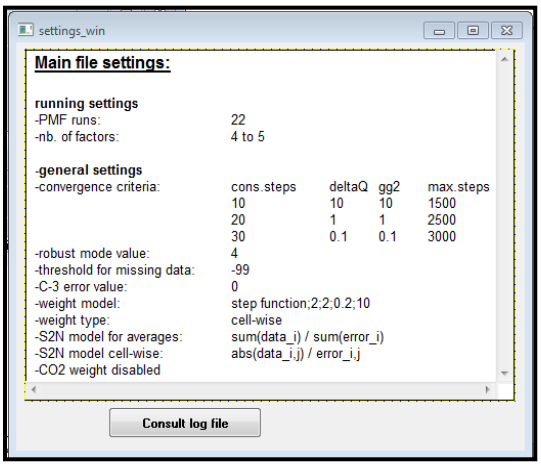


Figure 86 Main file settings of the actual PMF call.

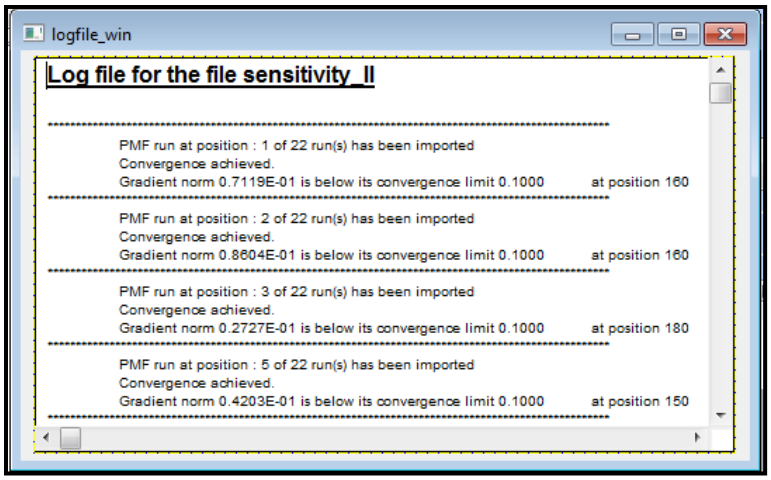


Figure 87 Log file of the current PMF call, here “sensitivity_II”.

7.5. Options on how to import the chosen PMF run(s)

With “II.b Import chosen HDF5 file(s) in IGOR”, only a sub selection of the PMF results were imported, enough to create the overview graphs. For further inspection, all information from the selected PMF run(s) from 7.2 or 7.3 is/are imported into IGOR and evaluated after pressing on “III.b Analyze chosen PMF run(s)” (Figure 61). I.e., for each PMF run the residual matrices, explained variation matrix, relative matrices over profile and time series for the profile

(contribution-weighted) and the time series etc. are calculated. There is also the option to calculate different time resolutions and to average PMF runs.

7.5.1. Averaging

SoFi Pro allows the user to average over several PMF runs. This is controlled via the dropdown menu “avg. PMF run(s).

Available options:

“no average”	only single PMF runs are imported and evaluated
“compute avg. (save single runs”	single PMF runs and average over these runs is computed and stored in SoFi. This option is relevant for a small number of PMF runs only
“compute avg. (only average”	average of PMF runs is computed and stored in SoFi only. This option is relevant when averaging over many runs that would cause an “out of memory” in IGOR when also stored as single PMF runs in IGOR

Note that when averaging over bootstrap runs (see 6.4.1) the repeats per run are first removed.

7.5.2. Averaging (resolution) and data options

The user can control the type in which resolution the result shall be imported. The default case contains the average for “raw” time resolution. “Hourly”, “daily”, “weekly”, “monthly” and “yearly” resolution can be added and will be performed for each single PMF run and also for the average over the single runs. Moreover, the user can control which data should be evaluated in SoFi (see Figure 88). For Figaero TD data, there is an additional option, “filter-based”, where the data is averaged per filter (thermogram) and matched to the provided filter timestamp.

By default, all data is imported, including all statistics.

Note that averaging over lots of PMF runs could cause the averaging process to be very slow. When disabling the checkbox “evaluate quantiles” from the averaging option “data stored of” list, the quantile statistics is not evaluated. The result will contain the average and the standard deviation only and lowers the computational time by a factor 2-3.

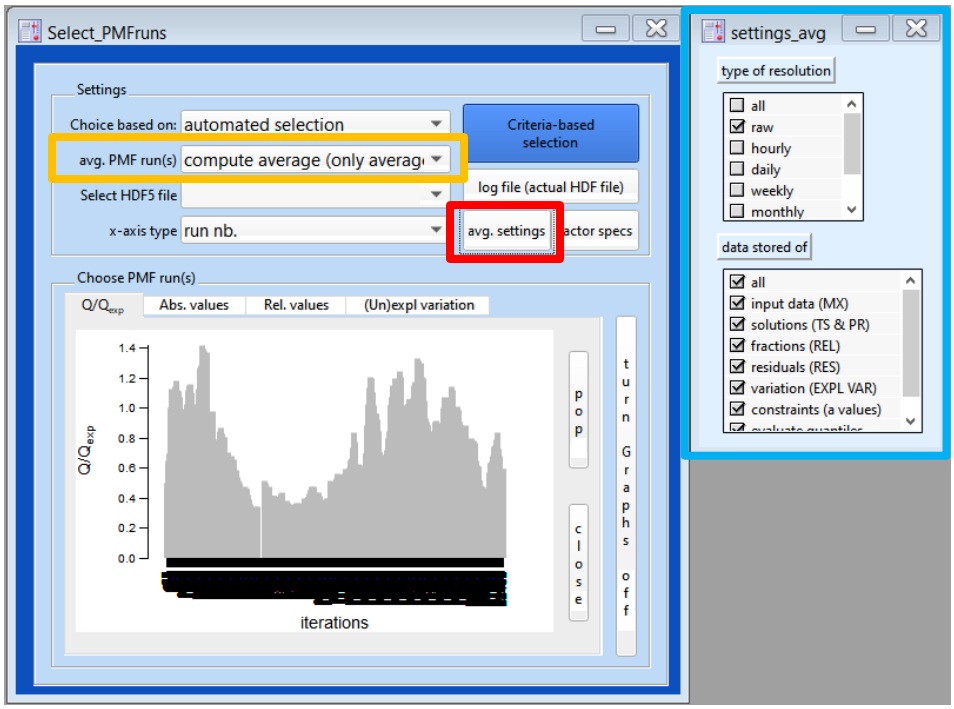


Figure 88 Avg. options (orange rectangle) and avg. subpanel (popped after pressing the button “avg. settings”, red rectangle).

7.5.3. Factor specification

Factor specification (for more details consult 6.3.1) can be accessed for the current HDF file also from the “Select_PMFruns” panel using the button “factor specs”. Moreover, factors can also be added, e.g., SV-OOA and LV-OOA can be combined to OOA on a new row in the “factor specs” wave (Figure 89). After averaging the PMF run(s), these combined factors can be used in the red result panel for further analysis.

Note that the combined factors MUST follow the list of real factors, i.e., they MUST be at the end of the list.

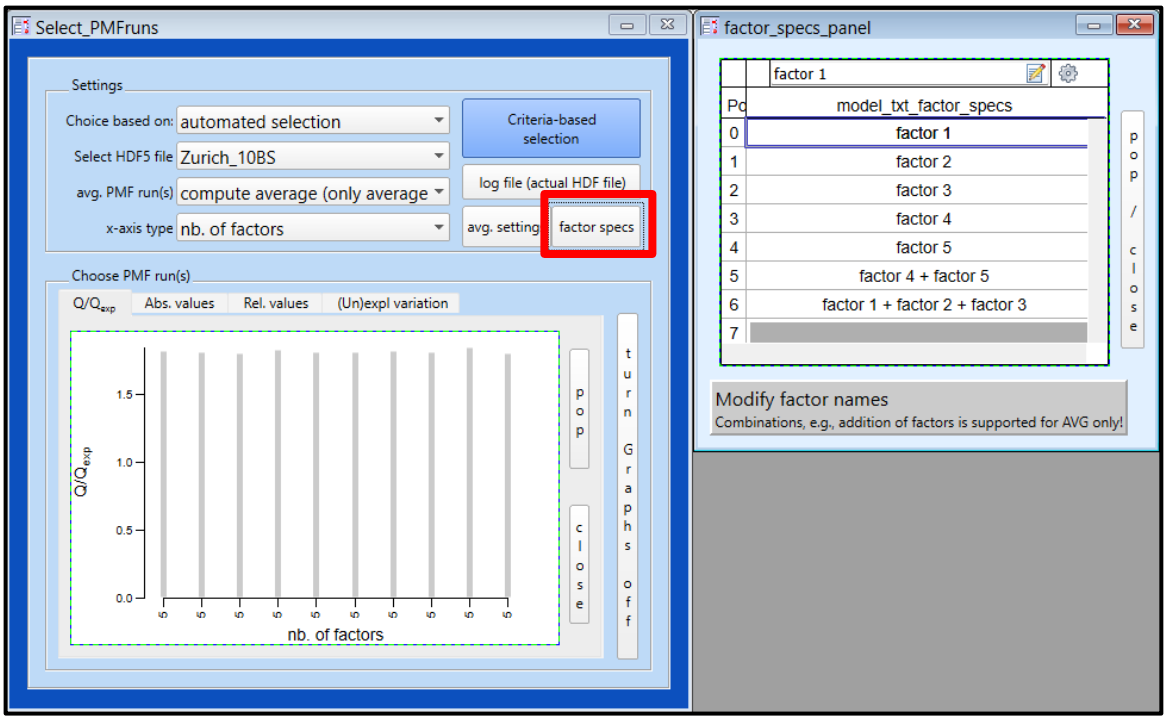


Figure 89 factor specs option for the modification of the factor names. Combination of two and more factors are also possible, in this example e.g. for SOA (factor 4 + factor 5) and POA (factor 1 + factor 2 + factor 3).

7.6. Technical information on the result folder

The next lines describe the structure of the result folders, the two main parameter files that keep all main settings of the current PMF call as well as the information stored in the result folder of a single PMF run. This is just for a better understanding, no waves should be changed here as it can cause fatal errors.

7.6.1. Structure of the result folders

root:SoFi:Results

":Variables"	All variables pertinent to the main result panel. The user must not change, delete or add anything in here.
":Overview"	The waves relevant for the averaged results (see 7.5)
":Final"	This folder contains the final time series and profile matrices as mean and standard deviation, as well as median and quantiles. This data is useful for post-SoFi analysis, e.g. wind and trajectory analysis using ZEFIR (Petit et al., 2017)
":Plot"	plotted waves

“:Graph_macro” macros for the recreation of the plots from the result panel tabs

“[HDF filename]” The result folders are listed here, e.g. sensitivity_I, abbreviated as sens_I hereinafter (see Figure 90)

“:[HDF filename]:Parameter” The parameter files (general data and error matrix, time series, variable information, **parameter files**, convergence criteria for ME-2, name strings, unit string, log file, group information)

“:[HDF filename]:Overview” All representations relevant to the PMF call (see 7.2)

“:[HDF filename]:Solutions” Contains the individual PMF runs of a PMF call, if these runs have been selected for the analysis (see 7.2)

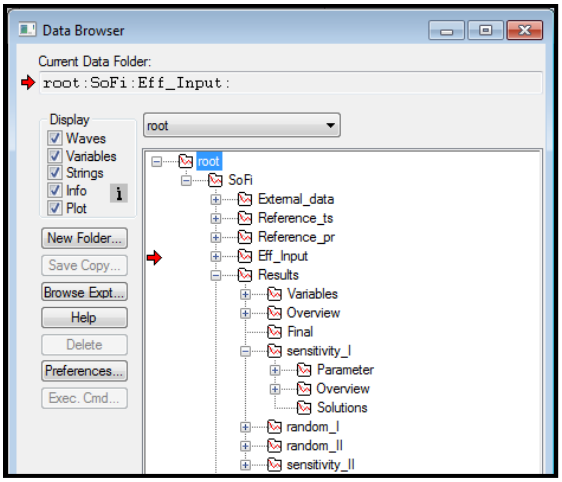


Figure 90 Data browser showing the Result subfolders, in particular for the “sensitivity_I” PMF call.

7.6.2. Parameter files

There are two main parameter files that declare all relevant information of the PMF call, solution_para and batch_file. Both are stored under root:SoFi:Results:[RUNNAME]:Parameter:.

7.6.2.1. **Solution para**

Solution_para is a text wave with 17 positions reserved for various general information:

”solution_para[0]” solver information
”solution_para[1]” normalization during the PMF iteration
”solution_para[2]” robust mode and its threshold value
”solution_para[3]” threshold value for the missing data
”solution_para[4]” matrix scaling factor
”solution_para[5]” C3 value
”solution_para[6]” weight model type including the parameters
”solution_para[7]” weight application type

”solution_para[8]”	CO2-related weight
”solution_para[9]”	mass closure
”solution_para[10]”	relative weighing (C-value)
”solution_para[11]”	fpeak value
”solution_para[12]”	a-value type and a-values of constrained information
”solution_para[13]”	reserved for pulling type
”solution_para[14]”	averaged S/N formula
”solution_para[15]”	cellwise S/N formula
”solution_para[16]”	reserved for statistical error propagation
”solution_para[17]”	rolling mechanism
”solution_para[18]”	multi-time
”solution_para[19]”	instrument type

7.6.2.2. **Batch file**

Batch_file is a numerical matrix containing the information for every PMF run.

batch_file[number of PMF runs][varying, depending on the nb. of constraints]

batch_file[0][0]	nb. of factor
batch_file[0][1]	iteration index for a fixed nb. of factors
batch_file[0][2]	rolling index
batch_file[0][3]	total index, i.e., row dimension of the batch_file
batch_file[0][>4]	information of constraints (over time and/or profile). Each constraint (time series, profile) receives a separate column.

7.6.3. Main information

The user defines which PMF runs should be considered for the analysis and upon pressing on “III.b Analyze chosen PMF runs(s)” the single PMF runs are imported (see 8.1). The result folder has the following name, e.g. for the first PMF run:

“:[HDF filename]:Solutions:run_0”

where the zero stands for the run nb., i.e., the first run for this case. A result folder contains typically the following information:

“curr log”	log file of the current PMF run (log file of ME-2)
“Input data”	Shows the PMF input for the current PMF run. This is relevant for e.g. the bootstrap strategy (6.4) or the rolling mechanism (6.5).
“ini_file”	Text wave containing the instruction file passed to ME-2.
“fpeak_matrix”	Relevant, if fpeak runs have been performed

The next result waves are stored in the corresponding resolution folders, i.e., “raw”, “hourly”, “daily”, “weekly”, “monthly” and/or “yearly”.

“sol_ts/pr”	are the PMF results
“pr/ts_constr”	Contains the information of the a-value constraints. It is a 3D matrix with nb. of factors in *y* and the lower and higher *a*-value limit in *x* for the first two layers. The third layer contains the *a*-value for every constrained model entry.
“REL_pr/ts”	Relative matrix for the time series and the profile matrix. The relative profile matrix is contribution-weighted.
“RES_mx”	residual matrix
“RESW_mx”	scaled residual matrix
“RESABS_mx”	absolute residual matrix
“RESWABS_mx”	scaled absolute residual matrix
“Q_mx”	*Q*-matrix
“EXPL_mx”	Explained and unexplained variation matrix factor with the *x* over time, *y* over the variables and *z* over the nb. of factors.
“RES_ts/pr”	averaged residual matrix over time / profile for all type of RES
“EXPL_tseries”	Explained and unexplained variation over time
“EXPL_profile”	Explained and unexplained variation over profile

8. SoFi – PMF run(s) analysis

8.1. Select a PMF run

The PMF_analyze panel (Figure 91) displays the results that have been calculated when pressing on “III.b Analyze chosen PMF run(s)” (Figure 61), and hence it reacts very fast upon changes. The panel is divided into three main tabs from which the various graphs can be plotted.

“General information”	contains four sections “Select PMF run for exploration”, “Info for all PMF run(s) in adjacent list”, “settings for chosen PMF run…” and “Export PMF solution”
“Overview plots”	Factor time series, profile as well as overview residual plots (absolute residuals, scaled residuals, absolute scaled residuals, Q)
“Detail plots”	Fraction plots, factor plots, scatter plots, residual analysis, residual histogram, correlation matrix, HR-data plots
“Averaged solution”	Plots for evaluating the PMF error, time-dependent factor profiles, time-dependent a-values. Only available for averaged solutions.

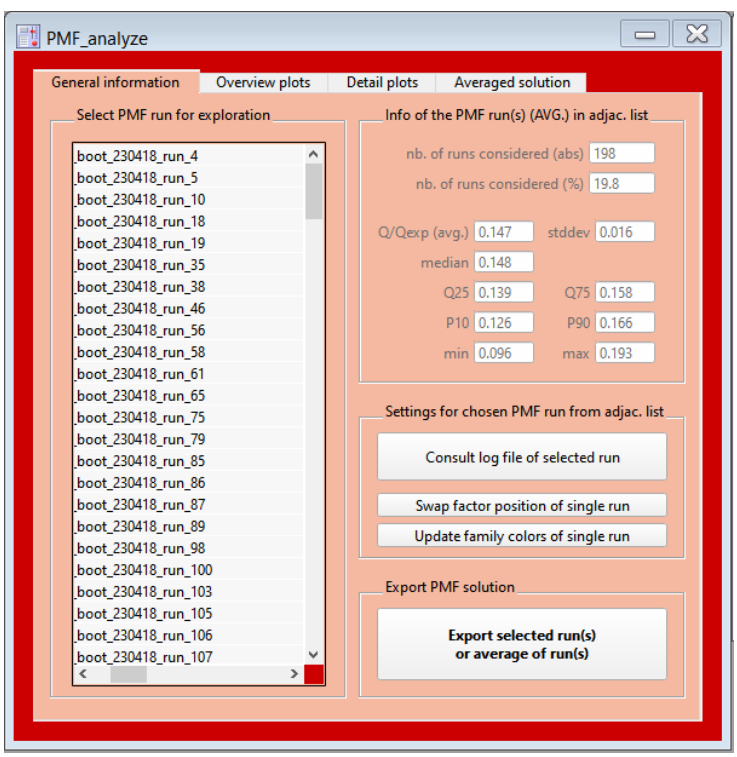


Figure 91 “PMF_analyze” panel – tab “General information”.

8.2. General information

8.2.1. Select PMF run for exploration

The user selects the PMF run (one run at the time) that should be inspected. A drop-down menu with the same list is also available from the subpanel “DATA”. This allows for a rapid scan, in case several PMF runs should be investigated.

When nothing is selected (pressing SHIFT on the selection removes the current selection) and if “compute average” is chosen (see **Fehler! Verweisquelle konnte nicht gefunden werden.**), the averaged solution is presented as a result.

8.2.2. Info for all PMF run(s) in adjacent list

This part shows some overview statistics of the averaged PMF runs, in case “compute average” is chosen (see **Fehler! Verweisquelle konnte nicht gefunden werden.**).

8.2.3. Settings for chosen PMF run from adjacent list

“consult log …” pops the log file written by ME-2 for the current run.

“Swap factor position…” If swapping occurred for unconstrained or loosely constrained factors, the user has the possibility to redefine the order of the factors. Swapping affects only the current/chosen PMF run in IGOR (Figure 92).

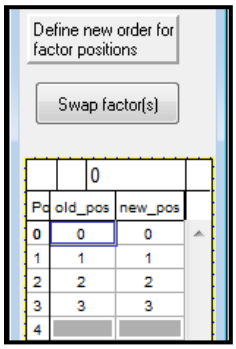


Figure 92 Swapping subpanel showing the old positions (left column) and the new positions (right column).

“Update family colors…” This button only appears if user-defined families are detected by SoFi (see 3.4.9). If the user-defined families are defined or updated after running PMF, the family colors can be updated here. This affects only the current/chosen PMF run in IGOR.

8.2.4. Export PMF solution

The set of environmentally reasonable PMF runs, defined as final PMF solution can be exported to .itx, .txt and .dat files. Depending on whether the user has selected a single solution or the average (no selection) in the overview, the corresponding solution is exported. If the user chose to export an average solution as an .itx file, it is possible to choose between exporting only the mean time series/profiles or the solution with all the statistics. Pressing on the "Export selected or average PMF runs(s)" button opens a dialog, where the user can define the file name and the file format. Profiles and time series are exported in separate files. The time series is exported in the time resolution as selected in the data subwindow of the analyze (red) panel. After pressing continue, a dialog for the file path is opened. An option for exporting the data in a Zefir-friendly format (inclusive averaging) will be added soon.

8.3. Subpanels graph and data

Once the PMF result panel (red panel) is popped two subpanels "graph" and "data" are popped too.

"Graph panel" contains options related to the graph, e.g. mean or median representation, mode of representation (line, point, cross, etc.), size of representation, etc.

"Data panel" contains options related to the type of data, e.g. time series or diurnal cycle

These subpanels (Figure 93), once popped, remain constantly available, even if the user changes tab. Changing tab will load them automatically with the new options for the current graph. The same subpanels were already presented in 5.3.1 for the data, 5.3.2 for time-dependent options and 5.3.3 for the graph options and the reader is referred to those sections to learn more about these subpanels.

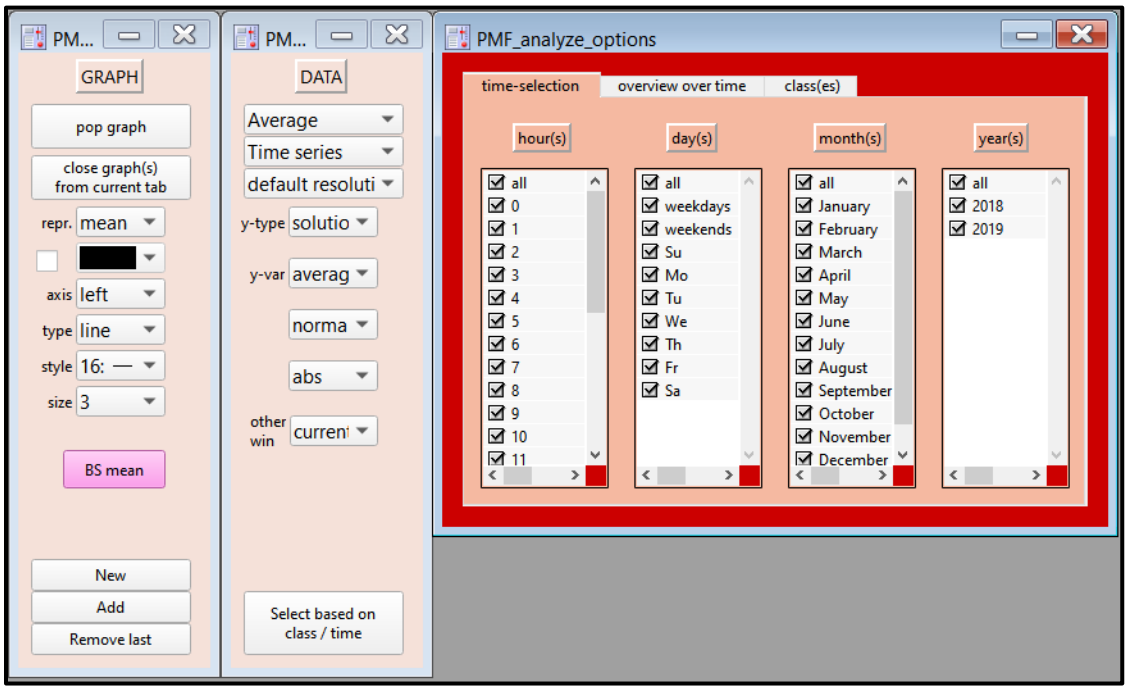


Figure 93 The two subpanels “graph” and “data” loaded with the settings for e.g., a time series standard plot and the time-dependent options for the factor time series (right).

The top dropdown menu of the DATA panel allows to select the PMF run or in case the average was calculated, the average that should be plotted without going back and forth to the “General settings” tab.

Note: Do not forget to press on “close the graphs” after having finished inspecting the results on a particular tab. Otherwise, SoFi accumulates too many waves and becomes slow.

8.4. Overview plots

Factor time series, profiles and the overview residuals can be visualized. SoFi will use the default Igor colors if not defined otherwise (see 3.4.8.2). The residuals graphic contains the residual, absolute residual, scale residual, absolute scale residual and Q residual over time or variables, respectively. Figure 94 shows some possible graphs that are governed from this tab. The spread of the average solution, e.g. standard deviation or quantile statistics, as shown for the residuals shown in Figure 94 can be controlled from the subpanel “GRAPH”.

For constrained information can be visualized by checking “a-values”. E.g. factor profiles were constrained in the solution shown in Figure 94; the shaded area spans the area between lower and higher constrains with the center of the area representing the anchor (constrained profile). A comparison between this gray area and the resulting profile (colored bar) shows in which direction (up or down) the solution was pulled during the in ME-2 run.

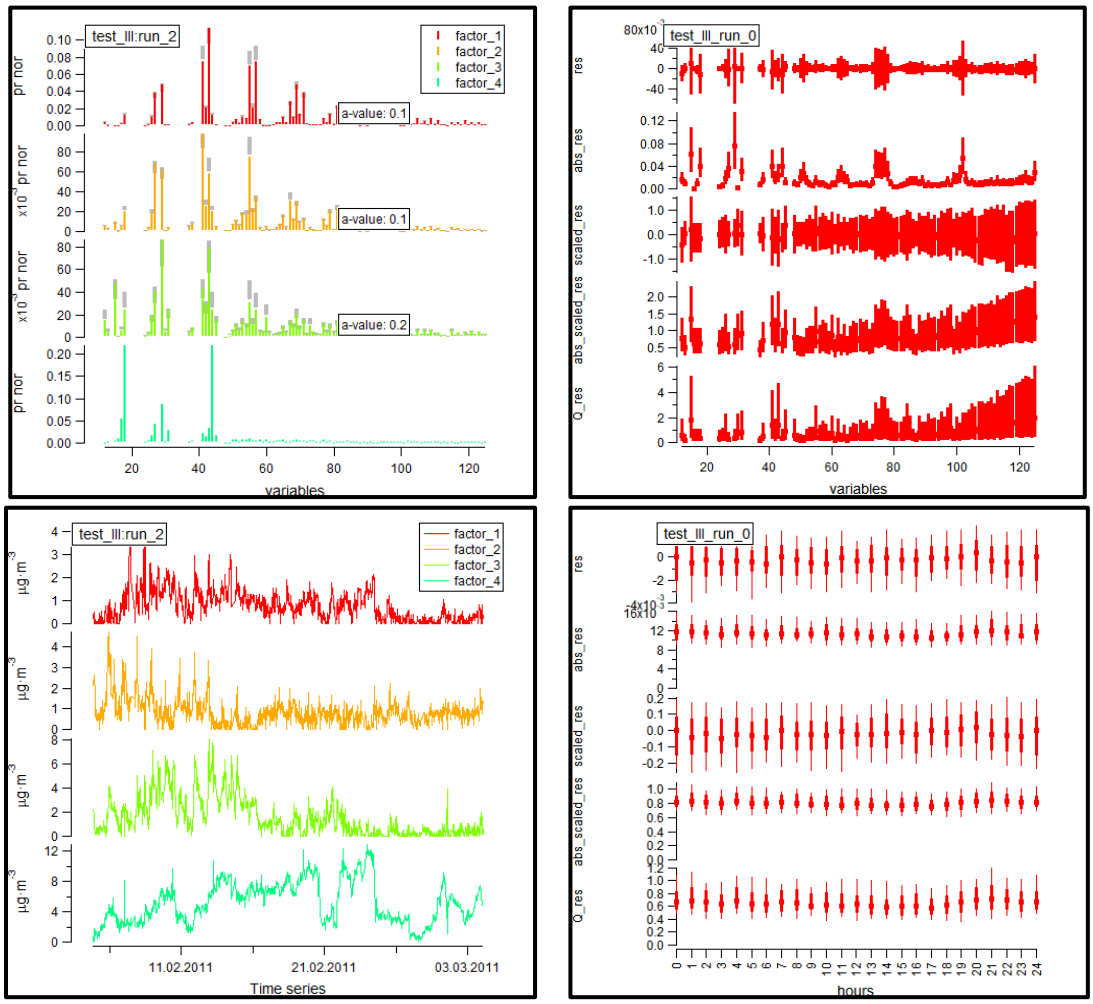


Figure 94 Factor profile (top left) and residuals over the variables (top right). The residuals over the variables are represented by the mean and interquartile range (IQR). Factor contributions (bottom left) and daily cycle residuals (bottom right) represented with the median, the bars for the IQR and the whiskers for the 10th to 90th %.

8.4.1. Time series specific options

If imported into IGOR (see 7.5.2), the user can switch between the various resolutions “raw”, “hourly”, “daily”, “weekly”, “monthly” and “yearly” resolution from the subpanel “DATA”. In addition, temporal information can be represented as cycle plots, i.e., daily, weekly, monthly and yearly cycle.

If the data resolution allows it, all time series plots can either be plotted as “normal” or as cycle plots. The available cycle plots are diurnal cycle, diurnal cycle over a week (i.e. the diurnal cycle is split into weekdays to better understand the average diurnal behavior on different days of the week), weekly cycle, monthly cycle and yearly cycle. For all cycle plots, the first value is repeated at the end. Weekday related plots start with Sunday.

For Figaero TD PMF in default resolution (i.e. the time series is the temperature dimension), these cycle plots are obviously not available. Instead, the thermograms can be averaged to into a single thermogram representing the average volatility behavior of a factor, using the option “desorption temp.”.

For Figaero TD PMF, the option “filter-based resolution” is also available. With this option, the individual thermograms (i.e. each measured filter) are averaged into one point. With the time series now representing the given time series, defined in the input, and not temperature anymore. Here, all the cycle plot from “normal” PMF are available again.

The user can switch between absolute and fractional representation of the time series.

8.4.2. Profile specific options

The first three options available for profiles in the “GRAPH” panel relate to the x-axis representation in the graph. With “x dim continuous”, a continuous x-axis is used, which is useful for data where no mass is included in the x-axis, such as e.g. metal or traditional offline PMF and where certain elements/variables were blacklisted. When checking “x-label”, instead of the numeric representation on the x-axis, SoFi will use the text information from the defined “pr_txt”.

“Factor specs” will add the defined factor names to the corresponding factor. If no factor names were defined, the factors will just be called “factor [nb]”.

The “contr. weighted var” option will add the weighted contribution of each variable on the right axis with black markers (Figure 95). This is especially interesting when the intensity of the different variables of the PMF input differ a lot (e.g., Xact data where there can be an order of magnitude difference in between the different elements). While the sum of all variables in profiles by default to 1 are normalized, the sum of the contr. weighted var. of one variable across all runs equals 1 (i.e.., the sum of the black markers at m/z 44 over all runs is 1).

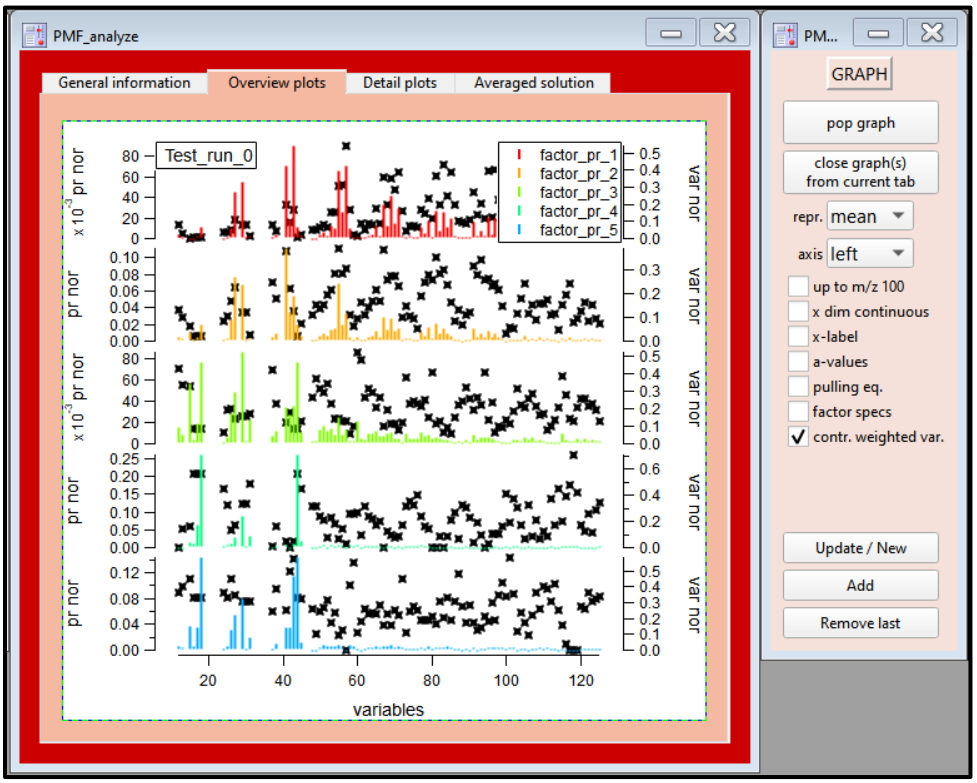


Figure 95 Overview plot with contribution weighted variables.

If user-defined families are specified for the chosen solution, the option “color with families” is available. Similar to the HR family plot (8.5.6.1), the variables are colored based on their family affiliation, however the variables are not stacked (UMR resolution) as in the HR family tab.

8.5. Detail plots

8.5.1. Fraction

Fractional representation in form of bar and pie charts is available over time and profile for contribution and (un)explained variation data. Bar graphs can be binned either over quantiles or over mass (including information of how often a certain mass bin is registered) for mass-dependent graphs. For pie charts over time series, external data can be added to the pie chart (e.g., inorganic species so that the pie chart represents total PM mass). Moreover, the user can plot HR family information factor-wise as pie charts. The user can choose between bar and piechart representation.

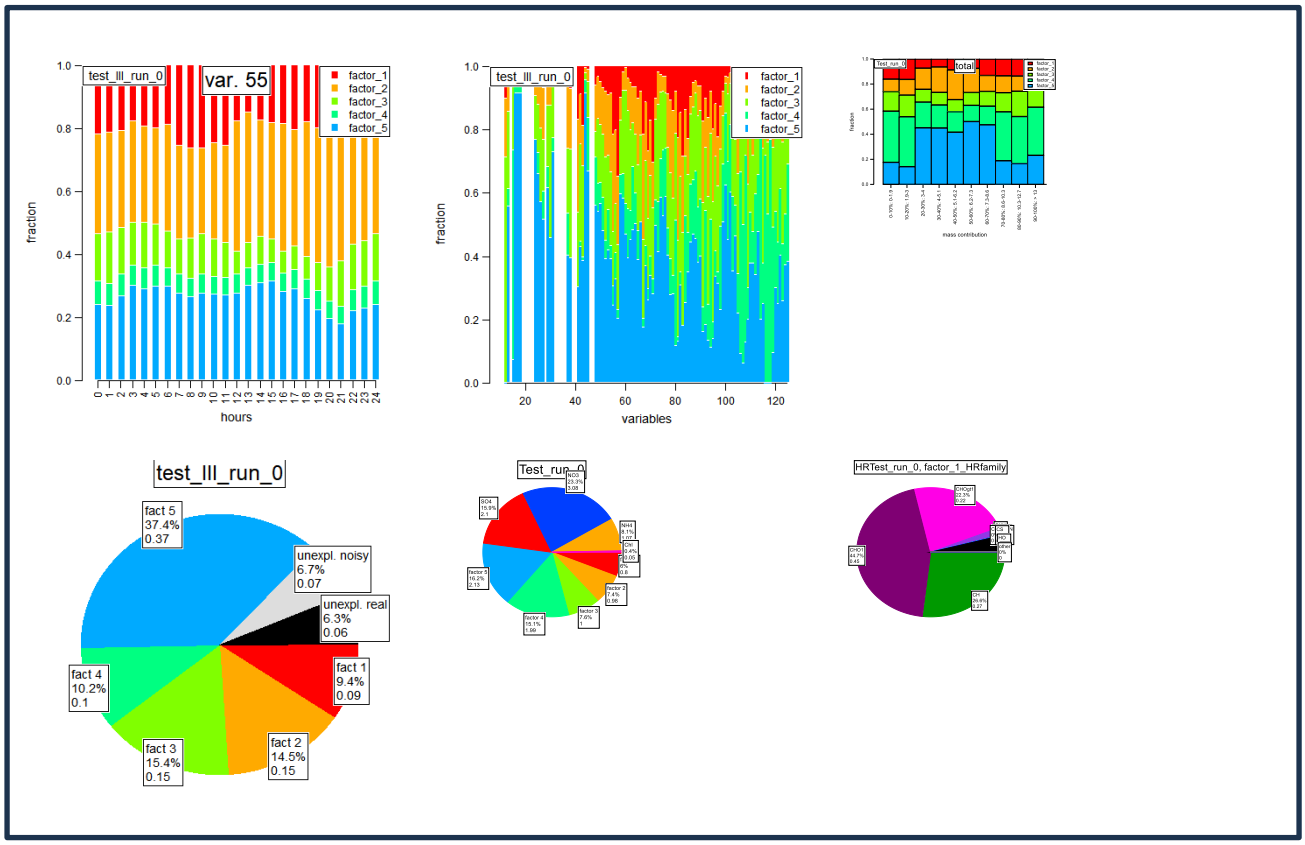


Figure 96 bar plot of daily cycle for variable m/z 55 (top left), contribution-weighted bar plot of profile plot (top middle), mass-dependent bar plot of contribution, binned over quantiles (top right), (un)explained variation as pie chart over a user-defined episode (bottom left), pie chart of total NR-PM, including organic PMF results and inorganic and pie chart of family contributions to factor 1 in a HR-AMS anaylsis.

8.5.2. Standard

The standard tab allows plotting the information over time (time series plot) or over the profile (variable plot) for a multitude of user-selected tracers. Figure 97 shows some possible graphs from the “standard” tab. The user also has the option here to plot the diurnals over a week, i.e. the diurnal of a variable/external/etc. for each day in the week in a single plot.

The user can add several traces form the current PMF run or from any other PMF run, by simply going back to the selection tab, selecting another PMF run and adding to the existing graph.

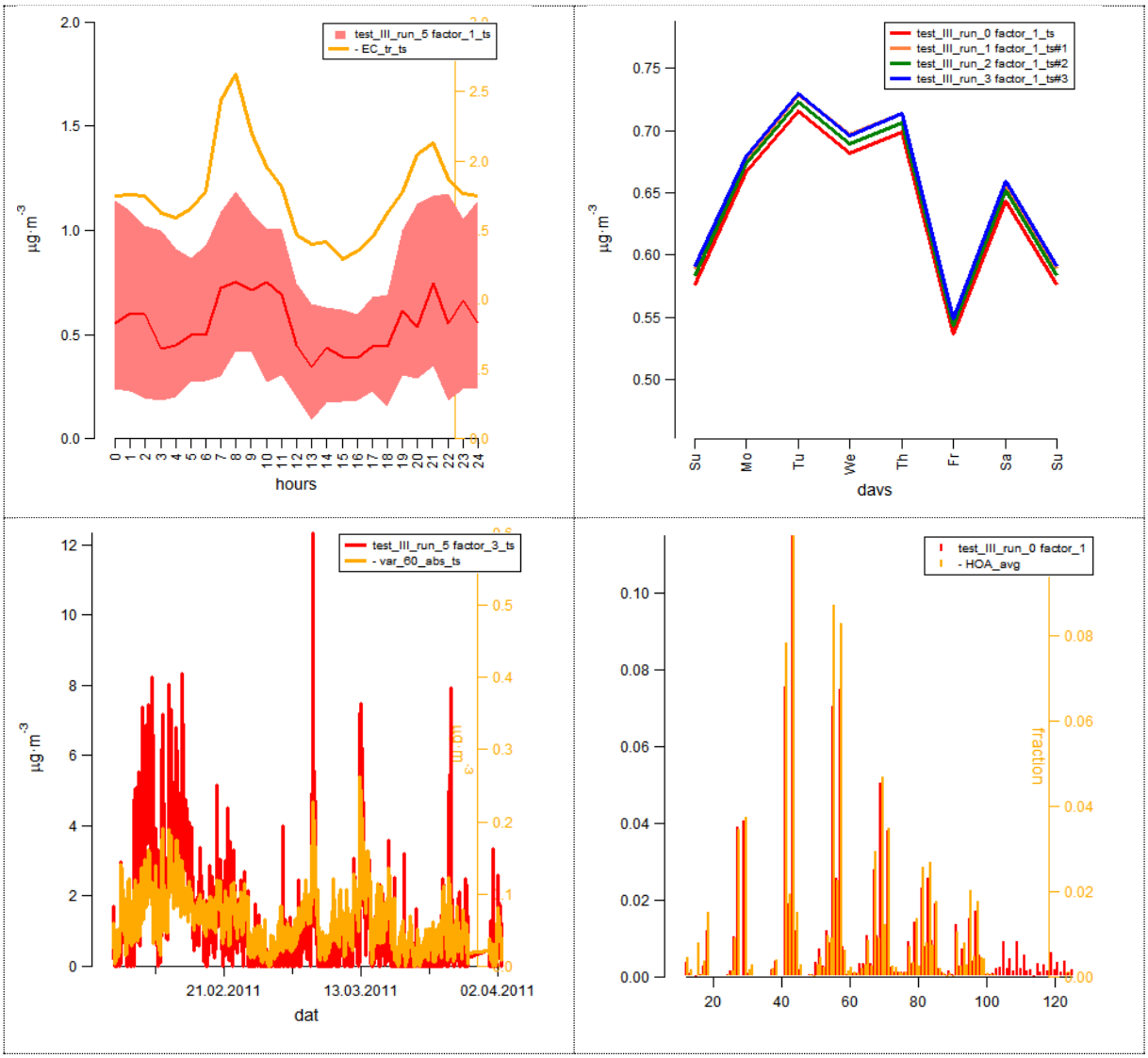


Figure 97 daily cycle of factor 1 (hypothetical traffic) and its tracer EC (top left), weekly cycle of factor 1 from several PMF runs (top right), time series of factor 3 (hypothetical BBOA) with its proxy m/z 60 (bottom left) and profile of factor 1 with an adjacent external traffic profile (bottom right).

Moreover, plots generated from the “Overview plots” or the “Fraction” tab can be loaded into the “Standard” tab. This allows the user to add other traces to the existing graphs for a better visualization and comparison. Choose either “overview plot” or “fraction plot” under the option win temp. for adding the plot from the overview or fraction plot. Pressing either on the tab “Standard” or on the “New” button will finally add the preexisting plot (Figure 98). Select the trace to be added from the available options and then press on the “Add” button for adding this trace. IMPORTANT: the option “win temp. “must be set to current, otherwise the new trace cannot be added (Figure 99).

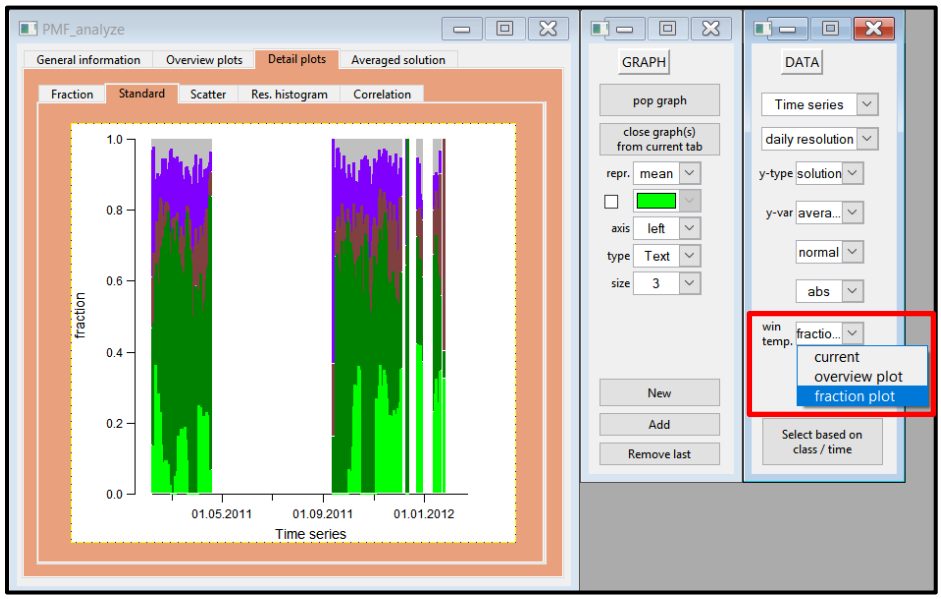


Figure 98 Plot a preexisting plot from the “Overview plots” or “Fraction” tab.

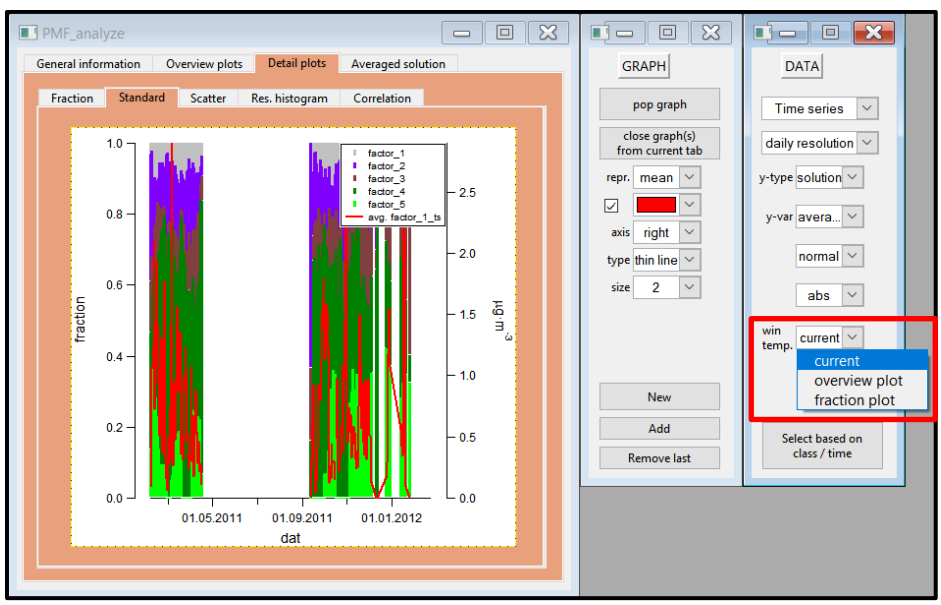


Figure 99 Add a new trace from the “Standard” tab to a preexisting plot.

8.5.3. Scatter

The scatter tab allows plotting the information as a scatter plot and color-coding it based on other information over time (time series plot) or over the variables (variable plot) for a multitude of user-selected tracers. Figure 100 shows some possible graphs from the “scatter” tab.

The user can compare traces between the current PMF run or also between various PMF runs, using the checkbox “hold” for the y-axis, x-axis or color-axis, respectively. This keeps the trace constant to the previously selected PMF run.

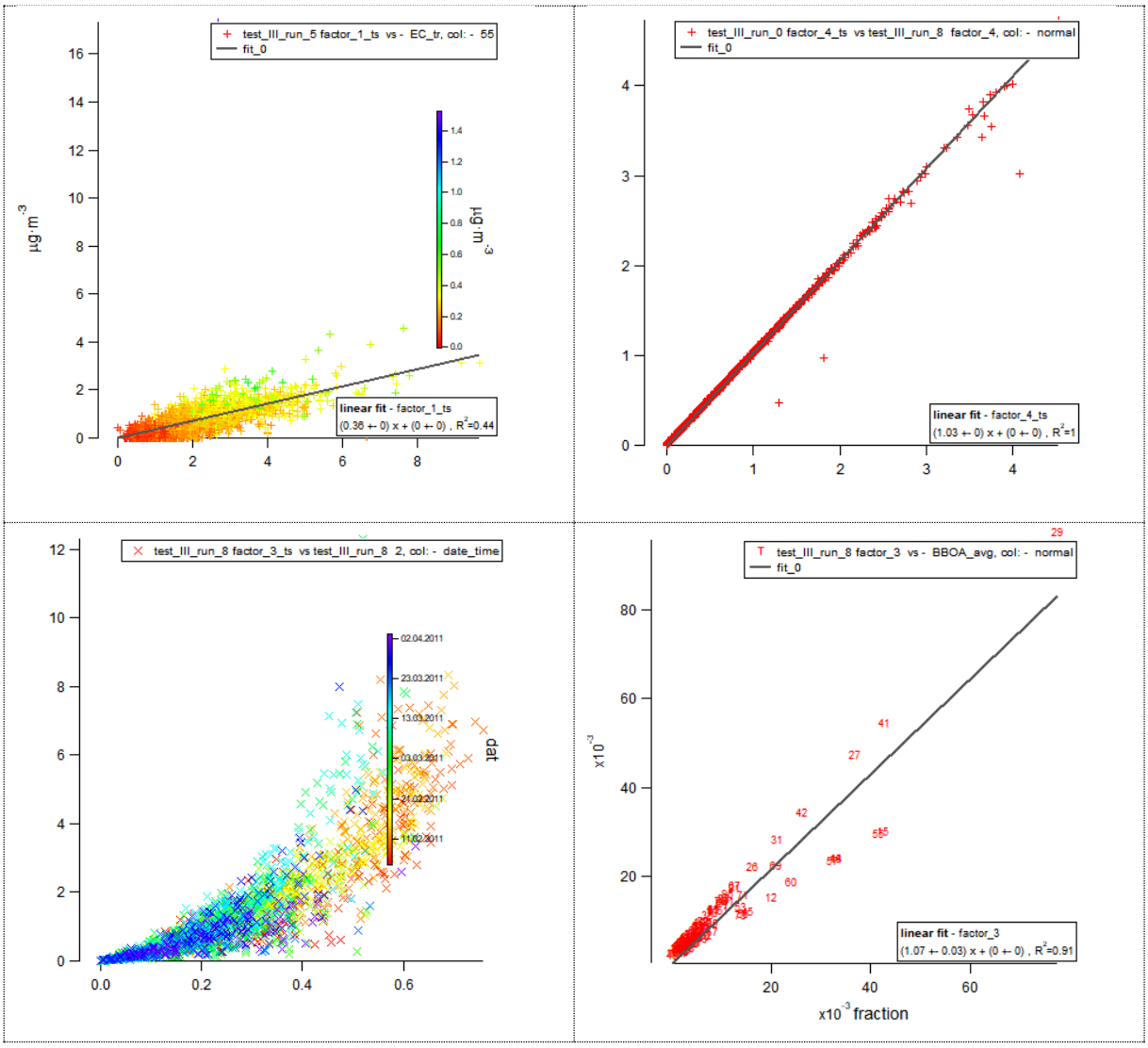


Figure 100 factor 1 (hypothetical traffic) vs. its tracer EC color-coded by its proxy m/z 55 (top left), factor 4 vs. factor 4 from another PMF (top right), factor 3 (hypothetical BBOA) vs. the explained variation of its proxy m/z 60 color-coded by time (bottom left) and factor 3 (hypothetical BBOA) vs. an external BBOA profile (bottom right).

8.5.3.1. **Option model subtraction**

SoFi Pro offers the possibility to compare a specific modeled outcome to the measured quantity that the model should ideally reproduce. In the case of ACSM data with primary and secondary organic aerosol sources the experimentalist can e.g. remove the primary contribution of specific variables and compare the modeled fractions of the secondary sources (typically the entries of the factor profile) to the secondary measured contribution. Consult Canonaco et al., (2015) for more details on its application.

<u>Example</u>

The fraction of m/z 44 and 43, i.e., *f*44 and *f*43 of the modeled secondary factors should be compared to the contribution of measured *f*44 and *f*43 from the secondaries only.

y-type and x-type must read "model subtraction". y-var and x-var are *m/z* 44 and *m/z* 43, respectively and the option "fraction" should be checked. The table on the right on Figure 101 contains the group names for each factor, here POA and SOA and "actual type" is the name of the group of factors that will be used for the comparison. The contribution of all factors that do not belong to the group "actual type" is automatically removed from the total measured contribution on the variables y-type and x-type.

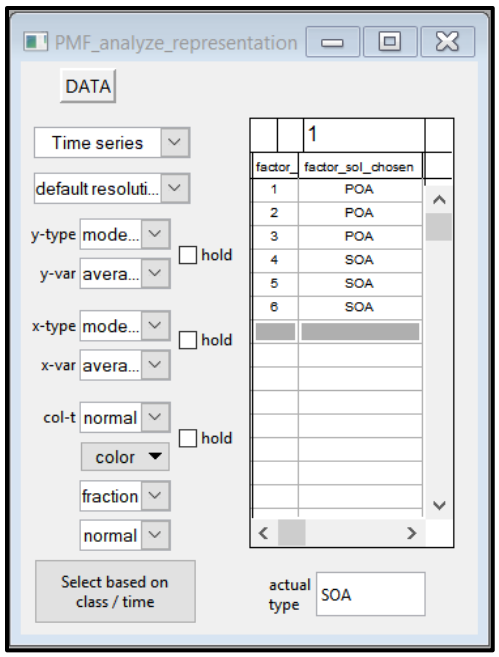


Figure 101 Model subtraction panel with the first three factors belonging to the group "POA" and the last three to the group "SOA". The factors being further analyzed are those defined under "actual type", i.e., SOA.

Given the linear relationship in PMF, the variability of the measured points (red points in Figure 102) is supposed to be captured by the linear combination of the modeled factor profiles (green points in Figure 102). This implies that the measured points that are properly reproduced by the PMF model are those that are on a line if e.g. two factor profiles are modeled. This is the case here, as light and dark green points represent two SOA factors form a rolling PMF run (see 6.5 for details on the rolling strategy).

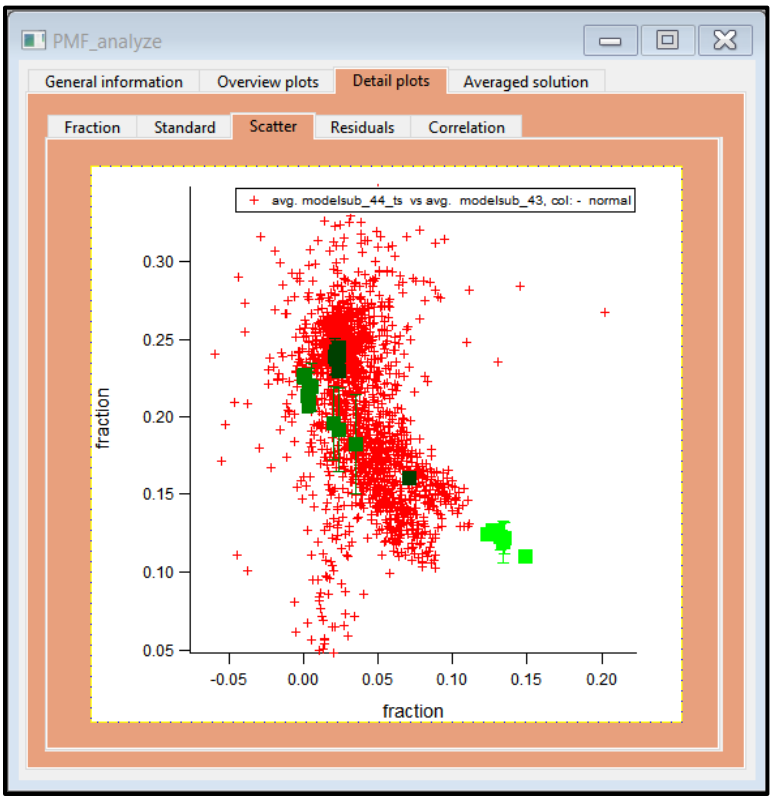


Figure 102 Outcome of the model subtraction option for *f*44 vs *f*43. The POA contribution of *f*44 and *f*43 has been removed leading to the red points. The light and dark green points are the fraction of the SOA factor profiles.

Note that when performing "model subtraction" plots on absolute values, the modeled fraction of the factor profiles cannot be compared to the measured contribution. Hence, only the measured variables after having subtracted the contribution from the factors that should not be further inspected are plotted.

Moreover, inspecting the shape of contribution subtracted measured points (red points in Figure 102) might also reveal information on how many factors should be used for the PMF model. If for example these red points lie on a line, only two factors are enough to capture their variability. For the case where the shape of these points is triangular, then three points are required. If the shape of the points is more spherical-like, then only one factor suffices.

8.5.3.2. **Option multilinear regression**

SoFi Pro offers the option to perform a multilinear regression between externals and PMF factors or externals and PMF factors plus externals. This could be used e.g. to perform a multilinear regression of eBC versus PMF combustion factors like traffic and biomass burning. y-type and x-type must read "multilinear regression". For the y-type, SoFi lists all externals. For x-type the user has to select (check) the factors (and externals) that should be used for the multilinear regression.

For easier navigation, the user can select in the dropdown menu for the x-var either only solutions or solutions and externals. SoFi supports multilinear regressions with up to 10 variables.

SoFi presents the result as a scatter plot where the external is plotted versus the sum of the fitting parameter times the factors. The fitting parameters are displayed in the bottom right corner. The multilinear fit can either be forced through zero or can be allowed to have an intercept (checkbox "MLR through zero"), in which case this one is also listed together with the fitting parameters.

8.5.4. Residuals

This tab hosts features relevant to the residual matrix. On one hand, the entire (scaled) residual matrix is accessible. On the other hand, residual histograms can be drawn for the total as well as for single variables or classes, if previously defined. In case of the C-value approach, the histograms of the different instrument classes can be inspected and compared here.

8.5.4.1. **Entire residual matrix**

The entire (scaled) residual matrix can be visually inspected. The adjacent slider that controls the color-code of the matrix allows to dynamically inspect various regions of the (scaled) residual matrix and to verify that the entries of the matrix are random, i.e., unstructured over time and over the variables throughout all ranges (see Figure 103). Furthermore, the explained variation can be visually inspected here as well, following the same principle as for the residual matrix.

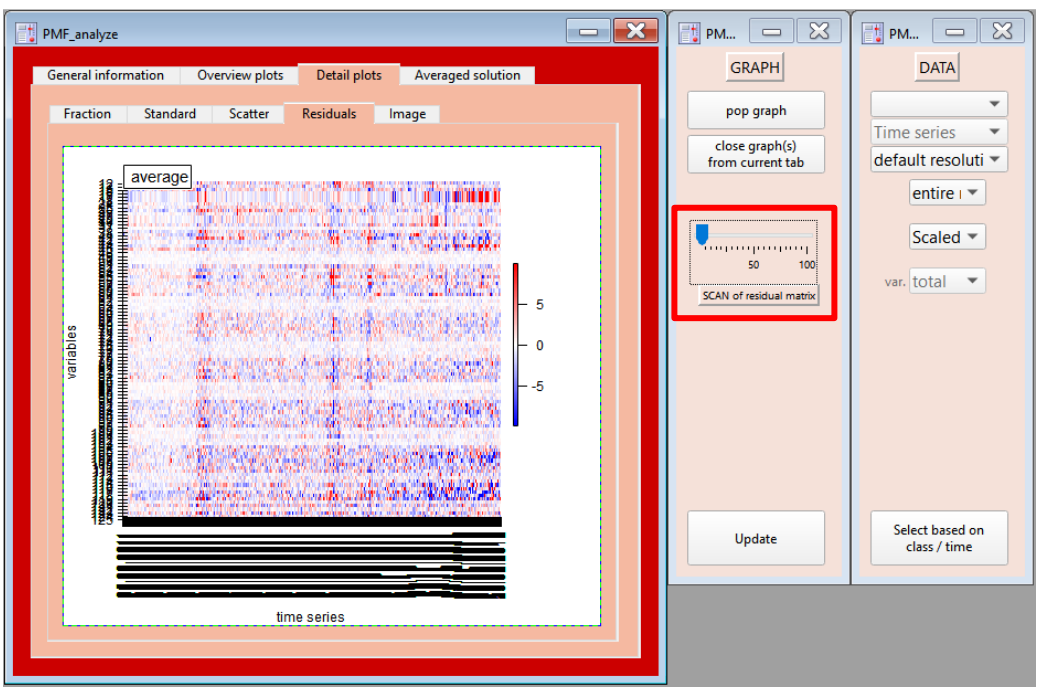


Figure 103 Subsection of the scaled residual matrix color-coded between -0.4 and +0.4, based on the user-defined slider "SCAN of residual matrix" value on the subpanel "GRAPH".

8.5.4.2. **Residual histogram**

Residual histograms are an easy and fast method for investigating whether the PMF result contains some systematic under- or overestimation (see Figure 104 for some examples). Ideally, the PMF result doesn't contain any systematic scaled residuals, i.e., the scaled residuals considering all input cells scatter around zero in a noisy way. This leads to a unimodal distribution, ideally Gaussian-shaped (the internally performed Chi squared statistical test tells the user the likelihood to the Gaussian-curve).

Following histograms are accessible

"residual"	the residual only is taken. This reveals how good the data is modeled. The shape is supposed to be unimodal. It is not expected to be Gaussian-like, as PMF doesn't directly fit the residual, but the scaled residual (see below)
"abs. residual"	absolute residual histogram. The histogram will be on the positive quadrant only.
"scaled residual"	scaled residual histogram. This quantity reveals how well the PMF model fitted the data. Ideally, the probability density function is unimodal and Gaussian-like. In addition, according to Paatero and Hopke (Paatero and Hopke, 2003), the scaled residual histogram is supposed to mainly scatter between +/- 3.
"abs. scaled residual"	absolute scaled residual histogram. The histogram will be on the positive quadrant only.
"Q residual"	Q histogram. The histogram will be on the positive quadrant only.

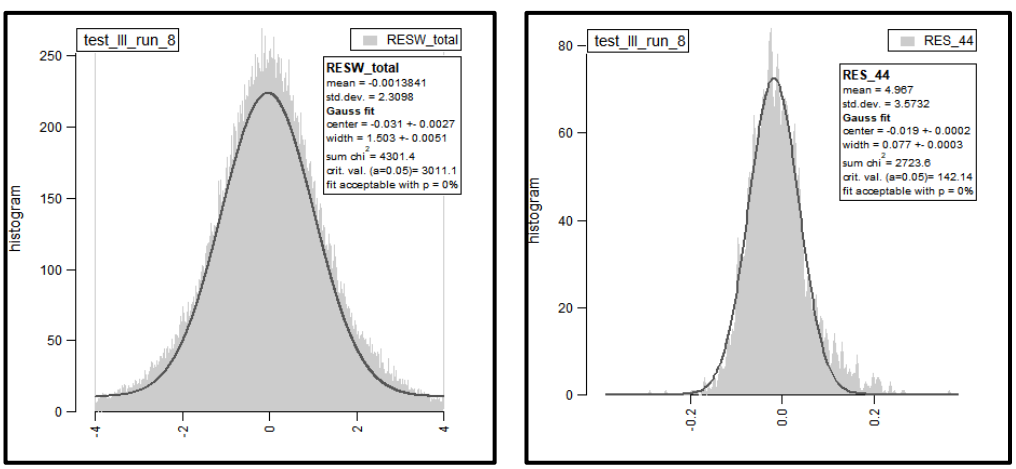


Figure 104 Histogram for total scaled residual (left) and residual of variable m/z 44. Note the extreme high bars at +/- 4 due to the robust mode. These lines are not considered for the Gaussian fit and its subsequent statistical significance test.

If the actual residual varies from this shape in that it is bi- or multimodal or showing some high spikes at negative or positive sides, further investigation of the PMF model is required. It might be that the number of factors is too high/low, if constraints were applied that they don't match the natural variability of the datasets, etc.

Relative weighting of classes (C-value approach)

In case the analyst has performed relative reweighting of the errors of certain classes, their (scaled) residual histograms are plotted separately and proper statistical tests can be performed to judge, whether it is likely that the (scaled) residuals have the same underlying distribution.

The following statistical analysis is directly accessible from SoFi:

"ANOVA (equal variance)" a one-way ANOVA test with the assumption of equal variance and the underlying normal distribution for all (scaled) residual sampling distribution (classical ANOVA testing with the F distribution or T distribution for two classes). In addition, Bartlett's test for the equal variance is also computed and the result represented in the adjacent text box.

"ANOVA (unequal variance)" a one-way ANOVA test with the assumption of the underlying normal distribution for all (scaled) residual sampling distribution (Welch test)

"ANOVA (non-parametric)" a non-parametric or rank one-way ANOVA test is performed. (Dunn-Holland-Wolfe test is performed that accounts for various classes with a different number of class members)

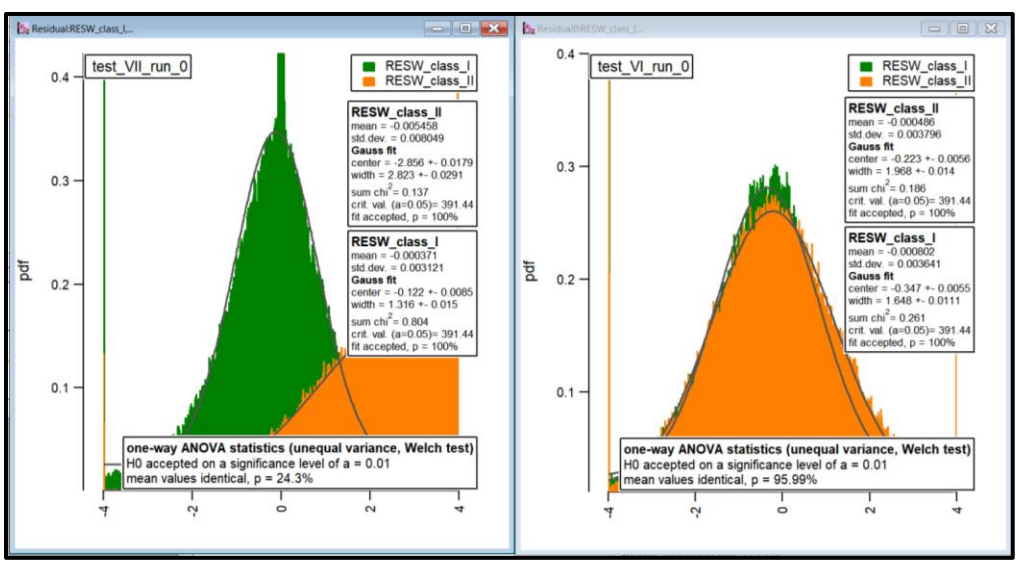


Figure 105 dataset split in two classes and errors unweighted (left), the scaled residual of the two classes is optimized using the C-value approach (equilibrate group scaled RES)

Based on the reported p values for each statistical test the analyst can judge how likely the test hypothesis H_0 is true, i.e., the mean values of the (scaled) residual distributions for each class arises from the same underlying distribution.

An example is given in Figure 105, where a dataset is artificially split into two classes (class_I in green and class_II in orange) and the errors of class_II were manually increased by 3 orders of magnitude. The left panel shows the PMF run where class_II is underestimated (errors are 1000 times larger than for the variables of class_I), as their scaled residuals are almost all positive. The green and the orange distribution are likely to occur from the same underlying normal distribution with ~24% probability (from a statistical point of view still satisfactory, as typical α-values are in the range 5%, 2% or 1%.). However, applying the dynamic equilibrate group scaled RES (C-value approach) made the scaled residuals for both classes even more similar with a p value of ~96%, i.e., both distributions arise from the same underlying normal distribution with 96% probability.

8.5.5. Image

From the drop-down menu on the first window next to the analysis (red) panel, the user can select the type of the image plot. Currently, SoFi offers correlation, wind analysis and hierarchical clustering image plots.

8.5.5.1. **Correlations**

Factor time series as well as factor profiles can be correlated to external data and/or variables and within each other. The following options are available:

“R_Pearson (squared)”	R-Pearson value useful for linear relationships
“R_Spearman (squared)”	also rank correlation, useful for any type of relationship
“Kendall tau (squared)”	rank correlation coefficient that analyzes concordant and discordant pairs
“R_uncentr (squared)”	cosine similarity
“Fisher transformation”	Fisher z-transformation of a Pearson correlation coefficient

For all these options the representation can be either in image, in a table or in a combined format (Figure 106). For the Pearson and Spearman correlation coefficients an additional table contains the p-values as well, to better judge on the significance level of the correlation.

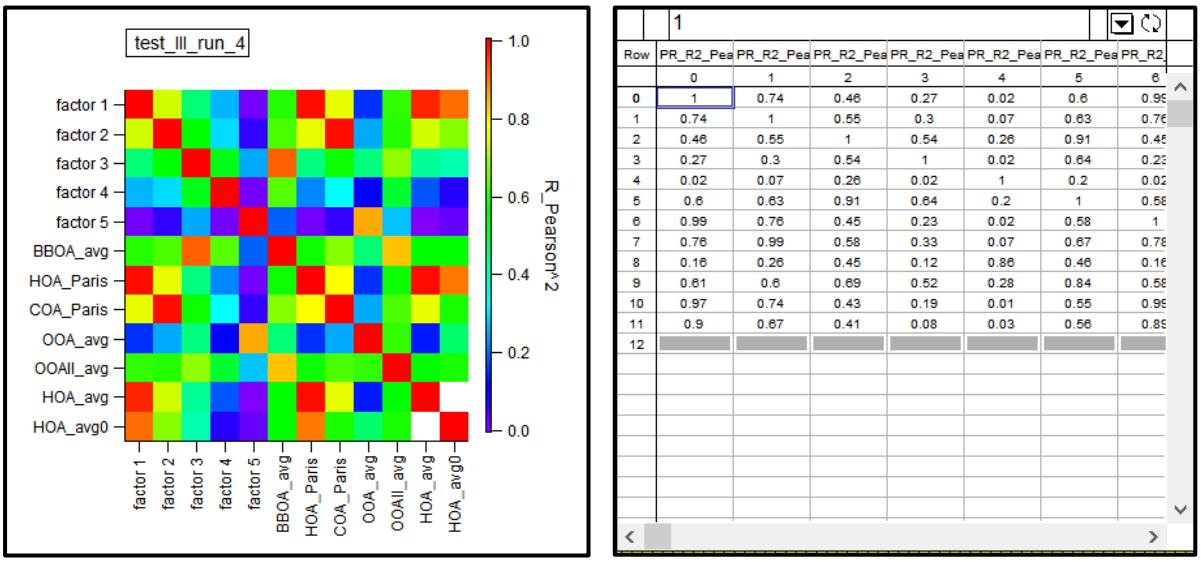


Figure 106 Correlation image plot and correlation table between the factor contribution and external tracers defined under the panel “Data”.

Tables on the subpanel “Data” allow the user to select the proper factors and externals or variables (Figure 107). Default value is that all entries are used for the correlation. With the “list” drop-down menu, the user can select the axis. With “data” drop-down menu, the user can switch between external data and variables (only time series). Y_list and x_list are independent of each other, so externals and variables can be correlated. SoFi will plot the correlation according to the position of the factor/external/variable in the list. To change the order, the entries can be dragged and dropped to the desired position. The selection and order of the externals/variables is saved, to reset it the “Reset list and selection” button should be used.

The subpanel “Graph” also contains type of the representation (table R values, table p values, image, image+R values, image+p values) and the critical value for a significance test based on the threshold. The user can vary the threshold on the same panel and can check immediately for acceptable minimum R values. This test is only available within the option R-Pearson and Kendall tau.

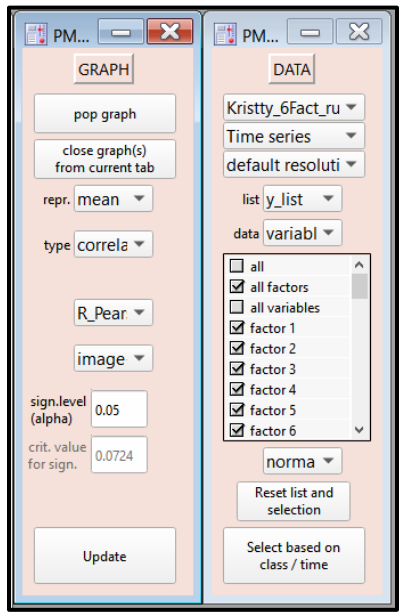


Figure 107 Graph (left) and data (right) subpanels for the section "correlation".

8.5.5.2. **Wind anaylsis**

SoFi offers the option of wind rose, pollution rose, joint probability and non-parametric wind regression (NWR) plots. The plots for the joint probability and the NWR are similar as they can be found in Zefir (Petit et al., 2017). All these plots require wind speed and wind direction. SoFi will look for waves named as they are generated in the treat external (yellow) panel. So by default, the wind_vel and wind_dir_degdep (direction dependent averaging, preferred) or wind_dir (normal average, not direction dependent) are used. If these waves do not exist, a pop-up appears where the user can select the wind speed and direction waves. The wind speed can be in either m/s or km/h. The result plot will be plotted in the same unit. The wind direction data needs to be in degrees with 0° representing the North azimuth. The user has the option between five different color gradients.

In general, a nonparametric regression is a method of estimating the mean value of a dependent variable given the value of one or more predictor variables. In the case of NWR, the dependent variable is the concentration of a pollutant *C,* and the predictor variables are the average wind direction ϑ and wind speed u. So the average pollutant concentration C is calculated as a function of wind speed u and wind direction ϑ using Kernel functions (Henry et al., 2009)

$$E(C|\theta,u) = \frac{\sum_{i=1}^{N} K_1\left(\frac{\theta - W_i}{\sigma}\right) K_2\left(\frac{u - U_i}{h}\right) C_i}{\sum_{i=1}^{N} K_1\left(\frac{\theta - W_i}{\sigma}\right) K_2\left(\frac{u - U_i}{h}\right)} \tag{27}$$

where C_i, U_i, and W_i are the observed concentrations of the pollutant and the resulting wind speed and direction at a given time point, respectively. The smoothing parameters σ and h, are estimated based on the rule of thumb (Silverman, 1986).

The Gaussian kernel is used for the wind direction

$$K_1(x) = (2\pi)^{-\left(\frac{1}{2}\right)} \exp(-0.5x^2)\,, -\infty < x < \infty \tag{28}$$

, while for the wind speed the Epanechnikov kernel is used

$$K_2(x) = 0.75(1 - x^2), -1 < x < 1 = 0 \tag{29}$$

However, there is no evidence that the results are particularly sensitive to the choice of the kernel function.

Wind rose

The wind rose shows the direction the winds blew from and the length of each "spoke" around the circle shows how often the wind blew from that direction. The user can select the bin width in which the speed should be grouped as well as the maximum speed. Furthermore, the calm speed, meaning up to which wind speed is regarded as calm/windless, can be adjusted.

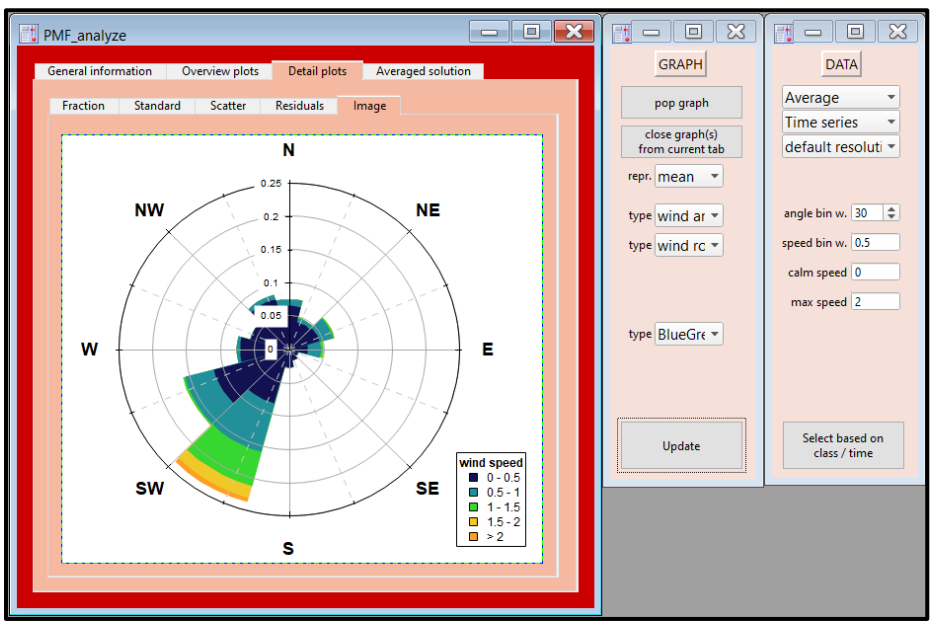


Figure 108 Wind rose. In the data subpanel (right) the user can define the angle bin width, the speed bin width and the maximum speed. By default, the calm speed is set to zero, but can be changed by the user.

Pollution rose

The pollution rose applies the same plot structure as the wind rose but substitutes a concentration time series for wind speed. Similar to the wind rose, the bin size for the angle has to be defined. Furthermore, the concentrations bin width as well as the maximum concentration has to be set. For a drop-down menu, the user can select between concentrations from solutions, externals and variables.

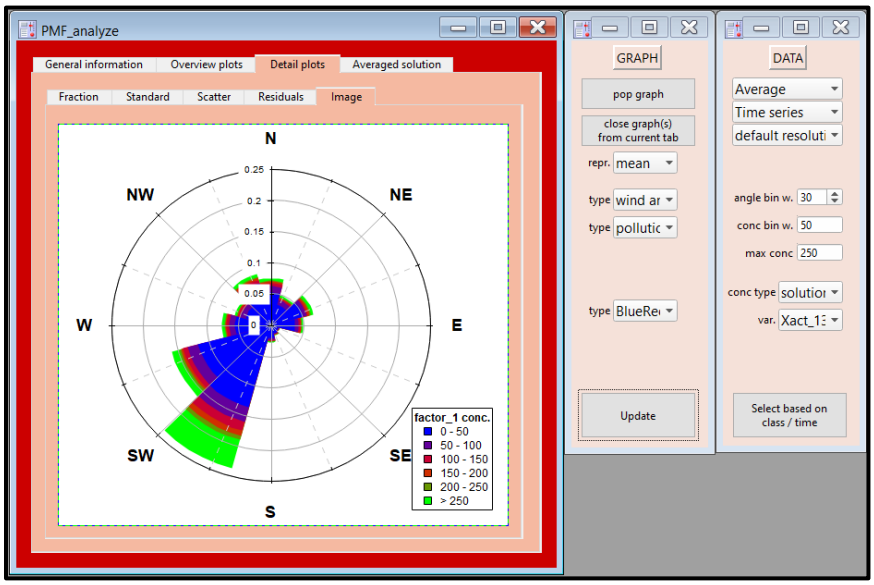


Figure 109 Pollution rose. In the data subpanel (right) the user can define the angle bin width, the concentration bin width and the maximum concentration. From the two drop-down menus the user can select which concentration should be used for the pollution rose (solution, external, variable).

Joint probability

As the name suggests, the joint probability shows the distribution of wind speed and wind direction. It is similar to the wind rose, which is the traditional way to visualize the joint distribution of wind direction and wind speed, however the joint probability is much more detailed. The user has to give the maximum speed (upper limit) to which the plot should be drawn. Since it is a polar plot, the user has to define angular and radial resolution, i.e., how many points should be drawn to make it look smooth.

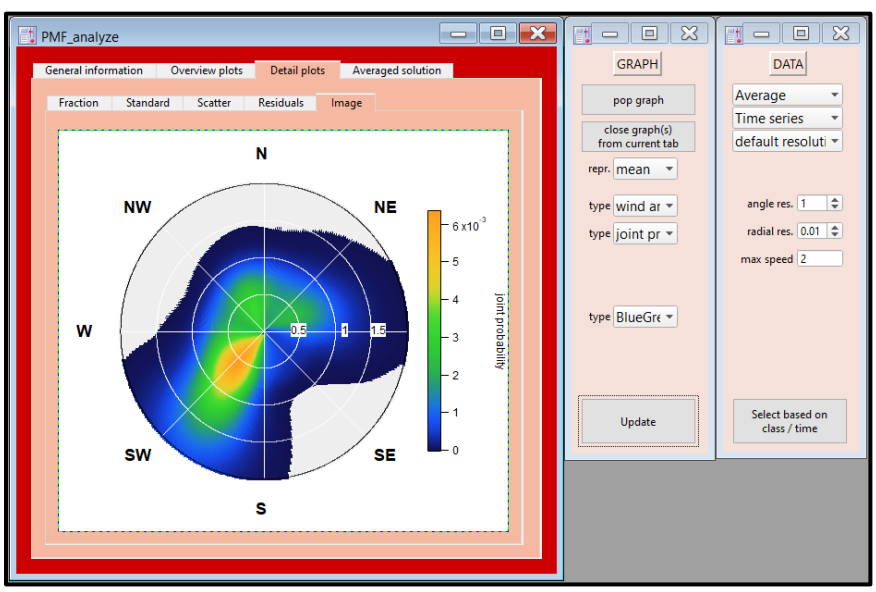


Figure 110 Joint probability. The user has to define the angular and radial resolution as well as the maximum speed.

NWR (nonparametric wind regression)

The NWR shows the detailed relationship between the selected pollutant and the wind direction and the wind speed. Aside from its better performance at separating source in the same general direction, the NWR method is more efficient than the pollution rose because it does not require choosing wind direction bins. For the NWR, the user only has to define the maximum wind speed. From the dropdown menu, the desired pollutant concentration (solution, external or variable) can be selected. Similar as for the joint probability, the angular and radial resolution have to be defined. The smaller the values, the higher the resolution, the longer the calculations. The grey areas are regions where the joint probability is too little, i.e., there are too few points for a reliable estimate. Some users might be familiar with this graph as it is similar to the one available in ZeFir (Petit et al., 2017).

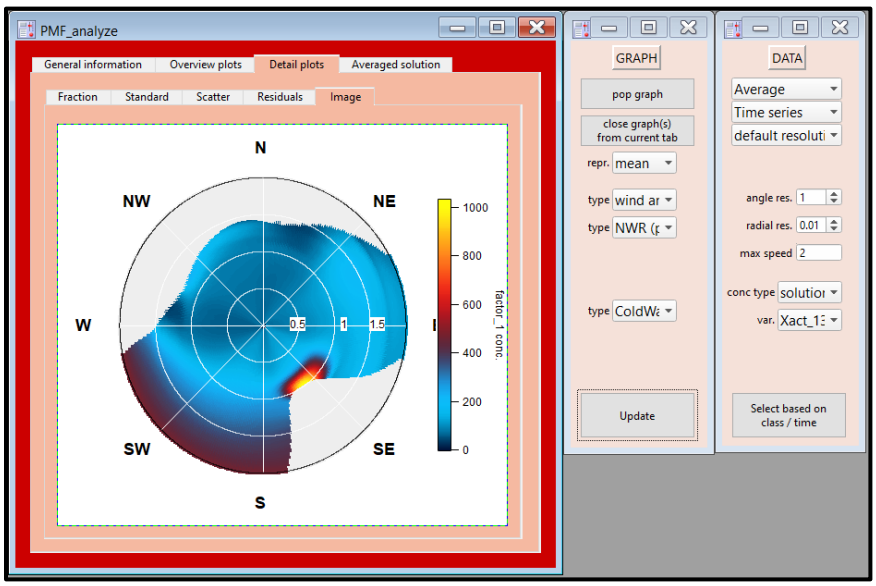


Figure 111 NWR. The user has to define the angular and radial resolution, the maximum speed as well as the pollutant.

8.5.5.3. **Hierarchical clustering**

This option is only available with Igor 9 or newer.

This option hierarchically clusters the selected factors and external or variable time series, as selected from the list. By default, the Euclidean distance is used for the clustering. The time series of the factors and externals can be pre-processed internally (recommended): mean center, normalize [0,1], center and normalize [-1, 1], X_i/SD(X), (X_i-mean)/SD(X). There are 3 methods available in SoFi to determine the dissimilarity between nodes in the dendrogram that represent more than one data vector ("linkage" method). By default, the single method is selected, where the dissimilarity is the dissimilarity between the two nearest elements of each cluster (Nearest

Point algorithm). The complete linkage method is also referred to as Farthest Point algorithm. The average method is the average of all the pairwise dissimilarities between points in each cluster.

The user can choose between different dissimilarities the clustering is performed over: Euclidean or Manhattan distance or a correlation-based (R Pearson) dissimilarity matrix. Hierarchical clustering is not NaN stable. So for the first two, SoFi needs to ignore all data points where at least one external/variable is NaN, for the latter SoFi can either follow this (more stable) or compute the correlation for all pairs based on the available data (might lead to non-meaningful correlations as not all correlations are based on the same number of points).

The hierarchical clustering plot (Figure 112) shows on the left-hand side the dendrogram. If the checkbox “color clusters” is checked, the different clusters are colored according to the chosen cut distance. On the right-hand side, the pre-processed time series of the factors and externals is plotted as an image plot.

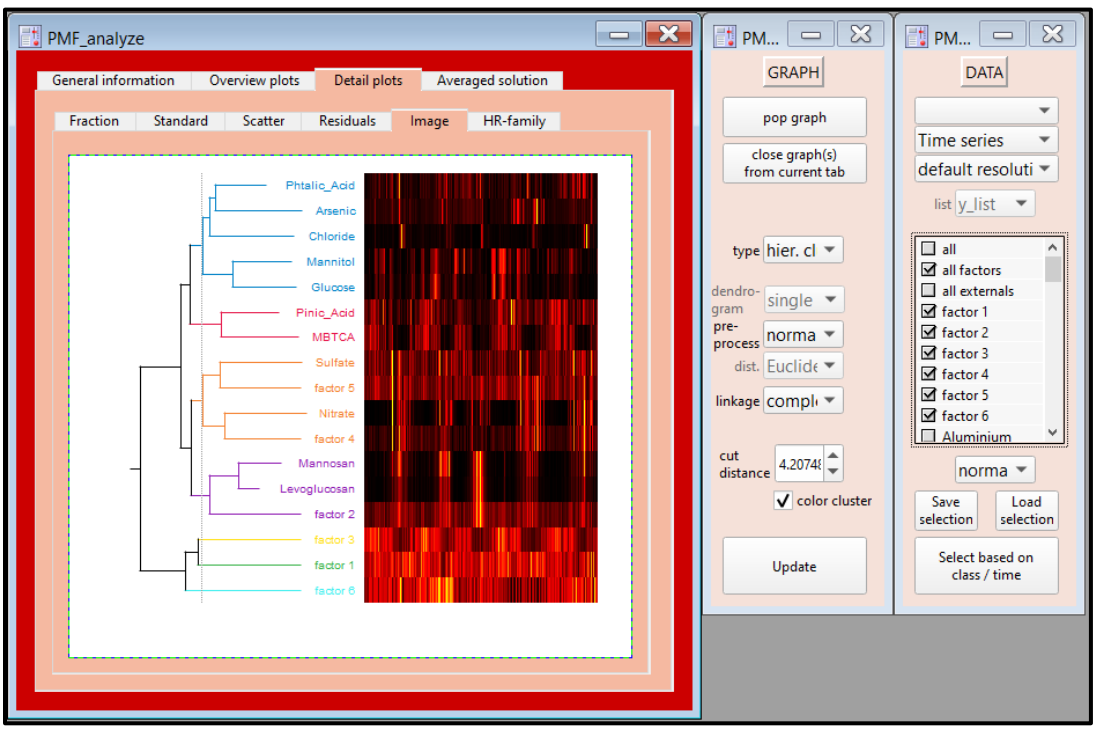


Figure 112 Hierarchical clustering.

8.5.6. HR plot

The tab “HR-family” is accessible in case the text vector containing the family information for the variables with the name “speciesfamilytext” is stored in addition to the PMF input data (import data, see 4.1). Creating or updating user-defined families after running PMF is possible, but the chosen PMF run or average has to be updated according to 8.2.3.

8.5.6.1. **Stacked profiles**

For HR-AMS data analysis the family-stacked version of the factor profile plot is supported in SoFi. Thea means that per UMR mass, all the families are stacked on top of each other, which makes it better readable than the option in the overview plot where the ions can also be color-coded according to the families. Moreover, all factors are equipped with the OM:OC, O:C, H:C, N:C and S:C calculation (Figure 113), which can be turned off if desired. The O:C and H:C ratios are given as mole fractions and the user can choose between one of the two options “Aiken ambient” (Aiken et al., 2008) and “Canagaratna ambient” (Canagaratna et al., 2015) method for visualizing these ratios.

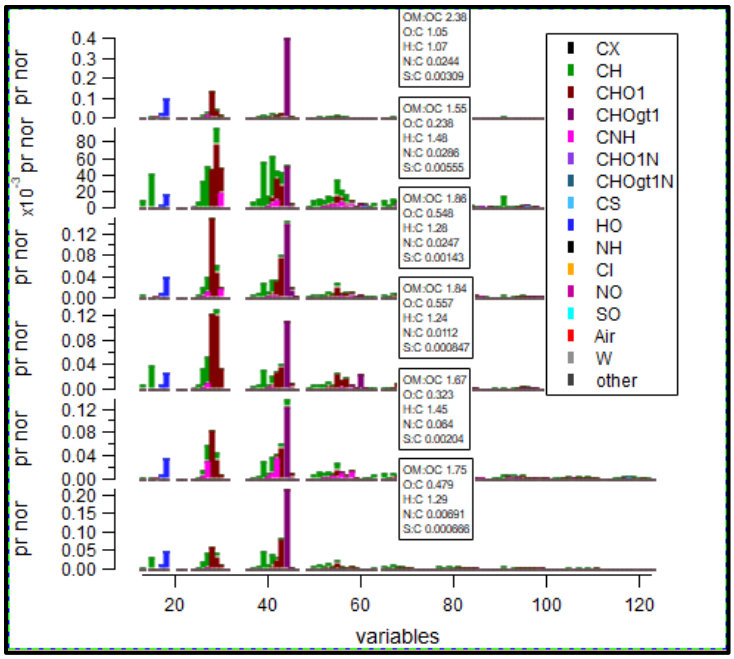


Figure 113 HR family stacked plot for HR-AMS analysis.

8.5.6.2. **Mass defect**

In high resolution mass spectra, the Kendrick mass (Kendrick, 1963) is often used to identify homologous compounds that differ by only a few base units and the Kendrick mass defect (KMD) analysis is often used for an easy visualization of composition. SoFi can plot the mass defect either according to Kendrick original definition of the mass defect or the more widely used definition in atmospheric chemistry by Aerodyne and Alton (Alton et al., 2023):

$$KMD_{\frac{m}{z}} = \frac{m}{z} * \frac{\text{Kendrick}_{\text{base}}}{\text{IUPAC}_{\text{base}}} - \text{round}(\frac{m}{z} * \frac{\text{Kendrick}_{\text{base}}}{\text{IUPAC}_{\text{base}}}) \quad (30)$$

The KMD is calculated as the difference of the measured exact mass multiplied by the ratio of the Kendrick base unit divided by IUPAC base unit and the rounded measured exact mass multiplied by the ratio of the Kendrick base unit divided by IUPAC base unit. Kendrick (1963) proposed CH_2

as a base unit, since then other units were used too. SoFi currently offers 4 base units to choose from (Table 8).

Table 8. Kendrick unit conversion options in SoFi.

Base unit	Kendrick base unit	IUPAC base unit
^{12}C (= rounding only)	12	12
$^{12}C^{1}H_2$	14	14.01565
^{16}O	16	15.9949
$^{12}C^{1}H_2{}^{16}O$	30	30.01055

The mass defect plot (Figure 114) can be color-coded by families (only for user-defined families), the marker size can be changed so that it is a function of the intensity. If a log function is used, there is the option to use a minimum intensity (floor) to stabilize log scaling if the data spans several orders of magnitude.

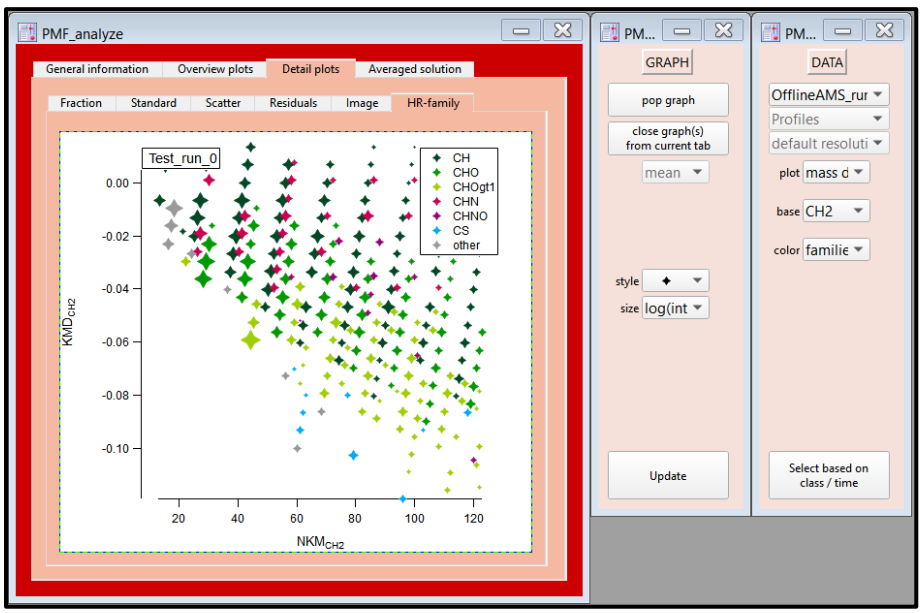


Figure 114. Kendrick mass defect plot, color-coded by user-defined families and markers scale with the log(intensity) of the signal.

8.5.6.3. **Van Krevelen plot**

The Van Krevelen diagram (Van Krevelen, 1950) is being used for the graphical interpretation of HR mass spectrometric data, in which the elemental ratios of individual compounds are plotted instead of those of the bulk material. In the diagram (Figure 115), hydrogen carbon atomic ratios (H:C) and oxygen carbon atomic ratios (O:C) are plotted against one another. The markers can be scaled to the intensity of the ion in the chosen factor. In addition, the markers can be colored

based on user-defined families and the carbon number. SoFi can read chemical formulae with up to one bracket (e.g., $(H_2O)_3H_3O$).

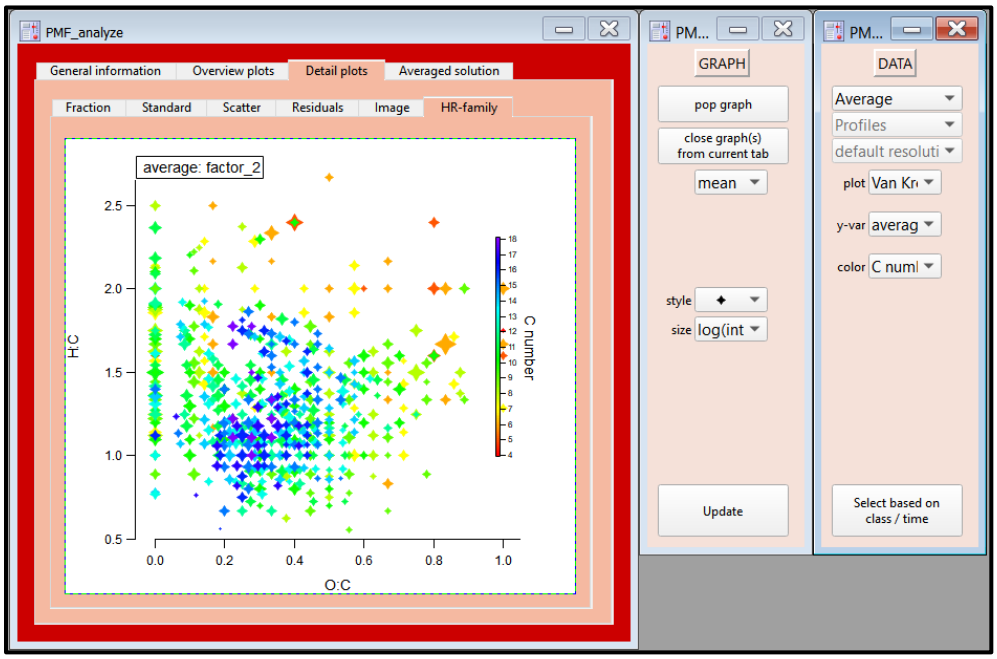


Figure 115. Van Krevelen plot, color-coded by the number of carbon atoms.

The intensity of the ions can also be scaled with a log function, where the option to use a minimum intensity (floor) exists. This is to stabilize log scaling if the data spans several orders of magnitude.

8.5.6.4. **Carbon number distribution**

The carbon number distribution plot bins ions based on the carbon number. Ions with the same carbon number are further separated into different subcategories, shown by the color code. SoFi supports a classification based on user-defined families, number of oxygen, H:C and O:C. The range for the H:C and O:C can be defined in the table below the color selection (see sample in Table 9). SoFi can read chemical formulae with up to one bracket (e.g., $(H_2O)_3H_3O$).

Table 9. Sample H:C list. The table should include only <, > and ranges. If a range is provided, the lower limit is included in the range, not the upper.

Table in SoFi	SoFi reads
< 1.1	< 1.1
1.1 - 1.3	>= 1.1 && < 1.3
1.3 - 1.5	>= 1.3 && < 1.5
1.5 - 1.7	>= 1.5 && < 1.7
> 1.7	>= 1.7

Since it could be hard so grasp how much each subcategory shares to the total factor composition, the user has the possibility to add summary pie charts (Figure 116, right). The pie charts are only properly presented when the graph is being popped.

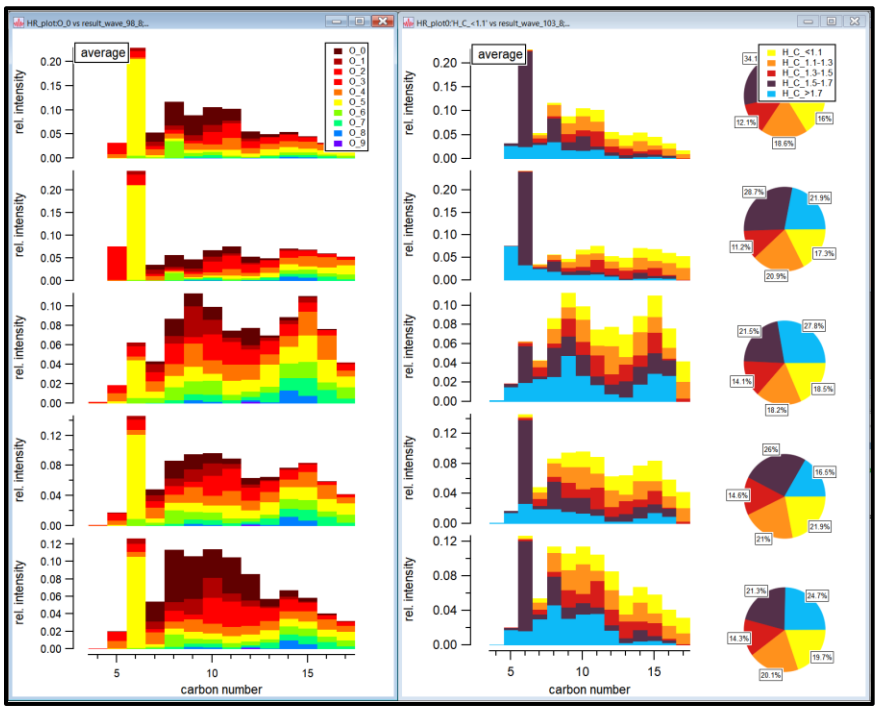


Figure 116. Factor-wise carbon number distribution, color-coded by the amount of oxygen atoms (left) and the H:C, including the pie chart summary for every factor (right).

8.5.7. Category plot

These plots provide a detailed inspection of the solutions, organized into pre-defined classes (categories) for better analysis and comparison. Concentrations, contributions etc. are shown category-wise. Category plots are displayed as bar plots, allowing for a clear visual comparison of the contributions of different categories. These plots are most useful for e.g., data from different stations or with different size fractions.

All time series visualization options available in the standard plots are also accessible here, Additionally, this section includes an extra feature that allows plotting the entire solution in a single view (Figure 117).

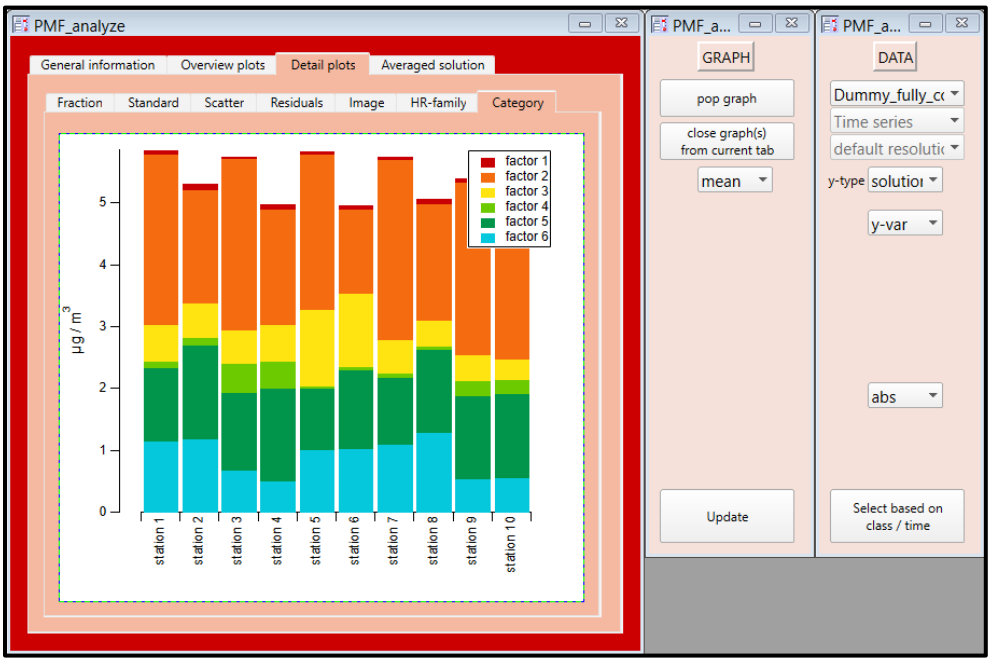


Figure 117 “Category” tab with the concentration of each factor in each class (category).

8.6. Averaged solution

This tab is useful for the analysis of the averaged solution in particular for rolling PMF runs. It contains statistical information on the degree of coverage of the PMF input data, a method for quantifying the PMF error, statistics on the time-dependent a-values and time-dependent factor profiles. However, the tab “error estimation” will also serve well when quantifying the PMF error for regular (non-rolling) PMF runs.

8.6.1. Repeats

When bootstrapping or performing a rolling PMF run some modelled data points might be absent, e.g. for only a few numbers of replicates when bootstrapping or very tight thresholds are applied for the criteria. The tab “repeats” reveals the coverage and how good the averaged PMF result represents the PMF input data. In addition, the amount of repeats over time is also plotted (see Figure 118). Ideally, the amount of repeats is uniformly distributed over the entire PMF input and strong differences suggest a still insufficient amount of selected PMF runs, systematic over-/underrepresentation of the data due to a non-ideal number of factors, or constraints, etc.

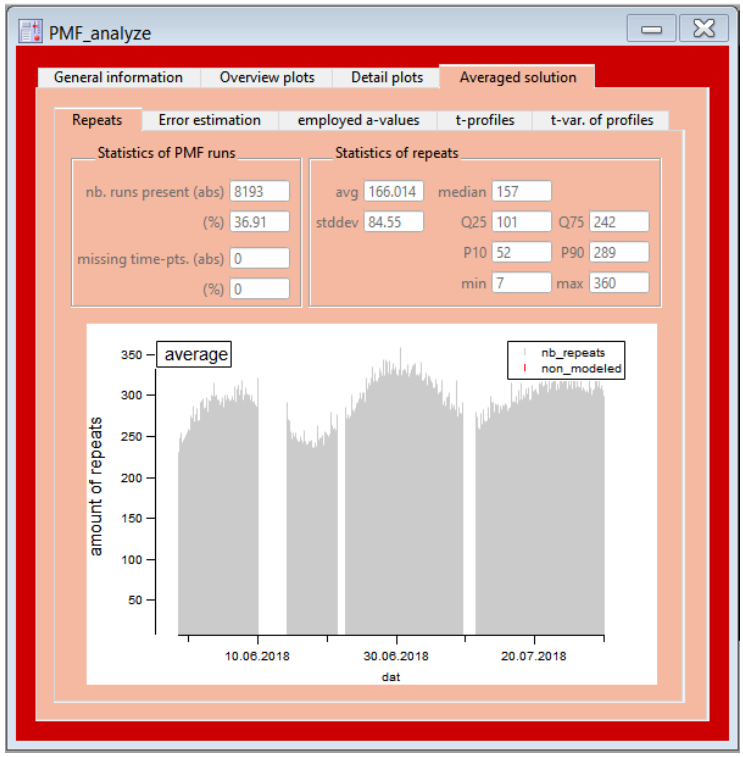


Figure 118 "Repeats" tab with some statistics on the PMF run and on the amount of repeated input points. This is particularly useful for bootstrapped or rolling PMF runs.

8.6.1. Mass error estimation

Applying resampling strategies such as bootstrap (see 6.4.1) and techniques aiming to systematically explore the rotational ambiguity such as a-value/ pulling (see 2.4) generate repeats for a timepoint *i* in the PMF solution. The variability within these points can be used to infer the rotational and statistical uncertainty. These two types of uncertainties are discussed in the following and are referred to as PMF error. It should be noticed that more types of errors might be added to the final PMF error, such as for example the model error (error of using a different number of factors, a different type of anchor for the a-value constraints) or the error related to the criteria such as the type of criterion and its related thresholds. The proposed PMF error in this study is given by the following formula:

$$PMF_{error,i} = 100 \cdot \left(\frac{spread}{mean/median}\right)_i \tag{31}$$

where the spread could be the differences between the third and first quartile or the standard deviation for a time-point *i*. The slope of a linear fit when plotting the spread against the mean or median value represents the relative error. Figure 119 represents both, the scatter plot and the probability density function (pdf) of the top formula for each timepoint *i* $\left(\frac{(q3-q1)}{mean}\right)_i$.

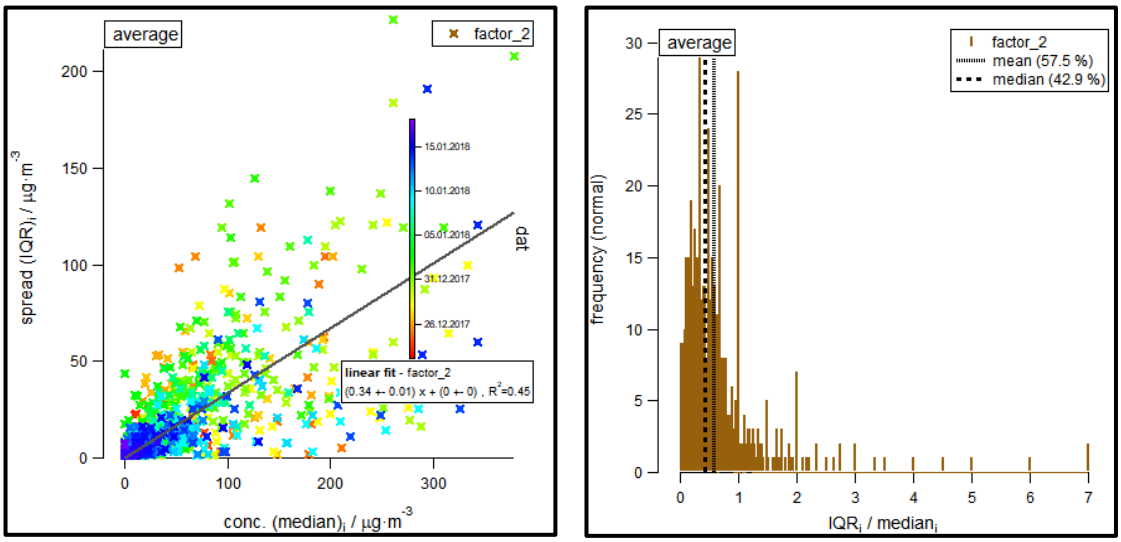


Figure 119 Spread in terms of IQR is plotted against the median value for factor 2 (left) and its corresponding density function plot (right). The mass error for this factor is half the slope in percentage, i.e., +/- 17 % on the scatter plot. On the frequency plot, the error is more represented by the overall median value of 42.9 %. Hence, the error is +/- 21 %, given possible small mean values for some points *i* causing extremely large values affecting the overall mean of 57.5 %.

8.6.2. Employed *a*-values

When performing random *a*-values in e.g. a rolling PMF run this tab visualizes the statistics of the employed a-values over time and as histogram (see Figure 120 and Figure 121). This graph reveals information on the degree to which the constrained information required to adjust to the PMF input data in time. High *a*-values with a small spread suggest that for these episodes the constrained information required more adjustment/freedom.

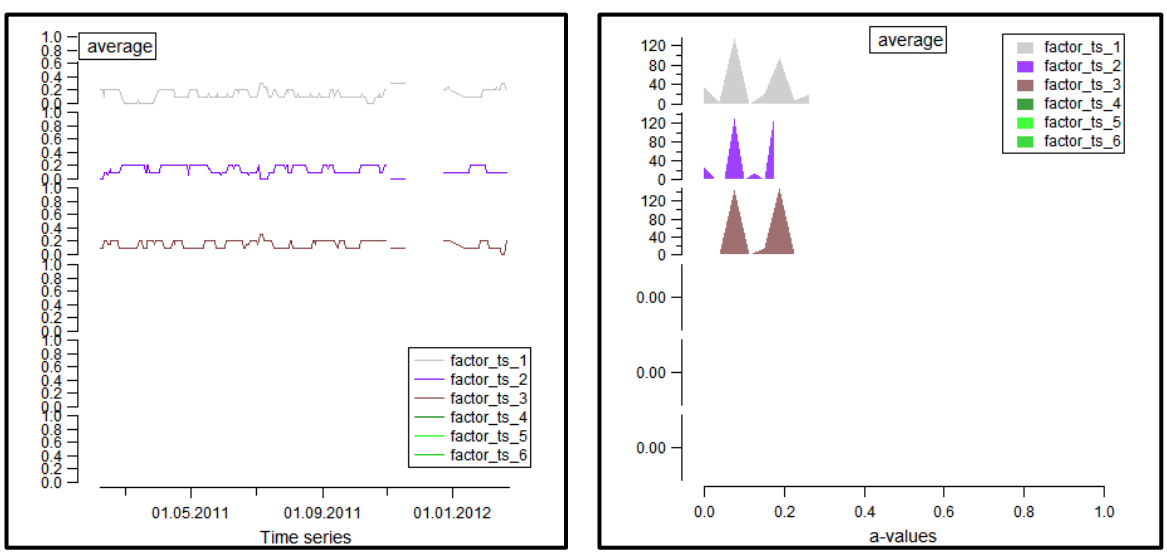


Figure 120 (left) effective *a*-values for the constrained factor profiles over time and (right) histogram representation of these a-values.

In addition to the mean, standard deviation, median and quantile statistics, the probability density function (pdf) can be visualized too.

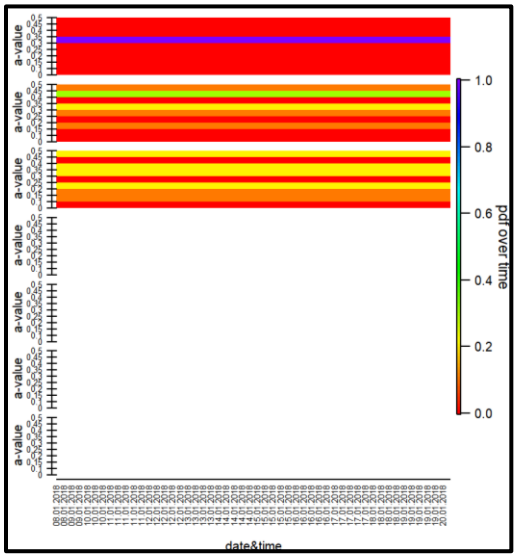


Figure 121 Probability density function of the a-values over time.

8.6.1. t-profiles

Rolling PMF aims to assess time-dependent factor profiles. This information is shown within this tab by plotting the factor profile over time as an image plot. The profiles can be normalized over the profile (sum of the variables equals one) or over time (each variable is divided by its mean value). Moreover, a 2-way ANOVA test is performed that tests all profiles for independence. The Null hypothesis (H_0) describes the assumption that all profiles present in the t-profile matrix are drawn from the same distribution. The results are visualized on the subpanel “GRAPH”. (see Figure 122)

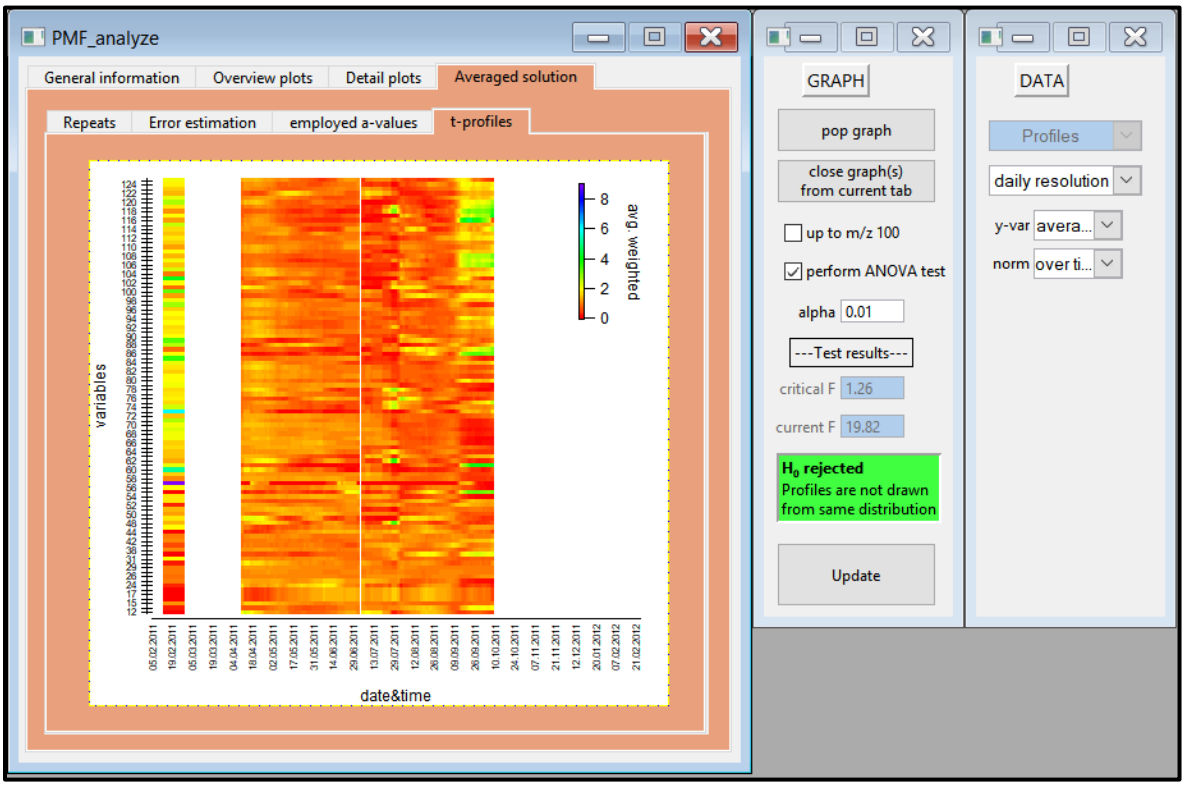


Figure 122 factor profile in the y-axis plotted over time. Variations in intensity of the variables reveal changes in the profile shape.

The user can also inspect the profiles by plotting averaged profiles over time. Drawing a marquee on the interested region (note that the x-axis only matters, i.e., the entire profile is drawn, even when not all variables (y-axis) are within the marquee, loads the marquee option after pressing on the right mouse button (see Figure 123). This process can be repeated until the desired averaged profiles can be directly compared together (see Figure 124).

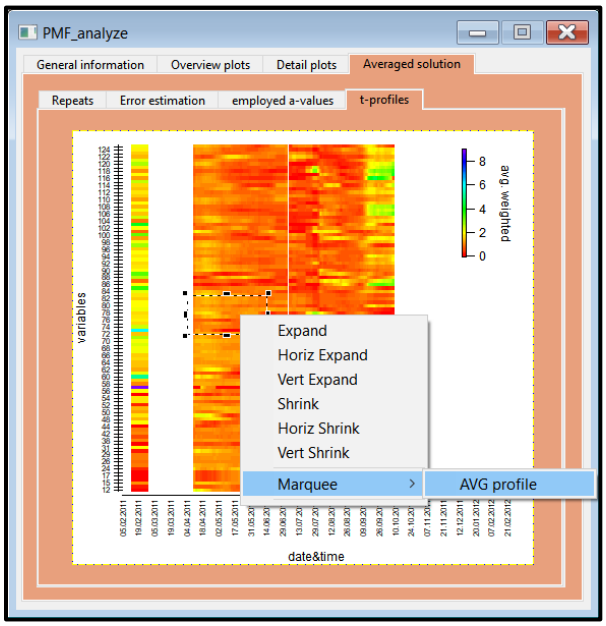


Figure 123 factor profile in the y-axis plotted over time including the option for plotting averaged profiles.

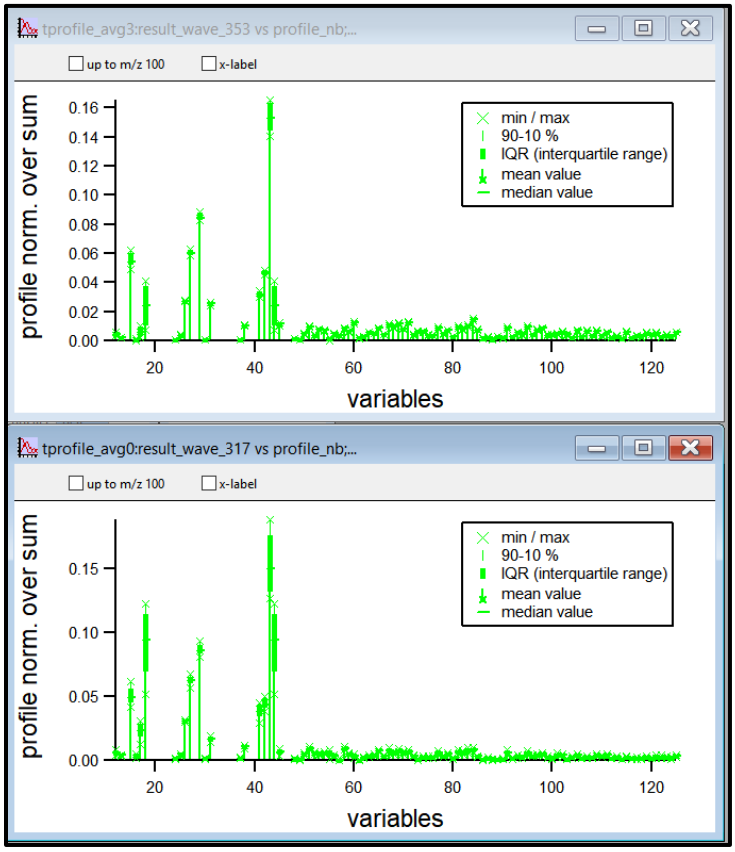


Figure 124 two averaged factor profiles including their statistic (top graph for the winter and bottom graph for the summer period).

9. Independent Analysis

9.1. Highlight interesting periods

The highlight panel provides a clear, customizable, and effective way to emphasize critical data periods, i.e. highlighting a fireworks period, make the distinguishment between heating/non-heating periods or highlight sever haze periods. Highlighted periods are set by a specifying a specific (continuous) time period and a period name. Additionally, color, opacity, and line size used to color-code the graph can be adjusted, ensuring that the highlighted sections are clear and visually distinct. Each period can exist from several sub-periods.

Depending on what kind of plot should be highlighted, the panel either accepts normal dates (European format with day, month, then year), numbers (like for diurnal plots or yearly plots) and text (same abbreviations as used in the weekly or monthly plots). The highlight panel can be accessed via the “highlight period(s) (tinted background)” button, available in the overview and standard plot.

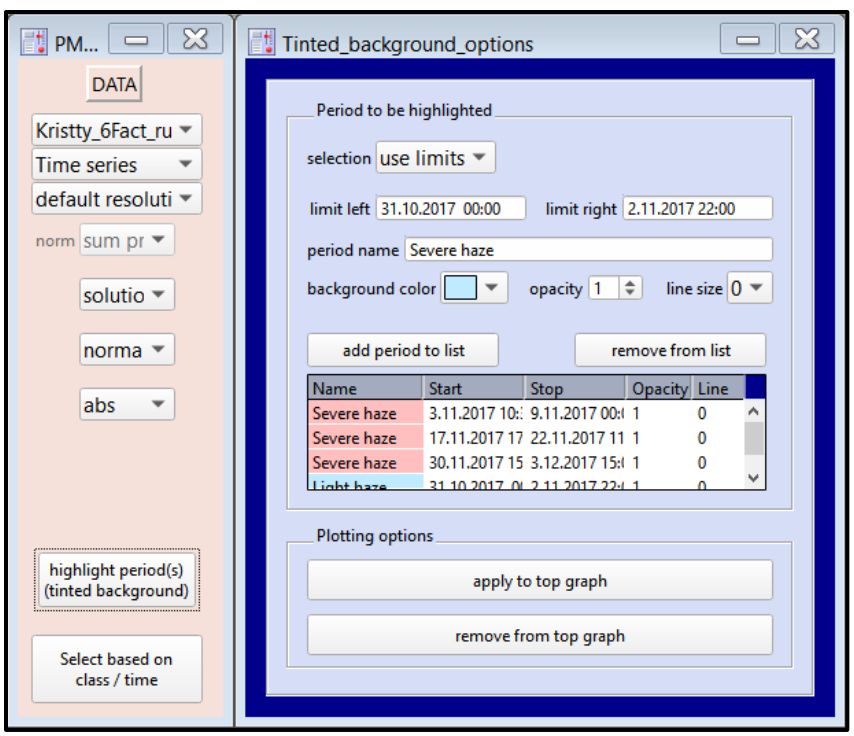


Figure 125 Highlight panel. Available for the time series overview plot as well as for time series standard pots.

9.2. Independent bootstrap analysis

For the explore and the analysis panel, SoFi offers an independent bootstrap analysis of the (multi-)linear fit parameters. This allows to assess the spread or variance of some statistics without relying on any assumptions for the underlying distributions, e.g., that the data is normally distributed.

The bootstrap panel can be opened through the small purple “BS” button which appears when the linear fit is activated for a scatter plot.

In the first tab, the bootstrap settings are being defined. This can be done either automatically, meaning that SoFi will use the variables, externals etc. selected in the explore and analyze panel, respectively. When the variables in the explore or analyze panel are changed, the bootstrap panel updates automatically.

SoFi also offers the option to manually define an x, y and time wave. These waves must all have the same length.

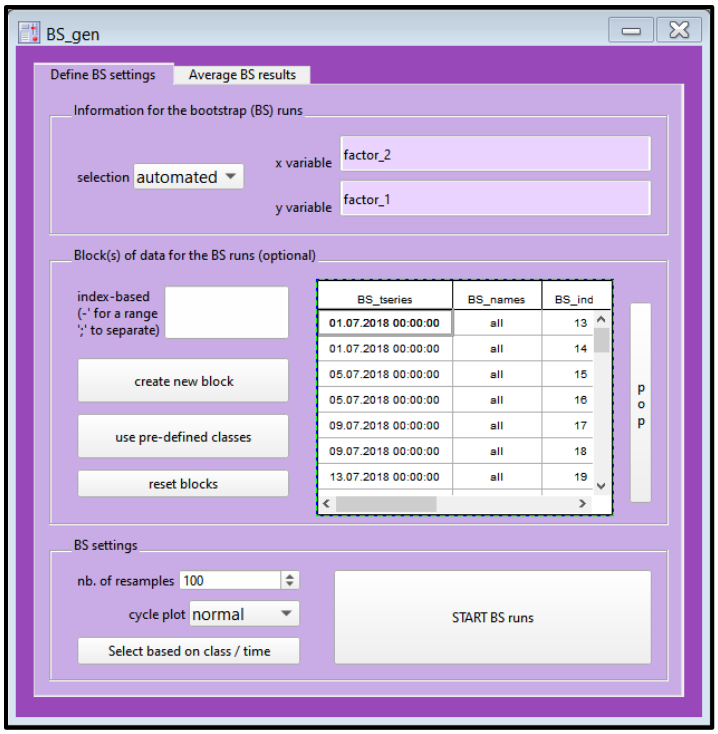


Figure 126 Bootstrap panel where the waves that are to be bootstrapped can either be selected automatically (as shown here) or manually. In addition, classes can be defined (optional).

Optionally, the user can define blocks. For each block, the points are bootstrapped separately before all blocks are set together again to rebuild the input. With this, it can be ensured that important events, seasonality etc. are still kept in the data. SoFi supports following syntax for creating blocks:

“Index based blocks”	use “–” for a range and “;”to separate single indices or ranges and press "create new block” to create the blocks and name them
“Class supported block”	if the user defined classes before, SoFi can use these classes to create blocks.
“Monthly blocks”	use the keyword “monthly” to create blocks month-wise. The blocks will automatically be named after the month and year of the corresponding time point
“Hours blocks”	use the keyword "hours" followed by the hours that should be used for the blocks. E.g. “hours17-20” will create one block that includes all points that are between 5 and 8 pm.
“Minutes blocks”	use the keyword "minutes" followed by the hours that should be used for the blocks. E.g. “minutes10-20” will create one block that includes all points that are between minutes 10 and 20.
“Fixed length”	use the keyword “block” followed by the number of point you want to have grouped together. E.g. “block10” will create new blocks every 10 points. Please note that this format is only supported for data with low time resolution.

By popping the BS_names, the block list can also be manually adapted.

Before starting the bootstrap analysis, the user can define the number of resamples, meaning how many times the analysis should be repeated. The goal is assessing the number of resamples for which the spread of the statistic doesn’t show substantial differences anymore. Furthermore, the user can select over which time point (select based on class/time) the analysis should be performed. In addition, the bootstrap analysis can also be performed over cycles to estimate the uncertainty of the cycles (besides the averaging statistics that can be accessed directly from the red analysis panel).

The results can be inspected in the second tab (Figure 127). The user has to define which fit parameter should be inspected. For simple time series (standard plot) the average concentration over all bootstrap runs can be inspected. For the (multi-)linear fits, the user can inspect the bootstrapped intercept, slope and correlation. For multilinear regression in addition the bootstrapped fitting parameters of the regression are available. The number of bins that can be defined refer to the histogram. To get an idea of how many points were used during each bootstrap repetition, the number of resampled points can be added to this plot. An overview of the statistics is displayed on the bottom right and can be popped to be easily copied.

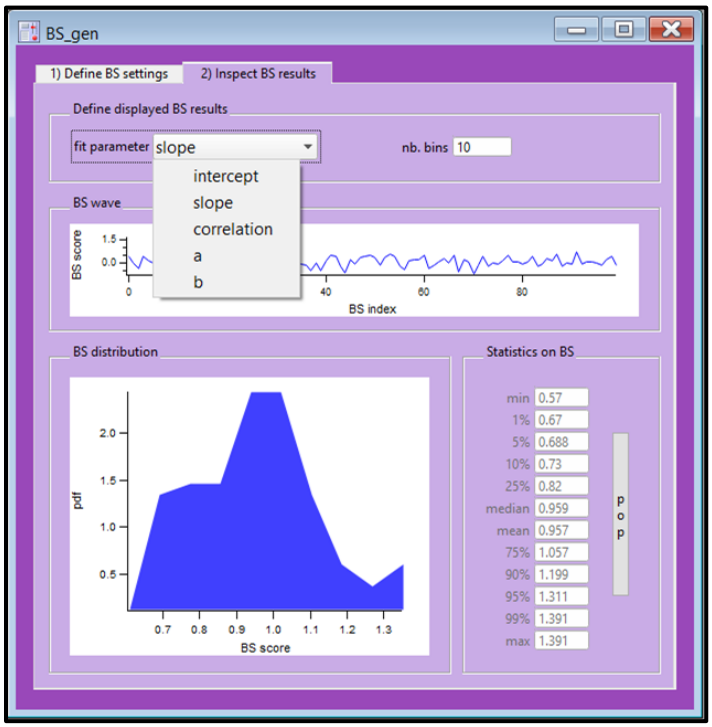


Figure 127 Inspect the results in the bootstrap panel. The user can inspect the bootstrapped results of the slope, intercept and correlation. In case of a multilinear regression (as here), also the fitting parameters thereof are accessible.

In case the bootstrap resampling was performed to estimate the cycle uncertainty, a third tab will automatically be available. In this tab, the cycle plot with the bootstrapped data is visualized. The user can choose between the standard representation (similar to what is available in the standard plot) or box and violin plots, respectively. The user has a wide range of options to best present the box plots, including but not limited to what should be plotted (e.g. the single data points, mean, what the whiskers stand for) and in which colors (Figure 128).

To access the statistics in a non-visual way, a statistics table can be popped.

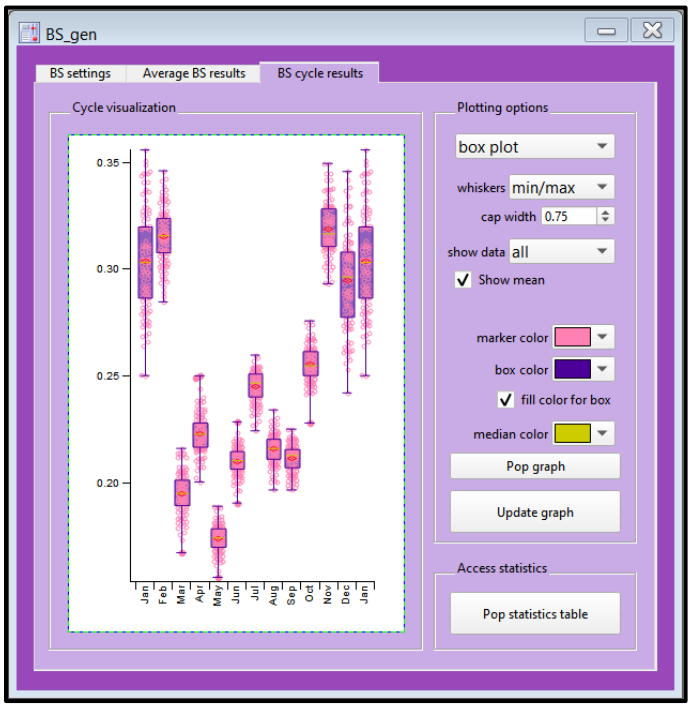


Figure 128 For bootstrap analysis of the cycles, the user has one more tab where the cylce plots can be inspected.

9.3. Marquee tool

The marquee panel allows the user to select point in a graph and highlight the same points in a different graph. For example, the user finds an interesting event in the factor time series and can highlight the same time points in graphs with tracers, temperature, RH etc. where the time series is not plotted.

Main options

The marquee tool is accessible from all graph windows starting with “var_expl”, “sol_nor” or “sol_scat”, i.e., all plots (popped or in the SoFi panel) created in the SoFi explore (green) panel or standard and scatter plot from the SoFi analysis (red) panel. Note: when closing the red panel and keeping graphs popped, these get renamed to “userwin” and are not accessible for the marquee tool. The marquee panel can be accessed by dragging a marquee and select “select period” or “set as active window” under “SoFi marquee” > “Marquee analysis tool”. This will open the marquee panel as shown in Figure 129.

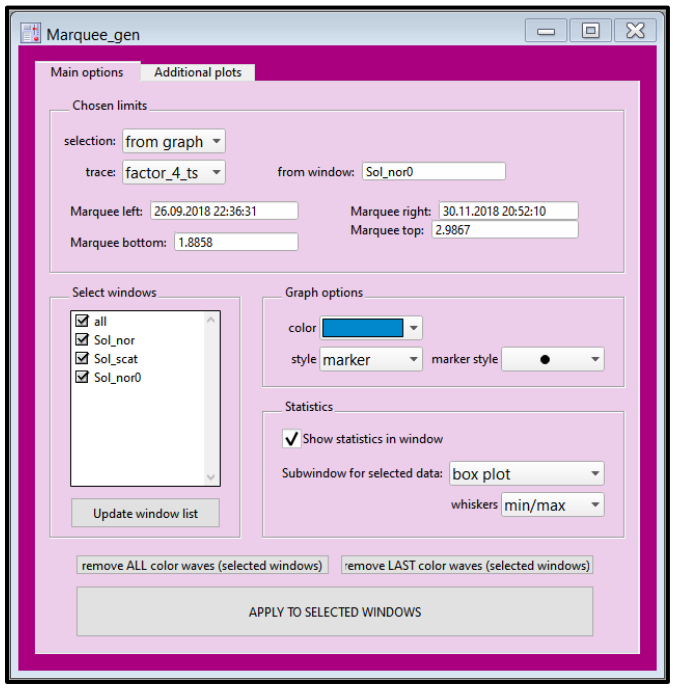


Figure 129 Marquee panel, where the limits and “base” window are defined (top), as well as in which other windows these points, that fit the limits in the defined window match, should be highlighted and in which way (color, statistics, detail plots).

The user needs to define which is the period of interest that should be taken as the “base” to color also the other windows. The window for the base wave is shown under “from window” and can be changed by clicking into any “sol_nor” or “sol_scat” graph, dragging a marquee and selecting “set as active window”. The trace automatically updates to the traces in the selected window and can be changed through the drop-down menu.

The user has three options to select which data from the selected window and trace should be used.

- “From graph”: the limits from the marquee when “set period” as applied. They are shown but can only be changed via the marquee.
- “use classes”: the user has the option to select classes that were created before PMF or add its own classes. The classes that were defined before PMF and that are also accessible for other analysis are shown in a list. One class at a time can be selected. The user has the option to add classes. For those, the user has to either provide a wave with indices (“new class from wave”) or write the indices into a newly created wave (“add new class manually”). Note, that the classes only work for graphs that contain the time series/profile as it was used for the PMF calculations, it will not work when the “select based on class/time” option was used for the plot because indices would not match anymore.

- “use limits”: the user can enter the limits him/herself. Date as well as numeric limits are accepted.

Under “select window”, the window(s) in which the data should be colored can be selected. The list shows all the windows that contain data with the same length as the trace selected as the base trace. It is updated if a different base window is selected. If new windows are being generated in the meantime, the list does not automatically updates, the user is asked to click the “update window list”.

The user can select different colors and markers/lines for the highlighted data via the dropdown menu under “graph options”.

It is possible to show some statistics on the selected data in each window. If the “show statistics in window” box is checked, number of points, mean, standard deviation, median and inter quantile range (IQR) are shown in each window. The statistics for the latest added points is shown. If other statistics should be shown, the user is asked to drag a marquee on one of the plots and select “update shown statistics” in the marquee analysis menu. There different layers (top layer is layer o) for which the statistics should be shown. This can either be done on the current window or for all windows that already show some statistics. Additionally, a subwindow with the selected data can be added to the graph. The options for the subwindow include linear regression, box plot and violin plot.

When “apply to selected windows” is clicked, for each window that is selected, the points corresponding to the chosen limits are extracted and appended as a new wave to the window. To remove the points from certain windows, “remove ALL color waves (selected windows)” should be clicked, to only remove the last one plotted in the selected window, select “remove LAST color waves (selected windows)”.

<u>Additional plots</u>

The user has an option for additional plots. In the plots from the main menu, the data shown in the subwindow is always the data selected. Here, the user has the option to also plot cycles of the highlighted data, scatter plots and pie charts. For these plots to be available, at least on marquee wave has to be applied.

For all options, either the subwindow can be updated or a new subwindow can be plotted next to the position of the first subwindow. The subwindows can be moved and re-sized in the tool mode (ctrl + t). The plots however can only be added if the tool mode is deactivated (again ctrl + t). For cycle and scatter plots, there is also the option to pop the plots into a separate window.

In addition, the control bar at the top of the graph, is always updated with the newest added subwindow data.

- **Cycle plots.** Available if the main graph includes a time axis (so not available for scatter plots). From the drop-down menu, the user can select the graph of interest. The next drop-down menu, shows all possible marquee waves (different periods that the user highlighted). The user has the option from daily to yearly cycles. As well as how the cycles should be displayed (mean, standard deviation, median, IQR, 10th / 90th percentile and min/max).

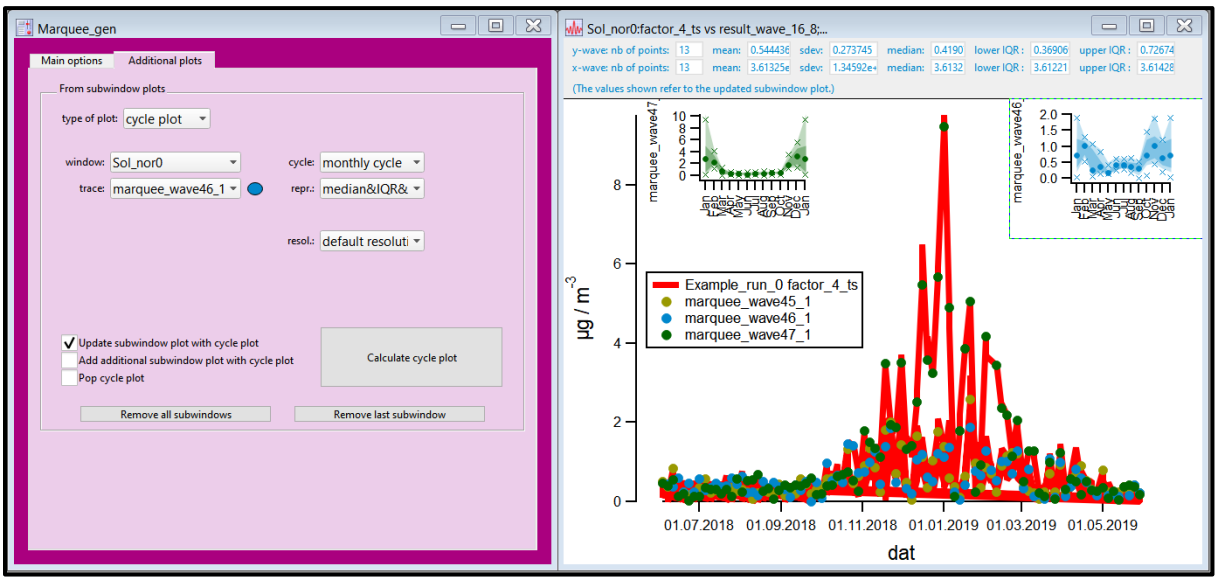


Figure 130 Time series of a factor, highlighted are different classes. Cycle plots show monthly cycles for the green and blue class.

- **Scatter plots.** Plot your highlighted data against any other data. The user again has to define the window and select the marquee wave of interest. If the main plot is a scatter plot, the user has to further specify whether the x- or y-axis should be used for the scatter plot in the subwindow. For the scatter plot, any time series can be selected: first specify the type (e.g. external) and then the variable (e.g. chloride).

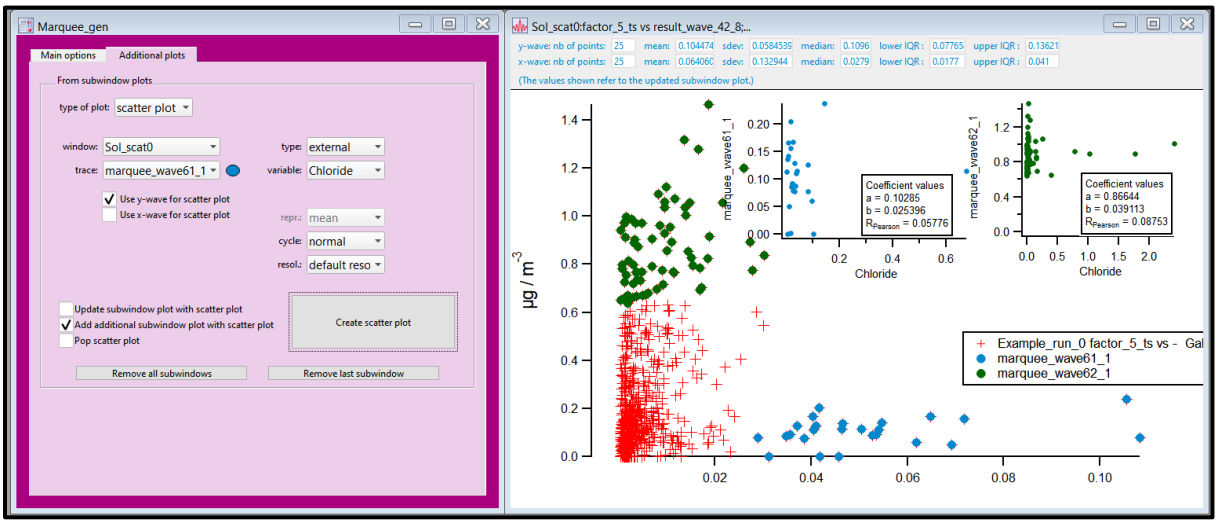


Figure 131 Subwindow scatter plot between the two subselection of the factor vs an external.

- **Pie charts.** Pie charts of the factor contribution (with or without expl. variation) for the highlighted period. The user can choose between contribution of the factors and the contribution of the factors including the unexplained variation. Furthermore, the user has control over the labelling options.

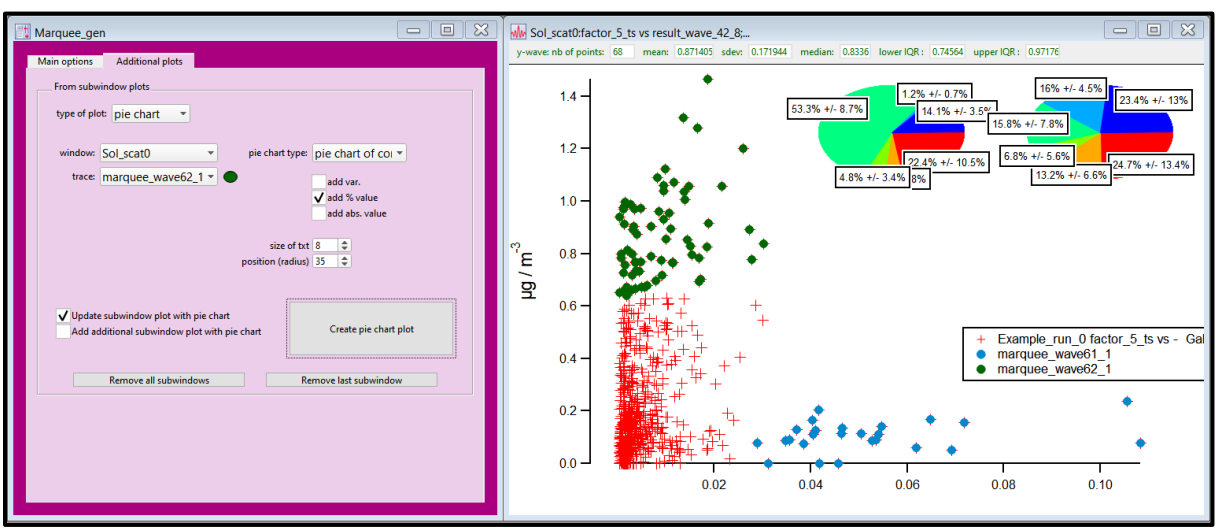


Figure 132 Pie charts for the two periods of interest.

9.4. Cluster panel

The cluster panel allows for clustering data either by K-Means or farthest point (FP) clustering. More details on the clustering algorithm of these two calculations can be found in the Igor help menu.

9.4.1. Cluster options

In SoFi, one can cluster different variables (including solutions, externals, expl. variations and much more), the PMF input and, if there are multiple PMF runs, variables across PMF runs. In the case of variables across PMF runs, there is also an option to use the clustering as part of the selection in the criteria panel. In a first step the input data has to be defined:

- Custom data: variables need to be defined by the user. Time series and profiles cannot be mixed, other than that there is no limitation on how much data can be combined. The single variables can be selected from the drop-down menu, similar to the one that can select the variables in the red analyze panel or can be typed in using the same style as in the blue criteria panel. After selecting/writing/ the variable name, press “write” and the variable appears in the list on the right-hand site.
- PMF input matrix: no further definitions required. The blacklisted input matrix will be clustered
- Variable across runs: only available if there is more than one PMF run available in the hdf5 file selected in the blue select PMF runs panel. In order to have full access to this function, the user first needs to define at least one image criterion and inspect it. E.g., to cluster all factor profiles of one factor over all PMF runs, the user first has to define the image criterion (profile, solution, factor x) for this factor, evaluate it and then inspect it, otherwise a reminder to do so will pop up. In the drop-down menu, all the image criteria the user can choose from are listed. Furthermore, in case the selected image criterion is based on profiles, the profiles can be scaled by ticking the checkbox for “scale factor profiles”. If checked, each variable is multiplied with a scaling factor $(m/z)^{sf}$ to give gradually more weight to patterns at the end of the mass spectrum. More details on this can be found in (Äijälä et al., 2017). Since the clustering is not NaN-stable, this option is only available for profiles and cycle plots as bootstrapped normal time series contains too many NaNs.

For custom data and the PMF input matrix option, the option selects based on class/time is available, where different classes or time points which should be considered for the clustering can be selected. In case of variables across runs, this option is not available in the clustering panel, instead the option should be used during the criteria definition.

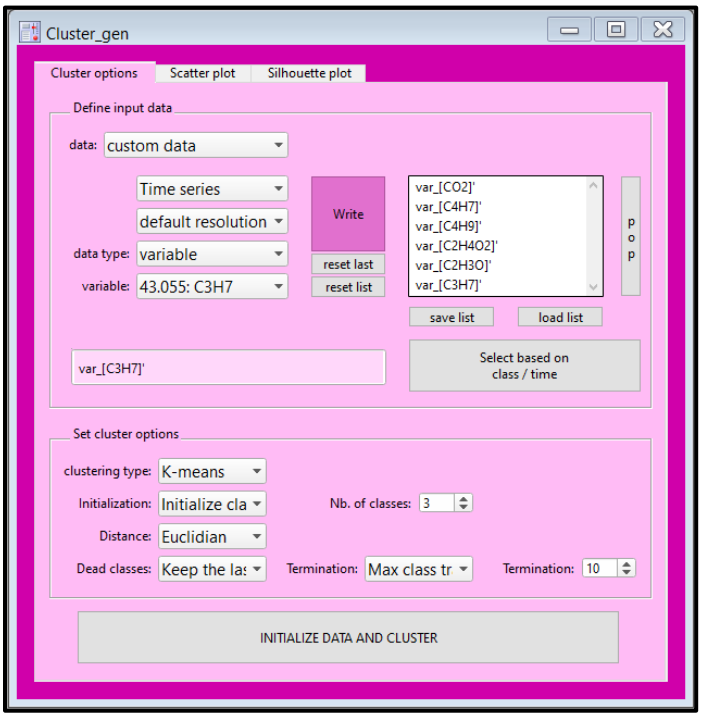


Figure 133 Setting the cluster options.

Next, the clustering options need to be set. SoFi offers two different clustering options: K-Means clustering and farthest point (FP) clustering. For clustering variables across runs, only the option of FP clustering is available due to the restrictions in Igor.

FP clustering:

- Termination: Determines when the iterations stop
    - Numbers of clusters (default): Terminates the calculation when the number of clusters reaches the specified value
    - Intercluster radius: Terminates the calculation when the maximum distance is less than or equal than the defined distance
- Number of classes/Intercluster radius: Specifies the termination number. The meaning of the number is determined by the termination method. If number of clusters are selected, the calculation is terminated when the number of clusters reaches the specified value. Note that this termination condition is sufficient but not necessary, i.e., the operation can terminate earlier if the farthest distance of an element from a hub is less than the average distance

K-Means clustering:

- Initialization: Specifies the initialization method
    - Initialize classes using values of a random selection from the population

    - Initialize classes using values of a random selection from the population (default)
- Distance:
    - Euclidian: Distance is evaluated as Euclidian distance (default)
    - Manhattan: Distance is evaluated as the sum of the absolute values (also known as Manhattan distance).
- Dead classes: Specifies how the algorithm should handle "dead" classes, which are those that lose all members in a given iteration
    - Remove the class if it loses all members
    - Keep the last value of the mean vector in case the class might get new members in a subsequent iteration (default)
    - Assign the class a random mean vector
- Termination: Determines when the iterations stop
    - User-specified number of iterations
    - Continue iterating until no more than a fixed number of elements change classes in one iteration (default)
- Termination number: Specifies the termination number. The meaning of the number is determined by the termination method
- Number of classes: Sets the number of classes in the data

To cluster the data, hit the button “Initialize data and cluster”. Note: Clustering is not NaN-stable. So if there are any NaNs in the input matrix, the whole row will be removed, i.e., when clustering custom data and some external time series is chosen that has NaNs at certain time points, there time points will not be considered at all.

Estimate number of clusters

This option should help to determine the optimal number of clusters based on two different methods: the elbow method and the silhouette method.

<u>Elbow method</u>

SoFi runs the chosen clustering algorithm for 1 to 10 clusters. As a result, the sum of squared distances (SSD) for each cluster number is plotted. The "elbow" in the resulting graph represents the point where the rate of decrease in SSD slows down, suggesting the optimal number of clusters.

<u>Silhouette method</u>

SoFi runs the chosen clustering algorithm for 1 to 10 clusters and calculates the silhouette scores for each number of clusters (can be quite time consuming). The optimal number of clusters is typically identified as the one that maximizes the average silhouette score across all data points.

Warning: this estimation might take some time, depending on the amount of data that should be clustered.

9.4.2. Standard plot

When clustering variables across runs, the standard plot is an option to explore how the e.g. profile looks across the runs and where the cluster centers are located. The input data is displayed as crosses, the cluster centers for each variable as bold symbol. The input data can be color-coded based on which cluster the single runs belong to.

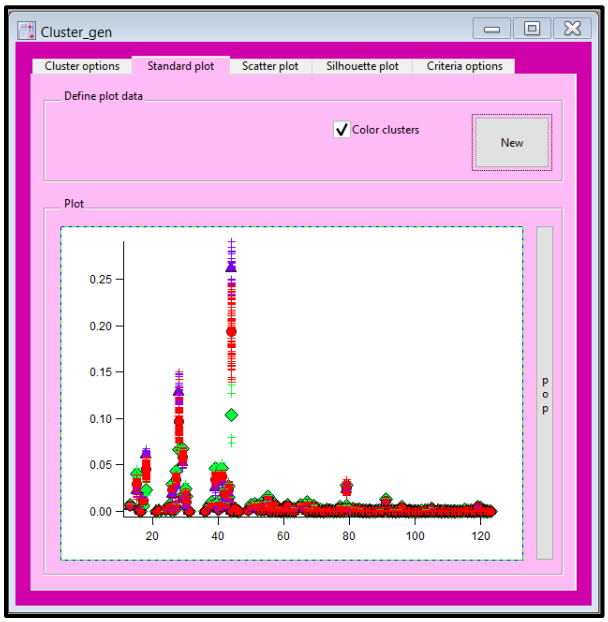


Figure 134 Standard plot. In this example, bootstrapped profiles were clustered. Cluster centers for each variable are indicated by bold symbols, the input data by crosses.

9.4.3. Scatter plot

In this tab, the user can get an idea of the different clusters based on a scatter plot. From a drop-down menu, the variable which is displayed on the x- and y-axis can be selected, respectively. Clicking the “New” button will create/update the graph. The data points from the input asre displayed as + and the cluster centers as **o**. To make the clusters more obvious, the box “color clusters” can be checked, so that the points are colored cluster-wise.

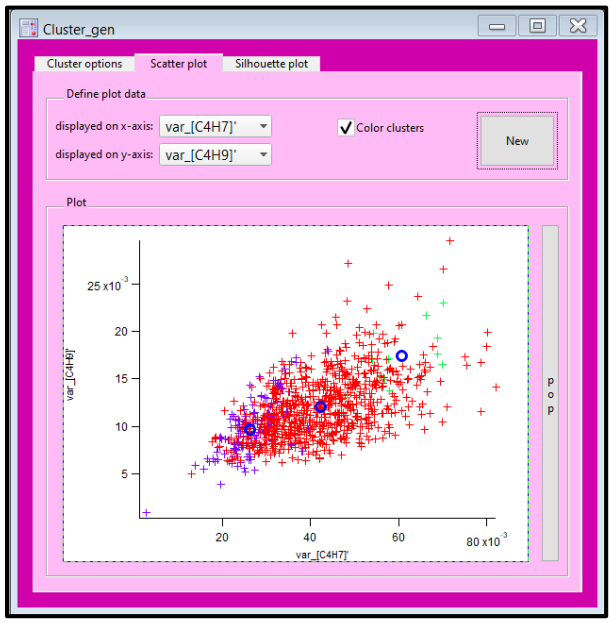


Figure 135 Scatter plot. Two variables can be selected to plot a scatter plot. The cluster centers are displayed in blue circles, the input data in crosses. In this case the input data is colored according to which cluster they belong.

9.4.4. Silhouette plot

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. The silhouette value of each point is defined as

$$s_i = f(x) = \begin{cases} 1 - \dfrac{a_i}{b_i}, & a_i < b_i \\ 0, & a_i = b_i \\ \dfrac{b_i}{a_i} - 1, & a_i > b_i \end{cases} \tag{32}$$

Where a_i is the average distance to all other objects of the same cluster A and b_i is the average distance to all other object of the closest neighboring cluster B. Silhouette values close to 1 mean that the within-cluster distance is much smaller than the between-cluster distance, so the point is likely grouped correctly, whereas silhouette values close to -1 implies that the point is on average more similar to the points of the neighbor cluster and likely misclassified.

The silhouette plot is created/updated by clicking on “new”. As the distance of each point to all neighbors in the same cluster as well as in all the other clusters has to be calculated, this could be time-intense for large data sets. After the calculation, the graph and the average silhouette value are automatically updated. Each group is represented by a separate color which corresponds to the same color/group shown in the standard and scatter plot. For further information, i.e., average silhouette value and number of points for each cluster, the user can select a cluster from the drop-down menu and the values will be updated for the selected cluster.

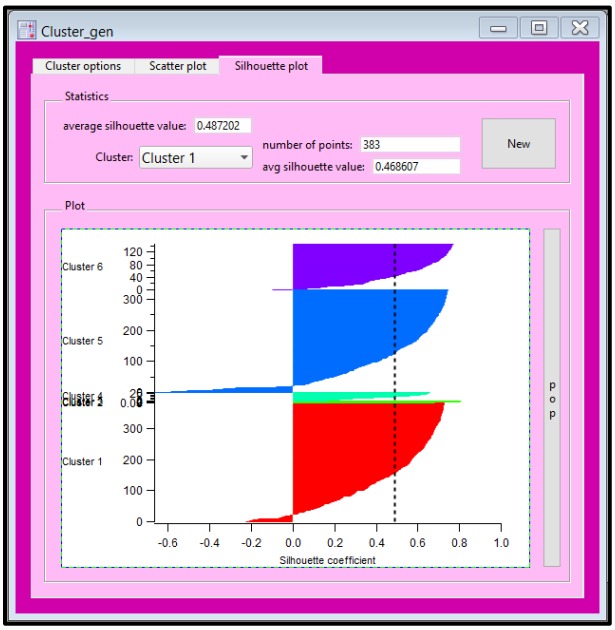


Figure 136 Silhouette plot. The silhouette value for each point in each cluster is displayed and color-coded according to the group. The black dotted line represents the average silhouette value. In the cluster drop-down menu, cluster 1 is selected and the number of points and the mean silhouette value for this cluster is displayed.

9.4.5. Criteria option

If variables across PMF runs, i.e, data from an image criterion from the criteria approach, are clustered, there is the option to select/unselect PMF runs based on the clustering results. A list of all the clusters is displayed. The silhouette plot needs to be inspected (plotted) to activate the cluster list. The user can select which clusters are regarded as “bad” solutions and should be excluded, i.e., not selected, in the criteria panel. By pressing “exclude” the selected cluster will be unselected in the criteria panel (Figure 137).

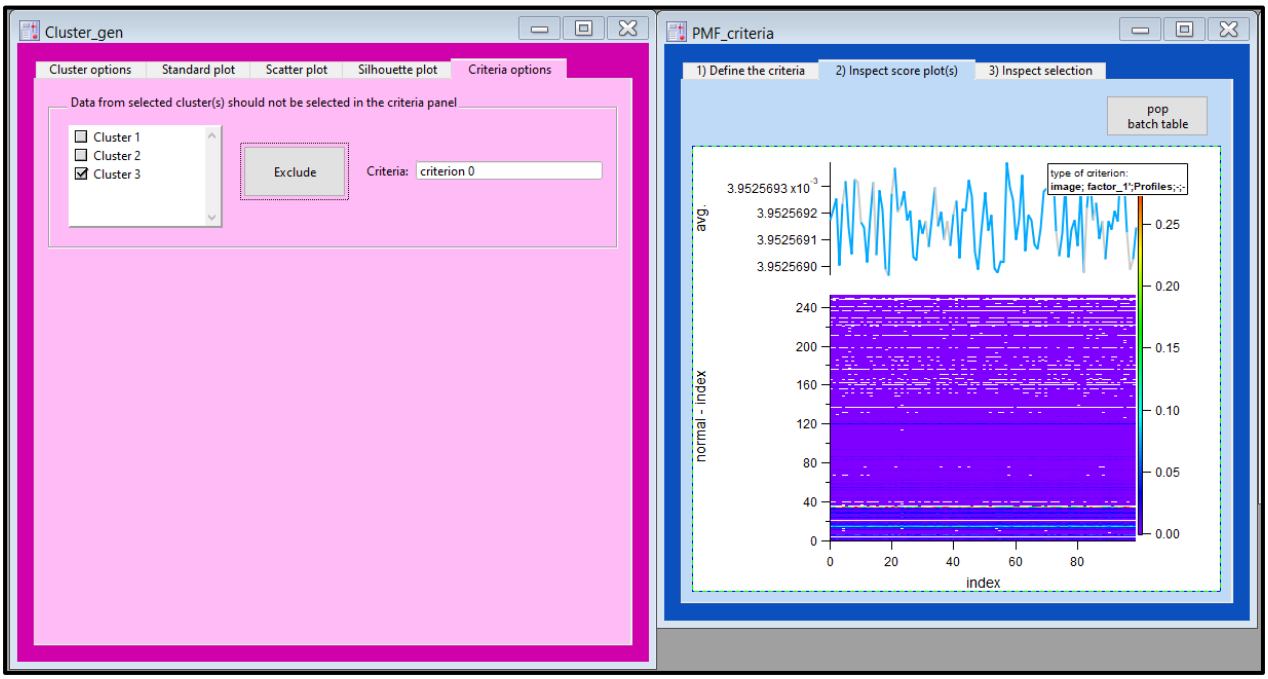


Figure 137 Unselect one or several clusters in the criteria panel. In this example all runs belonging to cluster 3 are unselected (grey) in the criteria panel.

10. Real-time module in SoFi (SoFi RT)

10.1. General considerations

10.1.1. What is SoFi RT

This section describes how to operate the SoFi real-time module, herein abbreviated as SoFi RT. SoFi RT performs automated and in real-time source apportionment analysis. SoFi RT is operational for separate and combined source apportionment runs of (Q or ToF)-ACSM and Xact. Moreover, there exist an automated import lane for BC data from the Aethalometer instrument where the source apportionment (Sandradewi et al., 2008) is automatically applied and the user will have access to black carbon data apportioned into biomass burning (BC_{bb}) and traffic (BC_{tr}). SoFi RT works only in SoFi 9 and does not need to be on the acquisition PC / laptop. The user is responsible for setting up a cloud account from a cloud provider, e.g., dropbox, onedrive, icloud, etc., that guarantees synchronized data folders between the acquisition PCs/laptops and the PC/laptop where the source apportionment analysis takes place (see Figure 138).

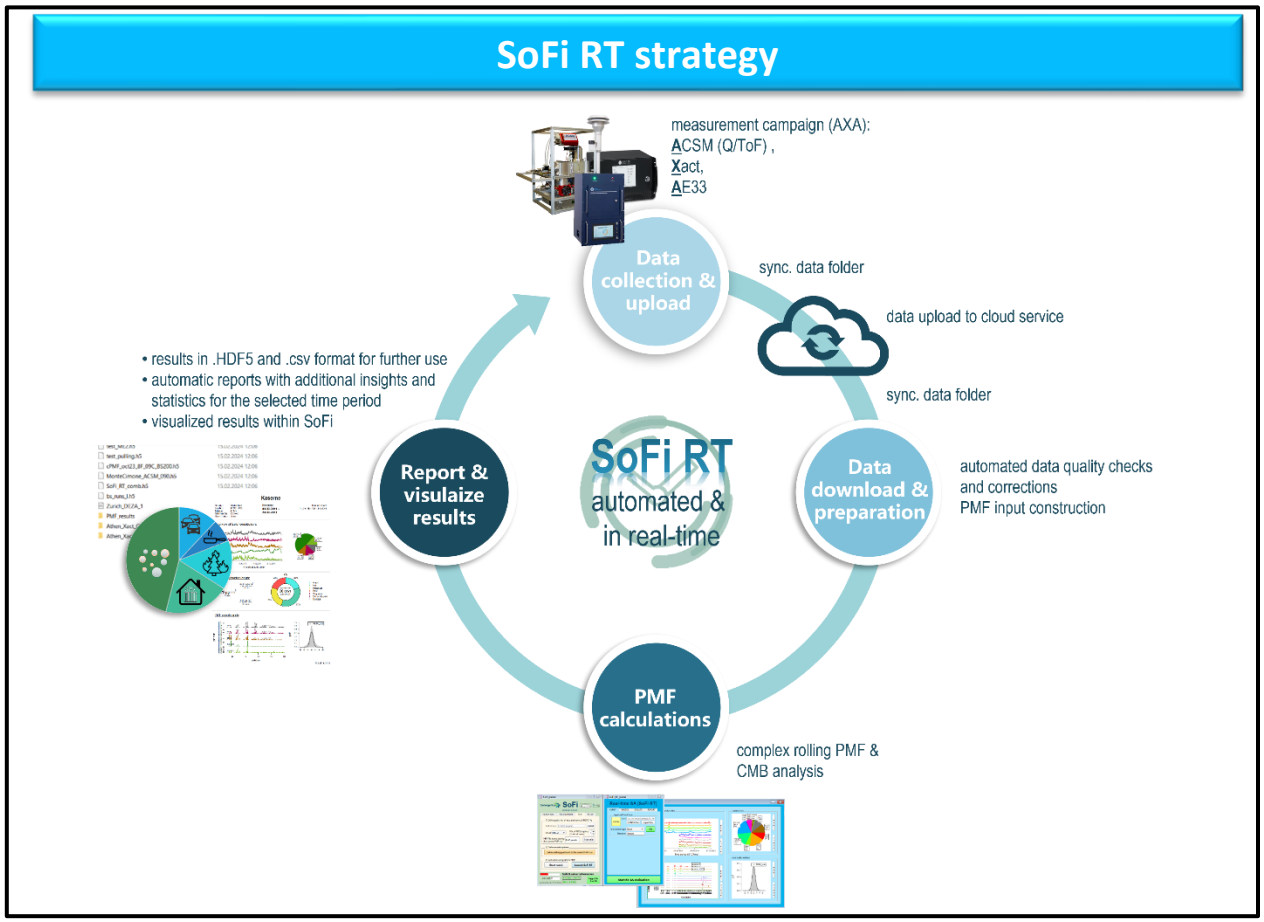


Figure 138 General scheme of working principle for data acquisition in SoFi RT.

10.1.2. Working principle

The data collected by the ACSM, Xact or Aethalometer is automatically processed (corrected and calibrated) and made ready for PMF by the manufacturer of the instrument (PMF input files). This is already the case for ACSM and for Xact data (both commercial types of instruments). When, these PMF input files are present in a synchronized folder, the same data is also available for SoFi RT, and the source apportionment (see chapter 2) can be conducted in real-time. SoFi RT performs a fully automated rolling PMF, where the user has full control over a multitude of correction and model settings that will be described in detail further down. SoFi RT performs two types of automated source apportionments, i.e., a more sophisticated one (rolling PMF), where several PMF runs are typically performed to assess the PMF error based on statistical and rotational uncertainty and a simpler one using the chemical mass balance (CMB) to get in real-time source apportionment results for the latest scans. Rolling PMF is carried out on a small and user-scalable input window with a length of e.g., 7-14 days and is moved on e.g., a daily basis, as the measurement proceeds. In between, CMB performs updates of the source apportionment, whenever a new measurement scan enters the synchronized folder. This allows the user/authority to be informed in real-time about changes in concentration of the various sources (see Figure 139, from Chen (2022)).

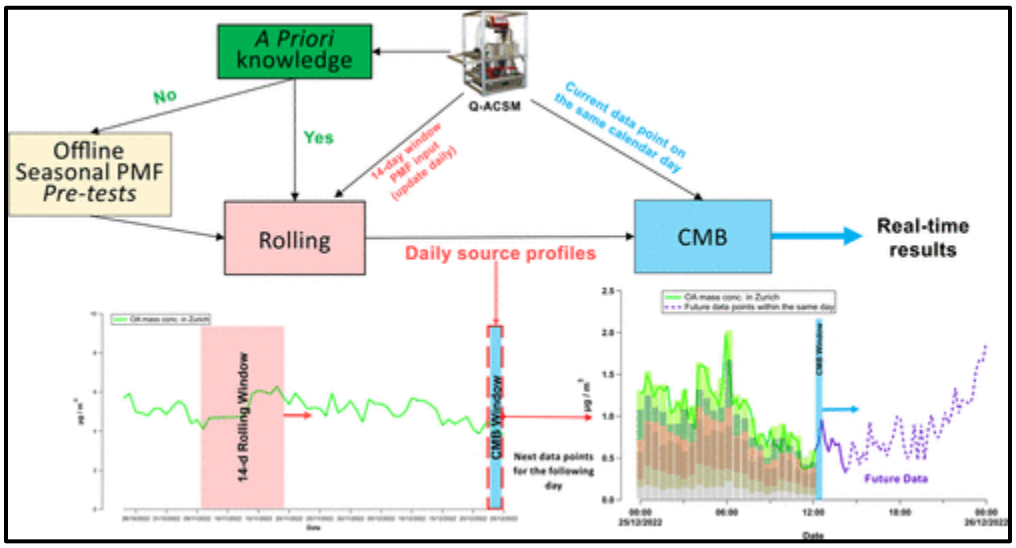


Figure 139 General scheme of working principle between advanced rolling runs and simple daily CMB runs in SoFi RT.

For the proper functioning of SoFi RT, all advanced functionalities of SoFi that are only available in SoFi Pro are necessary. That is why a SoFi RT license also contains a SoFi Pro license. SoFi RT produces source apportionment results that are visualized in a separate dedicated panel. However, all SoFi RT results, those from rolling PMF as well as those from the CMB runs, can always and anytime be inspected in detail using all features of SoFi Pro. This makes SoFi RT a powerful software module that remains fully in the hands of the user/modeler!

The next lines describe all necessary technical steps for successfully launching SoFi RT and performing semi-automated source apportionment with SoFi RT.

10.2. **Configuration file**

For real-time source apportionment there is the possibility to use a configuration file that stores all necessary information for setting up SoFi RT. The configuration file comes as RT_config.txt file and can be downloaded from our repository site.

Here are briefly the steps to have SoFi RT running using a configuration file:

1) Launch SoFi on a new IGOR experiment.
2) Go to the third tab "Run" and click on the blue button "Call SoFi RT" to launch the RT module.
3) Click on the button "CONFIG file" to select the configuration file.
4) The name can be changed, as the user is prompted to select the configuration file on the hard disk.
5) Click on the button "Start RT-SA evaluation"

Do not delete any lines from the configuration file but rather leave it blank after the :.

This procedure avoids all next manual steps from 10.3 to (not including) 10.5. Please read further details about inspecting and saving the SoFi RT results under 10.5.

It follows all lines of the configuration file:

[0-4]	File Information	
[5]	PATHS	empty
[6]	path of executable	e.g., "C:\ME2 engine" (see 10.3.2)
[7]	path of input files	e.g., "D:\OneDrive\Desktop\Input\ACSM" In case of combined, paths MUST be semi-colon separated (information on first tab of SoFi RT panel (see 10.4.1))
[8]	path for scan-wise results (CMB)	e.g., "D:\OneDrive\Desktop\Input\CMB_results" information on third tab of main SoFi panel (see 10.3.2)
[9]	path of factor profiles from file	e.g., "D:\OneDrive\Desktop\profiles.itx" If this line is empty, SoFi will search in database on our server for the name(s) of the factor profile(s) to be constrained (see line [23] below)
[10]	path of BC input files	e.g., "D:\OneDrive\Desktop\Input\AE33"

		This folder MUST contain daily .csv files generated through our batch file (see section 10.8).
[11]	path of other external data:	tbd
[12]	empty	empty
[13]	NAMES	empty
[14]	name of HDF result files (large files)	“test_config” name of the HDF file (see 10.3.2)
[15]	name of txt result file (single scans)	“CMB_result” prefix of the txt file containing the results for each scan (see 10.4.3)
[16]	name of prefix for input filename	e.g., “RTdata” In case of combined, it MUST be semi-colon separated
[17]	name of prefix for BC input filename	normally it is empty, but this can be changed here
[18]	name of prefix for other external data	tbd
[19]	empty	empty
[20]	DATA TYPE	empty
[21]	data type	“AMS UMR / Q-ACSM (without m/z 19 and 20)” information on first tab of main SoFi panel (see 10.3.1)
[22]	AMS specific	“exclude CO2 related variables” information on first tab of main SoFi panel (see 10.3.1)
[23]	instrument type	“ACSM” In case of combined, it MUST be semi-colon separated information on first tab of SoFi RT panel (see 10.4.1) prefix of the filename containing the input data (see 10.4.1)
[24]	empty	empty
[25]	MODEL OPTIONS	empty
[26]	restart RT	“0.1” Information on second tab of SoFi RT panel (see 10.4.2)
[27]	factor start	“4” Information on main settings from SoFi panel (see 10.3.2)
[28]	factor end	“4” Information on main settings from SoFi panel (see 10.3.2)
[29]	factor names	“HOA,COA,OOA1,OOA2” Information on main settings from SoFi panel (see 10.3.2)
[30]	nb. of iterations	“2”

[31]	resampling	Information on main settings from SoFi panel (see 10.3.2) “enabled”
[32]	rolling	Information on main settings from SoFi panel (see 10.3.2) “enabled”
[33]	window length	Information on main settings from SoFi panel (see 10.3.2) “5”
[34]	window shift	Information on main settings from SoFi panel (see 10.3.2) “1”
[35]	between PMF	Information on main settings from SoFi panel (see 10.3.2) “CMB after each scan”
[36]	stats tests (corr)	Information on second tab of SoFi RT panel (see 10.4.2) “disabled”
[37]	stats tests (comp.)	Information on second tab of SoFi RT panel (see 10.4.2) “disabled”
[38]	a value constraints	Information on second tab of SoFi RT panel (see 10.4.2) “enabled”
[39]	name of constraints over pr	Information on main settings from SoFi panel (see 10.3.2) “HOA,Paris,0.1;BBOA,avg,0.3” These are the names of the fingerprints stored on our repository site. Note the structure: [name],[additional information on the site],upper a-value;[name],[additional information on the site], etc. In case additional information on the site is provided, then SoFi searches for the profile from our database. Otherwise, in case a profile file path is provided, then SoFi will search therein.
[40]	name of constraints over ts	“” Same as for [23] but for the time series constraints.

Note that the positions are hard-coded, hence, don’t remove, swap, or do any other operations on these lines. Also, the structure:

“[txt]: [space] **[txt]**” is hard-coded.

Don’t change anything or remove any spaces apart from the txt information at the end of each line (**[txt]**).

10.3. First steps for manual process

10.3.1. Launching SoFi and preparing the input data

1) Launch SoFi on a new IGOR experiment.

Currently, SoFi RT requires a dummy input file with the same size of the variables used for PMF. This file is directly generated for Q-ACSM data as is stored in the synchronized PMF output folder and has the name: “RTdata_template”. In near future, a database for ACSM/AMS and Xact source profiles will be made available on the server of Datalystica and this step will be made fully automated. For Xact data for the moment, contact us directly to get a reasonable template that you can modify further. Or in case you are used to IGOR, generate four waves, i.e., a data matrix, an error matrix, the time series wave and the numerical wave of the variables. Make sure that the dimension of the columns of the data and error matrix is matching that of the numerical variable wave. The data and error matrix must contain only one row and the entry must be NaN for that row in the data and error matrix as well as for the time series. Finally, store this file as igor txt file and then import it back into SoFi with the PMF import panel (Figure 140 and consult 4.1.1 for more details).

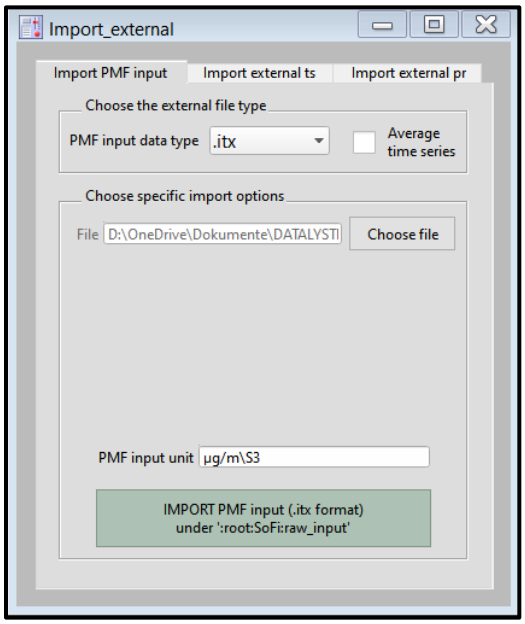


Figure 140 Import panel for the PMF input from an IGOR txt file.

2) Link the internal input strings, i.e., “Data matrix”, “Error matrix”, “Tseries nb”, “Variables nb”, “Variables txt” to the proper waves by clicking on “PMF input”. Similarly, use the import section to import potential fingerprints that should be constrained (Figure 141 and consult 4.1.3 for more details).

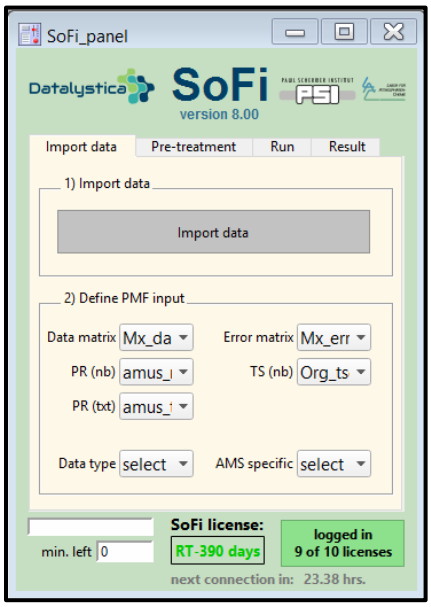


Figure 141 Main SoFi panel, import tab. The “Import data” button allows automatic import of the PMF input and external time series and profiles. Under “Define PMF input” the user has to define the PMF input.

3) Click on the button “initialize PMF input & check for missing data” to prepare the PMF input structure that is then passed to ME-2 (Figure 142).

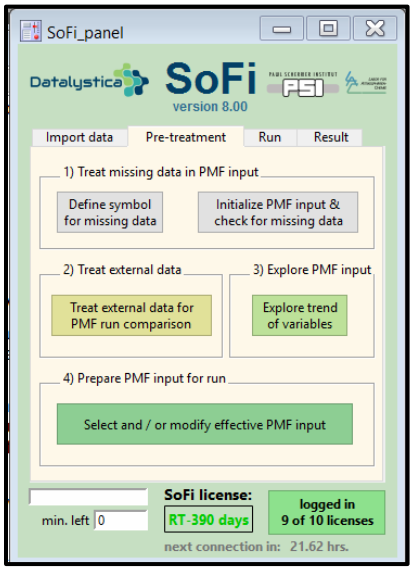


Figure 142 Main SoFi panel, pre-treatment tab.

4) Click on the button “Treat external data for the PMF run/result comparison” and extract the variables from your external profiles for the factor profile constraints (Figure 143 and consult 5.2.1 for more details).

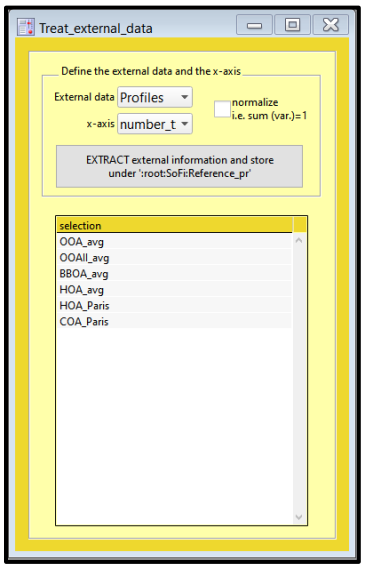


Figure 143 Extraction of reference profiles.

<u>!!!When importing mass spectra for AMS or ACSM data, make sure that these externals match the fragmentation table of the PMF input. The extract procedure in SoFi will fail otherwise!!!</u> Importing reference profiles via SoFi will take care of this automatically though.

10.3.2. Define main model settings

10.3.2.1. **Path, number of executables and HDF filename**

5) Define the location where the ME2.exe file resides by pressing the button "folder of executable" on the tab PMF run.

6) Increase the number of executables to its maximum value. This will increase the performance during computation.

7) Define a proper name for the PMF run under "HDF5 file name storing the current PMF run". If this name already exits, SoFi prompts the user to try with another name (Figure 144 and consult 6.1 for more details).

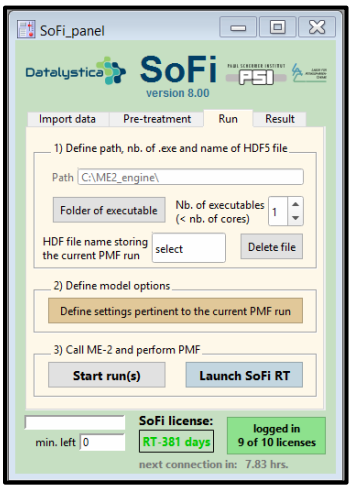


Figure 144 Main SoFi panel, PMF run tab.

<u>!!!The "Delete file" button removes the specified HDF5 from the OS. Despite a last warning prompt, the user should pay attention not to delete important information. SoFi does not control hidden back-ups of the HDF5 files!!!</u>

10.3.2.2.**Define number of factors and repeats**

Click on the orange button "Define settings pertinent to the current PMF run". This pops the panel governing all PMF settings (Figure 145). The user is referred to 6.2 for an overview over all supported PMF model settings.

Define number of factors and the number of repeats.

"starting factor" defines the minimal number of factors (p) investigated by the model.

"ending factor" defines the maximal number of factors (p) investigated by the model.

"nb. of PMF calls" is the number of repeats. In case bootstrap resampling is for example enabled, this option generates each time a new resampled PMF input.

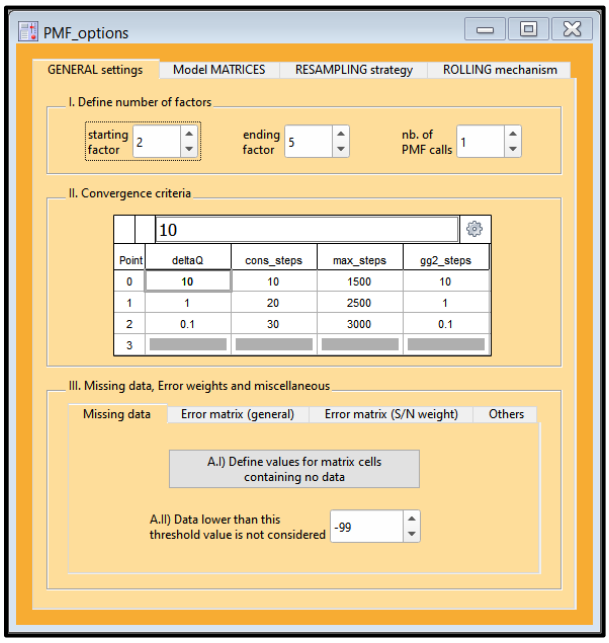


Figure 145 PMF settings, tab “General settings.

10.3.2.3.**Define seed for primary factor profiles**

Go to the 2nd tab of the orange panel to define profiles of known factors based on the previously imported external profiles. Use the following options (Figure 48 and consult 0 for more details): from vector (fv), all variables of profile and normalize.

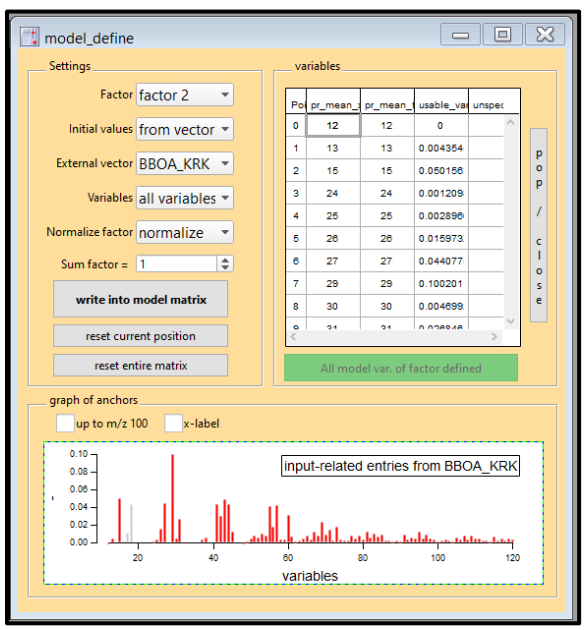


Figure 146 Panel for initial model values from vector. HOA_Paris profile normalized to 1 is going to be written on factor position 1 for the profile matrix F.

10.3.2.4.**Constrain a value for primary factor profiles**

Constrain the defined factor profiles with the a value technique.

To enable the *a*-value constraints, use the dropdown menu below the table. The *a*-values are only used by the ME-2 solver when the button “A-value constraints are enabled” turns green (Figure 147).

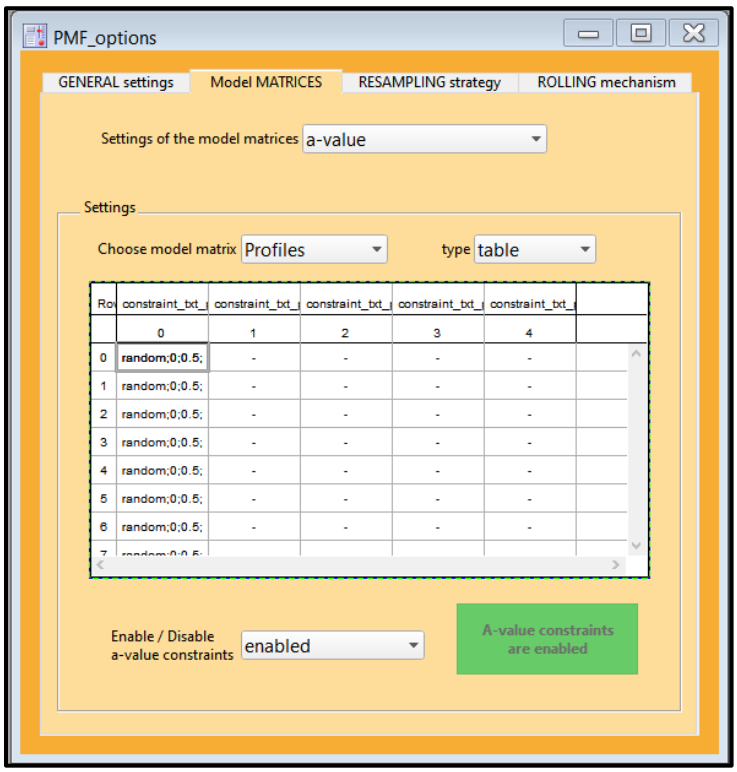


Figure 147 PMF settings, tab “Rotational ambiguity”, *a*-value.

10.3.2.5.**Enable random resampling**

Go to the third tab and enable random resampling (bootstrap strategy) by selecting the option “Enabled” and use the option unblocked (Figure 148). For more details on the bootstrap strategy consult the SoFi manual and the references therein under section 6.4).

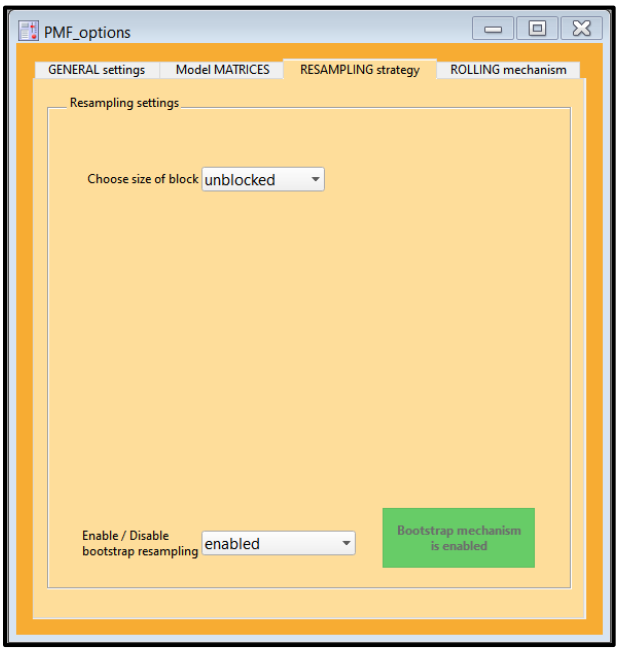


Figure 148 PMF settings, tab “Statistical error propagation”

10.3.2.6.**Enable rolling strategy**

Go to the fourth tab and select the option “Enabled”, together with window width and window shift (Figure 149). For more details on the rolling strategy consult the SoFi manual under section 6.5).

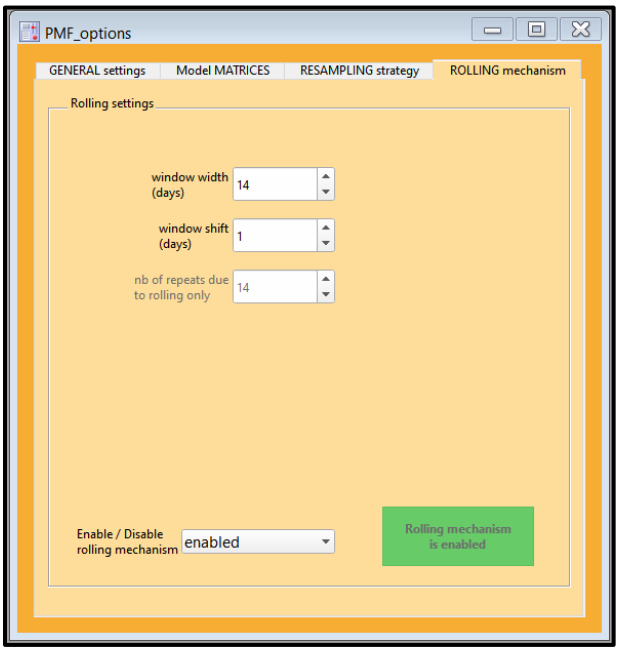


Figure 149 PMF settings, tab “Rolling mechanism”

10.3.3. Call SoFi RT

Click on the blue button “Call SoFi RT” to launch the RT module.

10.4. SoFi RT module

10.4.1. Define INPUT opt.

Figure 150 shows the first tab of the SoFi RT module for the model settings.

Input path and type

“offline / online button”	SoFi RT can be used online, i.e., the data is imported into IGOR in realtime or offline, the data is already present in IGOR as full PMF input. Then, SOFi RT performs an offline simulation of SoFi RT.
“Path”	Path must point to the synchronized folder containing the input data.
“type of instrument”	ACSM, Xact, Aethalometer. The user can turn them on using the “on” / “off” button independently.

Specific instrument options

“instrument options”	For ACSM data filtering options based on pressure and detection limit are supported. The filename is already hard-coded and comes directly form the addon of the acquisition software and does not require any further attention form the user. For Xact data, there exist two commercial instruments with different output file names. The user must pass the stem name of the sample file, e.g., “data”. For Xact 625, the elements are stored in three different files, “data1_data&time”, “data2_data&time” and “data3_data&time”. The user must only pass “data”. In case BC is enabled and BC data is automatically imported and treated in SoFi, black carbon data is apportioned into biomass burning (BC_{bb}) and traffic (BC_{tr}). The user has the possibility to enable or disable automatic correlation between possible traffic or biomass burning source with BC_{tr} and BC_{bb}, respectively for single and / or combined instrument settings. Default is that for ACSM data only, the correlation between the hypothetical HOA/BBOA and BC_{tr}/BC_{bb} is evaluated, respectively.

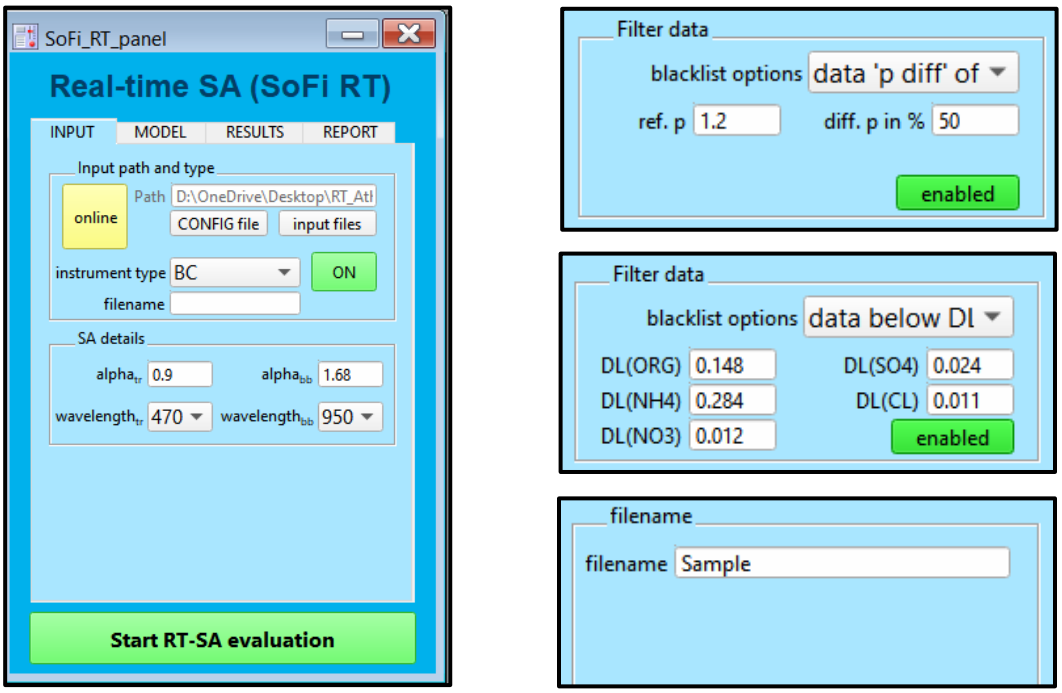


Figure 150 first tab of the SoFi RT panel (left) together with the options for the filtering mechanism based on pressure (top right) and detection limit (middle right) for the ACSM as well se the filename option for the Xact instrumentation (bottom right).

10.4.2. Define MODEL opt.

Figure 151 shows the second tab of the SoFi RT module for the model settings.

<u>Rolling settings</u>

"initialize PMF after…"	After this number of days, the automated rolling PMF analysis is started. The window width defined under rolling parameters (see 10.3.2.6) is linked to this value.
"PMF update every…"	After this number of days, the automated rolling PMF analysis is updated. The window shift defined under rolling parameters (see 10.3.2.6) is linked to this value.
"SoFi RT restart rate…"	tells how frequently SoFi RT is restarted. This is typically the scan rate of the instrument to get the latest scan and update the CMB result (in combination with the following option)
"between PMF updates"	No update (SoFi RT updates automated rolling PMF defined on top only).
	CMB after each scan (CMB is performed in addition after each new scan).

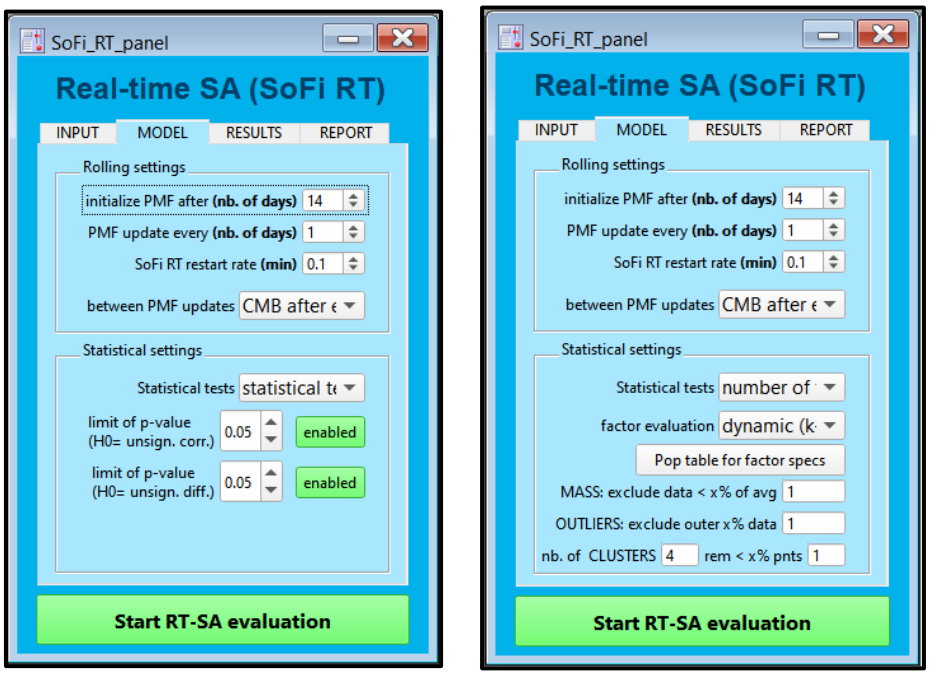


Figure 151 second tab of the SoFi RT panel showing the options for the statistical T-tests below (left) and for the dynamic factor evaluation below (right).

<u>Statistical settings</u>

"Statistical tests"	T-test limit of p-value for insignificant correlation or difference with other factors can be enabled or disabled
"factor evaluation"	default (from orange panel), the user should keep a static number of factors; dynamic (factor evaluation per window) based on cluster analysis (see below for more details).
"Pop table for factor specs"	opens the table where the information for the dynamic differentiation can be passed to SoFi RT
"MASS: exclude data…"	data below x % from the average value in a rolling window are excluded for Cluster analysis.
"OUTLIERS: exclude…"	data in the outer x percentiles in a rolling window are excluded for Cluster analysis.
"nb. of Clusters"	number of Cluster centroids used for the subsequent linearity analysis
"REMOVE Clusters…"	Cluster centroids with less than x % of datapoints are removed from the final Cluster centroids list.

10.4.2.1. **Dynamic factor evaluation**

In case the user selects the option “static”, the number of factors should remain constant, as there is no automated analysis when performing PMF runs with different number of factors in SoFi apart from an unspecific averaging control mechanism. This control mechanism guarantees that there is no averaging between PMF runs with different number of factors by allowing averaging only over PMF runs with most frequent number of factors on a timepoint-by-timepoint basis.

In contrary, when selecting the option “dynamic”, there is an additional statistical clustering analysis on a rolling window basis that allows to vary the number of factors. This has the advantage that scanning over various number of factors can be avoided resulting in much less PMF runs than compared to classical manual PMF analysis, where the modeler must first perform PMF runs with different number of factors and then evaluate in *post*-analysis which factor solutions are most the adequate ones over the measurement campaign. Clicking on the button “Pop table of factor specs” (Figure 151) opens the table where the information for the dynamic differentiation can be passed to SoFi RT. The main analysis is performed on a user-defined two-dimensional space based on two key variables that are relevant for a factor separation at the time.

<u>Example with *m/z* 44 and *m/z* 43 for ACSM data</u>

In case the measured data scatters in this space with a triangular shape, two factors are at least required to capture the variability of the data. In their fingerprints these two factors will most likely have their *m/z* 44 to *m/z* 43 ratios on the edges of this triangle (red points in Figure 152, left). On the contrary, when data is aligned, one factor with its *m/z* 44 to *m/z* 43 ratio on the edge of the line is sufficient to capture the variability (red point in Figure 152, right). The shape of the data is captured by performing the farthest point clustering algorithm with at most 10 clusters. 10 cluster centroids as upper limit has been selected has shown to perform rather well based on previous ACSM studies for two stations. Too little cluster centroids would make the line rather unstable, and too many points would very often lead to statistically significant correlations resulting in static number of factors. The ideal number of factors is determined dynamically from the algorithm based on intercluster distance analysis. However, the effective number of clusters in not so crucial, as a final linear regression on the cluster centroids (in black) is performed to verify whether the cluster centroids are aligned or not (Figure 152). The acceptance of the linearity of the cluster centroids is based on a linear correlation test using a customizable p-value threshold (0.05 as default value).

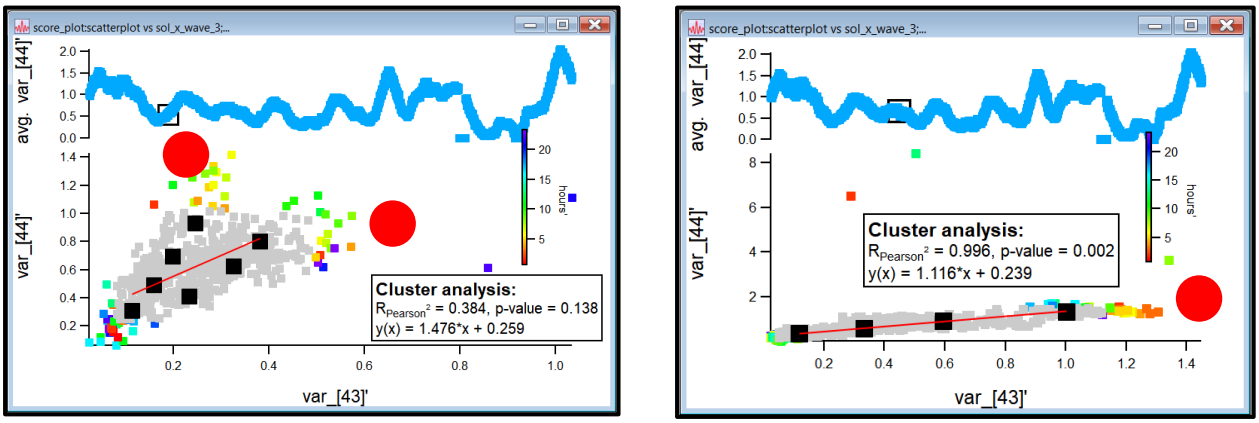


Figure 152 f44 vs f43 space with measured points (gray) and clustered centroids (black). The cluster centroids are not aligned and therefore two factors are required to capture the variability of the data (left), whereas they are aligned, and one factor only is required for the other example (right).

For the above example, the user would pass the information 44 vs 43 (44 and 43 refer to the txt information in the variable txt wave) in the first column of the popped table (Table 10). Similar cluster analysis can also be performed on other factors, e.g. 55 vs 57 for COA or 44 vs 60 for BBOA. The second column of the table refers to the name of the factor that can dynamically be disabled/enabled. In the event, where the cluster centroids are statistically aligned, this factor is disabled (in this example LO-OOA would be missing and only one OOA is modeled). Otherwise, LOA-OOA is present and then a PMF runs are performed for this rolling window with MO-OOA and LO-OOA (see also 6.3.1). The factor names must be defined in the PMF options panel (consult 7.5.3)

Table 10 options for clustering algorithm (left) and full list of factors (right) from the PMF options panel

RT_fact_specs:RT_factor_specs

R0 C0 | 55 vs 57

Row	RT_factor_specs[]	RT_factor_specs[]	RT_factor_specs[]	RT_factor_specs[]
	0	1	2	3
0	55 vs 57	COA		
1	44 vs 60	BBOA		0.003
2	44 vs 43	LO-OOA		
3				

factor_specs_table:mo...

HOA

Point	model_txt_factor_s
0	HOA
1	COA
2	BBOA
3	OOA;MO-OOA
4	LO-OOA
5	

The third and the fourth column of the table can be used to pass lower limits in terms of fractions, for the key variables. Data below this threshold is not used for the clustering analysis. The third column is reserved for the lower limit for the first variable, in case a reasonable threshold is known form *a priori* information or from literature data. For example, 0.003 (Cubison et al., 2011) in the fourth column in Table 10 stands for the lower limit of *m/z* 60 / total OA (i.e., *f*60). Note that SoFi RT supports absolute and fraction entries in the table. For absolute values, use the

variable name as in the variable txt wave. When fraction should be used use a “f” in front of the variable name: *f*VARNAME.

The current settings from Table 10 represent the following possible factor combinations summarized in Table 11:

Table 11 overview over potential list of factors and effective number of factors based on Cluster centroids analysis.

Key variable space	Cluster centroids aligned	Effective factors
m/z 55 vs *m/z* 57	Yes	no COA, only HOA
	No	HOA and COA used
m/z 44 vs *m/z* 60	Yes	no BBOA, only OOAs
	No	BBOA and OOAs
m/z 44 vs *m/z* 43	Yes	OOA
	No	MO-OOA and LO-OOA

The various combinations from the effective factors in Table 11 result in the possible scenarios that are evaluated by SoFi RT for each rolling windows:

- 2-factor solution: HOA+OOA
- 3-factor solution: HOA+MO-OOA+LO-OOA
  HOA+BBOA+OOA
  HOA+COA+OOA
- 4-factor solutions: HOA+BBOA+MO-OOA+LO-OOA
  HOA+COA+BBOA+OOA
  HOA+COA+MO-OOA+LO-OOA
- 5-factor solutions: HOA+COA+BBOA+MO-OOA+LO-OOA

Note that for these settings HOA is always present and OOA is either split into MO-OOA and LO-OOA or remains OOA. The user has full control over which factors and how many factors could be potentially evaluated and on their combinations, by defining which factors could be disabled based on the key variable plots under the options for the clustering algorithm (Table 10).

10.4.3. Define RESULT opt.

Figure 153 shows the third tab of the SoFi RT module for the graph options.

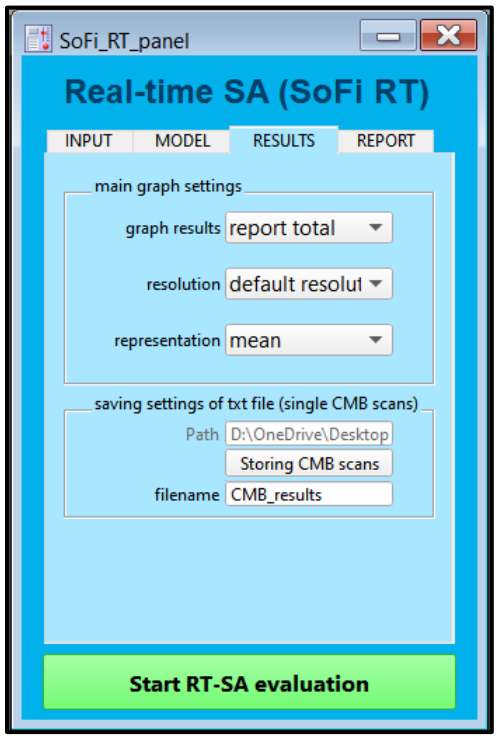


Figure 153 SoFi RT, graphs options tab

Main settings

"graph results"	report total (the entire PMF solution including the latest CMB updates), report PMF calls only (only the PMF solution), report last PMF call only (last PMF window), report CMB scans only (only the CMB runs form the actual day)
"resolution"	default, hourly, daily, weekly, monthly and yearly resolution.
"representation"	several types of representations can be used, i.e., mean, mean&stddev, median, median&IQR, median&IQR&10_90%, median&IQR&10_90%& min_max

Saving settings of txt file (single CMB scans)

"Path"	shows the chosen path for storing the daily CMB txt files.
"Storing CMB scans"	select the path from the OS.
"filename"	prefix of the daily CMB txt files.

Figure 154 shows how the daily CMB files look like. The Filename is composed of the user-defined prefix and contains the current date, here "CMB_results_2024_03_07.txt". SoFi copies the current date and time together with the modelled source contributions, here: HOA, COA, BBOA, MO-OOA, LO-OOA (ACSM) and traffic, dust, coal industry, biomass, chloride and sulfates

(Xact) for a combined PMF run. In this example the BC lane was also active and hence BC_tr and BC_bb are also imported and copied into the file.

*CMB_results_2024_03_07 - Editor

Datei Bearbeiten Format Ansicht Hilfe

date&time	HOA	COA	BBOA	MO-OOA	LO-OOA	traffic	dust	coal_industry	biomass	chloride	sulfates	BC_tr	BC_bb
2024/03/07 00:00:00	nan	nan	nan	nan	nan	0.0733	0.0102	0.157	0	0.0025	0.516	0.0425	0.144
2024/03/07 01:00:00	nan	nan	nan	nan	nan	0.055	0.00286	0.159	0	0.00201	0.467	0.0909	0.121
2024/03/07 00:01:22	0.0292	0.693	0.364	0.83	0.825	nan	nan	nan	nan	nan	nan	0.0929	0.12
2024/03/07 00:11:28	1.23e-11	0.288	0.0917	0.422	0.72	nan	nan	nan	nan	nan	nan	0.106	0.122
2024/03/07 00:21:28	0.0256	0.225	0.103	0.355	0.704	nan	nan	nan	nan	nan	nan	0.163	0.103
2024/03/07 00:31:28	0.0373	0.281	0.04	0.35	0.746	nan	nan	nan	nan	nan	nan	0.0679	0.13
2024/03/07 00:41:28	0.0539	0.203	0.127	0.336	0.791	nan	nan	nan	nan	nan	nan	0.0471	0.131
2024/03/07 00:51:29	0.022	0.257	0.114	0.321	0.789	nan	nan	nan	nan	nan	nan	0.0567	0.124
2024/03/07 01:01:28	0.0284	0.276	0.0584	0.395	0.717	nan	nan	nan	nan	nan	nan	0.114	0.103
2024/03/07 01:11:28	0.057	0.25	0.0161	0.338	0.673	nan	nan	nan	nan	nan	nan	0.109	0.0951
2024/03/07 01:21:28	0.0776	0.353	0.0753	0.384	0.81	nan	nan	nan	nan	nan	nan	0.078	0.105
2024/03/07 01:31:28	0.046	0.389	0.121	0.424	0.752	nan	nan	nan	nan	nan	nan	0.0601	0.112

Figure 154 Section from a possible CMB txt file.

10.4.4. Report

Figure 155 shows the fourth tab of the SoFi RT module for the report. The user can choose which quantities should be shown in the report.

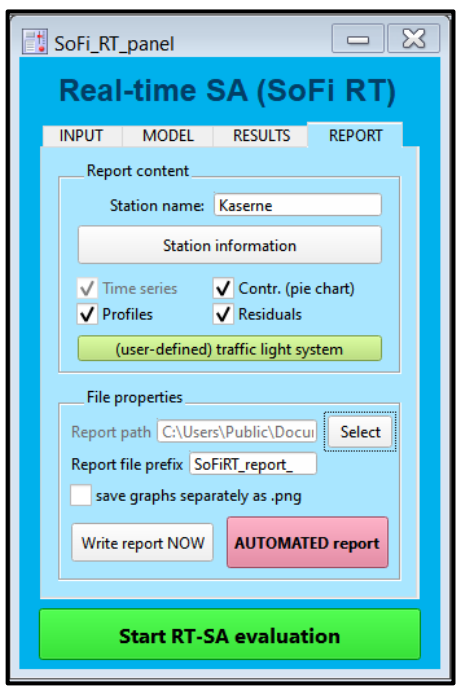


Figure 155 SoFi RT, report options tab

<u>Report content</u>

"Station name" Enter the name for the sampling site

"Station information" The user can enter station information, such as country, coordinates, altitude, station code and classification (urban, rural, etc.) and these information will result in the report file.

Visualized results	The following results can be visualized in the report: Contributions (pie charts, by default), time series, profiles, residuals
“air quality limits”	SoFi allows to summarize the average daily concentration into a grouped donut chart. The limits for the different categories as well as the colors can be customized.

<u>File properties</u>

Figure 156 shows an example of the report file currently available in SoFi RT.

“Report path”	Path for the storage folder of the report(s).
“Report file prefix”	File name of the report. SoFi will automatically add the date and time to the defined name
“Write report NOW”	Generates the report for the current PMF result that has been imported and evaluated by SoFi RT.
“AUTOMATED report”	Weekly or monthly report automatically generated and stored on the OS (available soon).

The generated report will be saved as .pdf under the defined file path. Furthermore, the main graphs can be stored as .png files. If the checkbox is checked, then SoFi will generate a new folder under the defined file location with the current date as name, in which the graphs will be stored.

The top part of the first page of the report contains a general overview over the air quality, measured by the instrument(s). It includes the current pollution level as well as an overview over the last few days. Firstly, of the average concentration measured by each instrument as well as its data coverage. Secondly, a synchronized average from where all instruments measured simultaneously (averaged to one-hour averages). This includes an overview of daily averages, classified by the pollution level.

The bottom part of the first page of the report contains an overview of the sources that SoFi resolved. In the pie chart on the left-hand side is the current source contribution, on the right-hand side the contribution over the last few days (datapoints where all instruments measured simultaneously) shown.

On the second page, more details on the source apportionment, like factor time series (general and diurnal), factor profiles as well as the scaled residuals are presented.

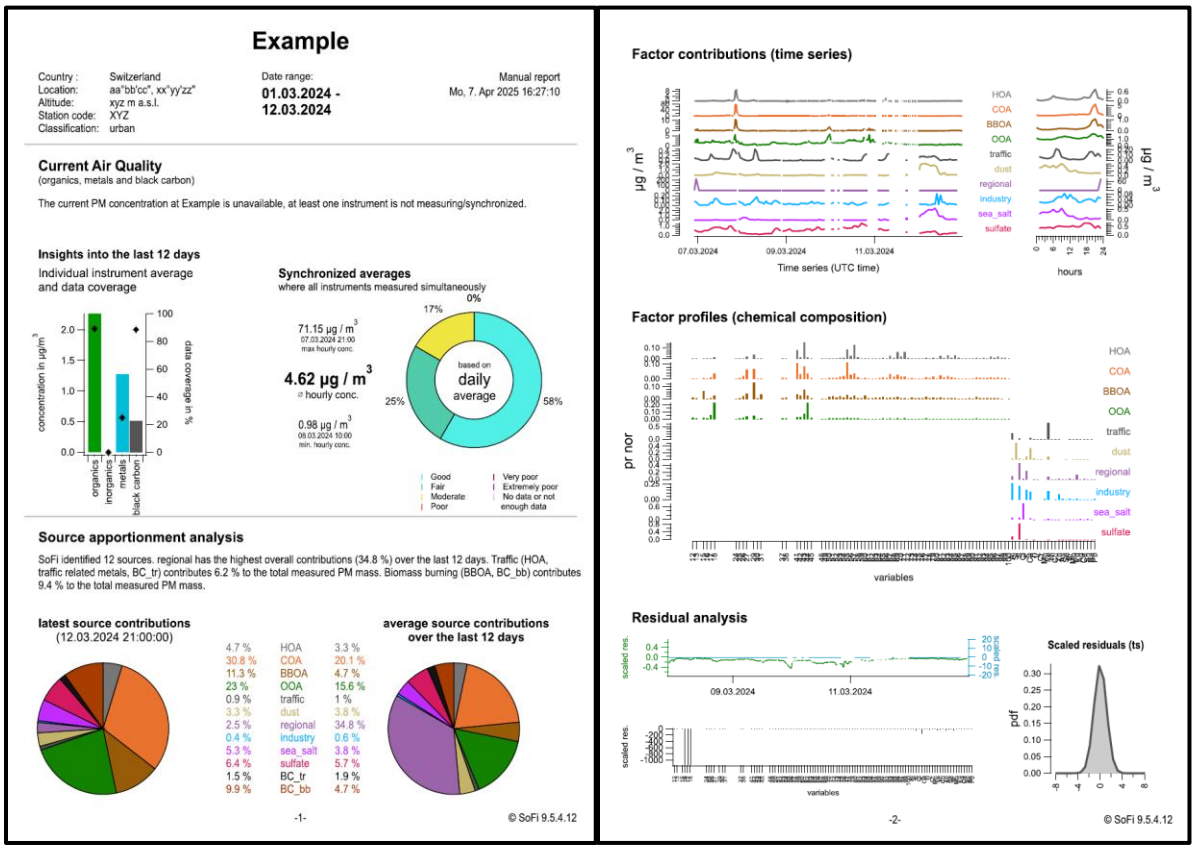


Figure 156 Example of the report file

10.5. **Launch SoFi RT**

10.5.1. Start/stop SoFi RT

The “Start RT-SA evaluation” button can be pressed form any tab of the SoFi RT panel for launching SoFi RT (Figure 153). After pressing on the green button, the SoFi RT result panel pops on the right side and SoFi RT will start running (Figure 157). SoFi RT is resume-stable and in case SoFi RT recognizes an existing and non-empty HDF file, the user will be prompted and can decide whether resume the run or reinitialize everything.

SoFi RT has a main endless loop with idle time therein as defined under “SoFi RT restart rate (min)” (10.4.2). During this idle time, SoFi RT can be stopped, and the results inspected or various model and/or graph options changed. When SoFi RT is evaluating and the color of the button is red and gray, it cannot be interrupted.

Start RT-SA evaluation **green button** (SoFi RT can be started), **red button bold text** (SoFi RT is running, but it can be interrupted), red button gray text (SoFi RT is running, it cannot be interrupted)

There is absolutely no harm in interrupting SoFi RT and restarting it. It is necessary, every time the user wants to inspect in more detail the overview plots on the SoFi RT graph panel, or if the user wants to modify the graph options on the SoFi RT settings panel.

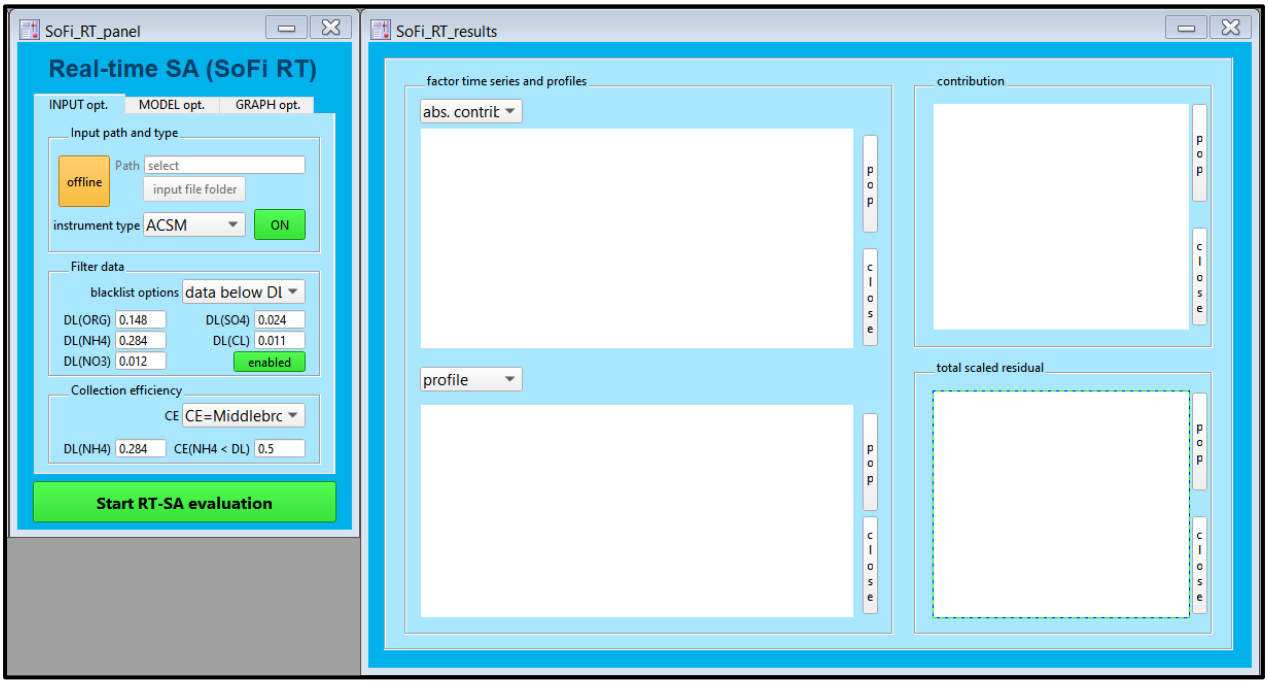


Figure 157 SoFi RT settings panel (left) with the result panel (right)

10.5.2. SoFi RT overview results

The user can inspect main overview plots on the SoFi RT result panel (Figure 159) after interrupting SoFi RT as just discussed.

<u>Graph options</u>

“factor time series”	abs. contribution, rel. contribution, residual, daily cycle
“factor profiles”	profile, residual
“contribution”	pie charts of the factor contributions
“total scaled residual”	scaled residual for all points. It’s supposed to be unimodal around 0, possibly Gaussian-shaped.

When starting SoFi RT for the first time, a folder named SoFi_RT including a subfolder Variables are created under “hard drive (e.g., C):\ME2 engine\” on windows. The folder SoFi_RT\Variables contains a hard-coded IGOR txt file called “SoFi_RT_criteria”. This file will be automatically generated form SoFi and contains the list of user-defined criteria for extracting the proper PMF runs during a SoFi RT run. At the end of the first SoFi RT cycle, i.e., when all repeats from the first automated rolling window are done, before SoFi pulls the proper PMF runs for averaging, if no

such file exists, SoFi automatically generates this file with a dummy criterion that monitors the averaged contribution of factor 1 and considers all PMF runs for the PMF average. The user is supposed to interrupt SoFi RT during the next idle period and to change the criteria using the criteria panel (see 7.3).

To get a tone warning for when the idle time starts, the user can enable the “beep” present in IGOR from the main settings menu of SoFi (see Figure 158 and read section 3.4.6).

Note that each SoFi_RT_criteria file should be dedicated to a separate SoFi RT run with a particular set of instruments. If for example several SoFi RT runs are performed from the same PC and the suite of instruments varied, then the user is supposed to change the criteria. This can be done by either deleting or moving the old file into another folder, or by modifying the criteria during an idle phase of SoFi RT. When re-initializing SoFi RT, the user will be asked whether the new and local list of criteria should be saved under windows or not.

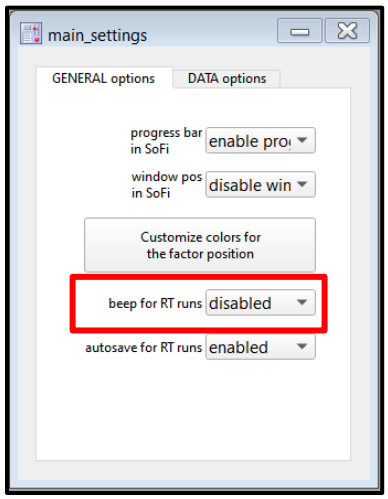


Figure 158 main SoFi menu with the beep option highlighted in the red rectangle.

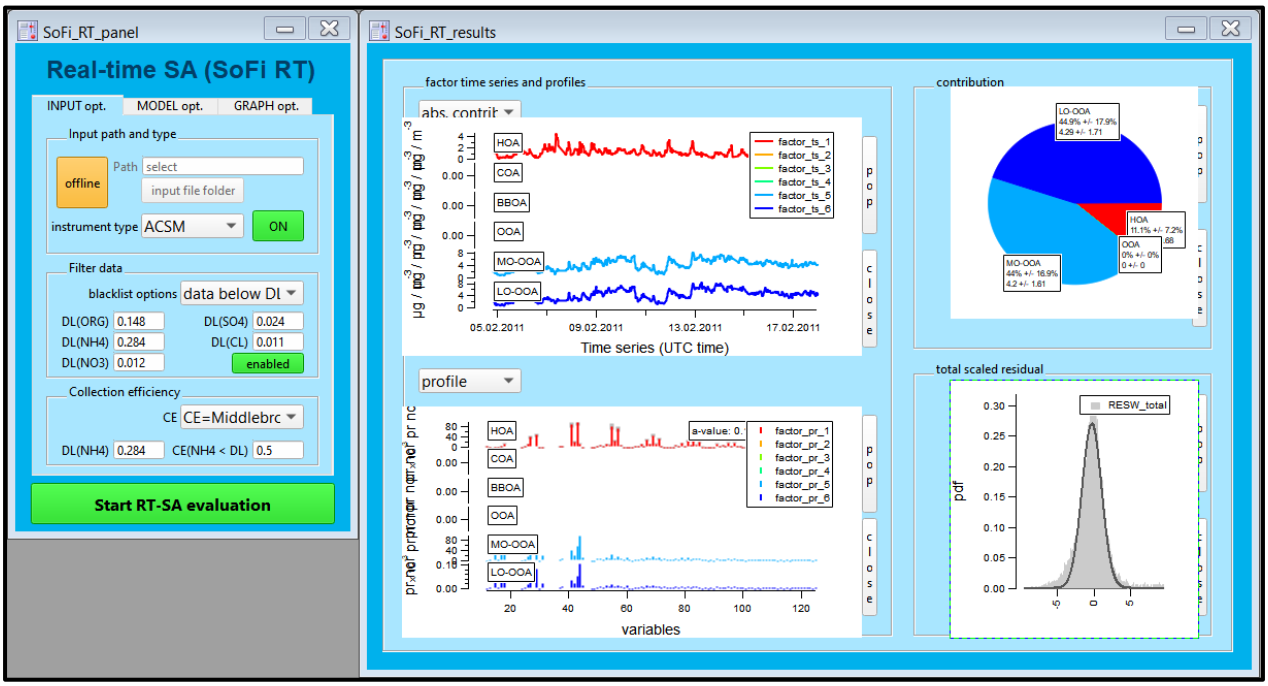


Figure 159 SoFi RT result panel with some Q-ACSM results

10.6. **Inspect SoFi RT results with SoFi Pro**

SoFi RT allows to have full control over the automatically selected PMF runs. The user can inspect all PMF and CMB runs performed through SoFi RT after having stopped SoFi RT. If SoFi RT shows some results on the SoFi RT result panel, this implies that the HDF file has been successfully loaded and the criteria panel processed in SoFi. Therefore, the user could only click on the button “III.a Select PMF run…” to see which runs have been selected and by clicking on “III.b Analyze chosen PMF run(s)” directly inspect the chosen PMF runs with no delay (Figure 160).

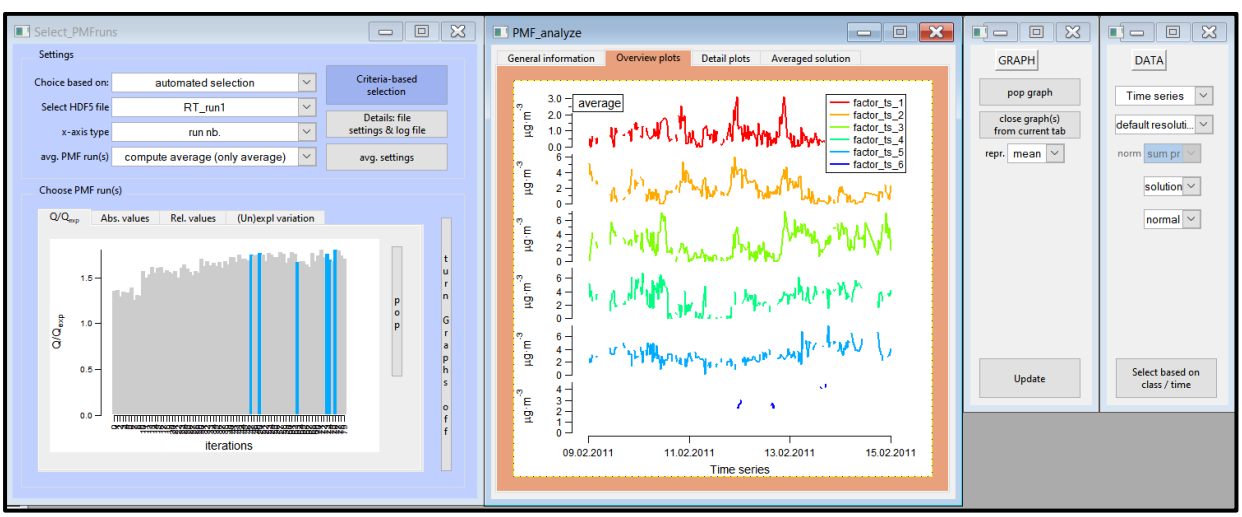


Figure 160 SoFi RT result inspected manually in SoFi.

10.7. Recommendations for ACSM data

Following the recommendations from (Crippa et al., 2014) for offline AMS/ACSM data analysis, when performing automated and in real-time rolling PMF analysis, the strategy also consists in constraining the factor profiles of the primary organic aerosol sources (POA) that are known to be present for the measurement location and to leave one up to two unconstrained factors for the assessment of the oxygenated organic aerosol (OOA). These factors, POA and OOA factors, do not necessarily need to be present all the time. When using the dynamic factor evaluation, SoFi RT automatically enables or disables the user-defined factors (Table 10). Note that currently, the range of dynamic factors is limited and there is a minimum to a maximum number of factors depending on how the user sets up the dynamic factor enabling or disabling (see 10.4.2.1 for more details and an example).

10.8. Generate daily BC files from AE33

SoFi RT requires daily .csv files containing absorption data measured by the Aethalometer, currently working for AE33. To generate these daily files, some settings have to be defined on the acquisition software of the AE33 together with the generation of a batch file that triggers the proper generation of these files. We here refer to the manual AE33_data_SoFi_instructions written by Dr. Manousakas Manousos Ioannis together with the batch file and an example file that can be found on our homepage.

SoFi RT requires daily .csv files containing absorption data measured by the Aethalometer, currently working for AE33. To generate these daily files, some settings have to be defined on the acquisition software of the AE33 together with the generation of batch files and a task scheduler that triggers the proper generation of these files. The reader is referred to the manual AE33_data_SoFi_instructions written by Dr. Manousos I. Manousakas at Paul Scherrer Institute (PSI) in 2022 together with the batch files and an example file that can be found on our homepage (https://datalystica.com/downloads/).

In case the experimentalist decides to use a different package that the one proposed here, please note the following when extracting the information from the AE33 acquisition software:

- The BC filename contains the information of the day in this exact format: “prefix or empty”+year+_+month+_+day.csv, e.g., 2024_03_21.csv.
- The columns in the file represent the following:
  - Instrument serial number (last 4 digits)
  - date (“dd-mm-yy”)
  - time (“hh:mm”)
  - BC1, BC2, BC3, BC4, BC5, BC6, BC7
  - FlowC (lpm)

Example:

0804,"01-Mar-24","00:00", 939, 900, 840, 823, 781, 758, 796, 5.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0

11. List of definitions and abbreviations

BBOA	biomass burning OA
COA	cooking OA
HOA	hydrocarbon-like OA
ME-2	Multilinear Engine second version
OA	organic aerosol
OOA	oxygenated OA
OS	Operating System
PMF	Positive Matrix Factorization
PMF2	solver for the bilinear positive matrix factorization algorithm
PMF3	solver for the trilinear positive matrix factorization algorithm
PMF call	Calling ME-2 to solve the PMF algorithm for certain settings, e.g. varying the number of factors, exploring the a-value, etc.
PMF run	Refers to a single PMF run out of a PMF call
PMF result / solution	Represents all PMF runs from one or more PMF calls that are environmentally reasonable and must be considered after a thorough rotational ambiguity exploration and the propagation of the statistical uncertainty.
POA	Primary organic aerosol
SA	Source apportionment
SOA	Secondary organic aerosol
SoFi	Source Finder software
SoFi Pro	key that enables the advanced utilities in SoFi
SoFi RT	key that enables the fully automated real-time SoFi software

12. Literature

Äijälä, M., Heikkinen, L., Fröhlich, R., Canonaco, F., Prévôt, A. S. H., Junninen, H., Petäjä, T., Kulmala, M., Worsnop, D., and Ehn, M.: Resolving anthropogenic aerosol pollution types – deconvolution and exploratory classification of pollution events, Atmos. Chem. Phys., 17, 3165-3197, http://10.5194/acp-17-3165-2017, 2017.

Aiken, A. C., DeCarlo, P. F., Kroll, J. H., Worsnop, D. R., Huffman, J. A., Docherty, K. S., Ulbrich, I. M., Mohr, C., Kimmel, J. R., Sueper, D., Sun, Y., Zhang, Q., Trimborn, A., Northway, M., Ziemann, P. J., Canagaratna, M. R., Onasch, T. B., Alfarra, M. R., Prevot, A. S. H., Dommen, J., Duplissy, J., Metzger, A., Baltensperger, U., and Jimenez, J. L.: O/C and OM/OC Ratios of Primary, Secondary, and Ambient Organic Aerosols with High-Resolution Time-of-Flight Aerosol Mass Spectrometry, Environ Sci Technol, 42, 4478-4485, http://doi.org/10.1021/es703009q, 2008.

Alton, M. W., Stark, H. J., Canagaratna, M. R., and Browne, E. C.: Generalized Kendrick analysis for improved visualization of atmospheric mass spectral data, Atmos. Meas. Tech., 16, 3273-3282, https://10.5194/amt-16-3273-2023, 2023.

Brown, S. G., Eberly, S., Paatero, P., and Norris, G. A.: Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results, Science of The Total Environment, 518-519, 626-635, https://doi.org/10.1016/j.scitotenv.2015.01.022, 2015.

Canagaratna, M. R., Jimenez, J. L., Kroll, J. H., Chen, Q., Kessler, S. H., Massoli, P., Hildebrandt Ruiz, L., Fortner, E., Williams, L. R., Wilson, K. R., Surratt, J. D., Donahue, N. M., Jayne, J. T., and Worsnop, D. R.: Elemental ratio measurements of organic compounds using aerosol mass spectrometry: characterization, improved calibration, and implications, Atmos Chem Phys, 15, 253-272, https://doi.org/10.5194/acp-15-253-2015, 2015.

Canonaco, F., Crippa, M., Slowik, J. G., Baltensperger, U., and Prevot, A. S. H.: SoFi, an IGOR-based interface for the efficient use of the generalized multilinear engine (ME-2) for the source apportionment: ME-2 application to aerosol mass spectrometer data, Atmos Meas Tech, 6, 3649-3661, https://doi.org/10.5194/amt-6-3649-2013, 2013.

Canonaco, F., Slowik, J. G., Baltensperger, U., and Prevot, A. S. H.: Seasonal differences in oxygenated organic aerosol composition: implications for emissions sources and factor analysis, Atmos Chem Phys, 15, 6993-7002, https://doi.org/10.5194/acp-15-6993-2015, 2015.

Canonaco, F., Tobler, A., Chen, G., Sosedova, Y., Slowik, J. G., Bozzetti, C., Daellenbach, K. R., El Haddad, I., Crippa, M., Huang, R.-J., Furger, M., Baltensperger, U., and Prévôt, A. S. H.: A new method for long-term source apportionment with time-dependent factor profiles and uncertainty

assessment using SoFi Pro: application to 1 year of organic aerosol data, Atmos Meas Tech, 14, 923-943, https://doi.org/10.5194/amt-14-923-2021, 2021.

Chen, G., Canonaco, F., Tobler, A., Aas, W., Alastuey, A., Allan, J., Atabakhsh, S., Aurela, M., Baltensperger, U., Bougiatioti, A., De Brito, J. F., Ceburnis, D., Chazeau, B., Chebaicheb, H., Daellenbach, K. R., Ehn, M., El Haddad, I., Eleftheriadis, K., Favez, O., Flentje, H., Font, A., Fossum, K., Freney, E., Gini, M., Green, D. C., Heikkinen, L., Herrmann, H., Kalogridis, A.-C., Keernik, H., Lhotka, R., Lin, C., Lunder, C., Maasikmets, M., Manousakas, M. I., Marchand, N., Marin, C., Marmureanu, L., Mihalopoulos, N., Močnik, G., Nęcki, J., O'Dowd, C., Ovadnevaite, J., Peter, T., Petit, J.-E., Pikridas, M., Matthew Platt, S., Pokorná, P., Poulain, L., Priestman, M., Riffault, V., Rinaldi, M., Różański, K., Schwarz, J., Sciare, J., Simon, L., Skiba, A., Slowik, J. G., Sosedova, Y., Stavroulas, I., Styszko, K., Teinemaa, E., Timonen, H., Tremper, A., Vasilescu, J., Via, M., Vodička, P., Wiedensohler, A., Zografou, O., Cruz Minguillón, M., and Prévôt, A. S. H.: European aerosol phenomenology − 8: Harmonised source apportionment of organic aerosol using 22 Year-long ACSM/AMS datasets, Environment International, 166, 107325, https://doi.org/10.1016/j.envint.2022.107325, 2022.

Crippa, M., Canonaco, F., Slowik, J. G., El Haddad, I., DeCarlo, P. F., Mohr, C., Heringa, M. F., Chirico, R., Marchand, N., Temime-Roussel, B., Abidi, E., Poulain, L., Wiedensohler, A., Baltensperger, U., and Prévôt, A. S. H.: Primary and secondary organic aerosol origin by combined gas-particle phase source apportionment, Atmos Chem Phys, 13, 1-47, https://doi.org/10.5194/acp-13-8411-2013, 2013.

Crippa, M., Canonaco, F., Lanz, V. A., Aijala, M., Allan, J. D., Carbone, S., Capes, G., Ceburnis, D., Dall'Osto, M., Day, D. A., DeCarlo, P. F., Ehn, M., Eriksson, A., Freney, E., Hildebrandt Ruiz, L., Hillamo, R., Jimenez, J. L., Junninen, H., Kiendler-Scharr, A., Kortelainen, A. M., Kulmala, M., Laaksonen, A., Mensah, A., Mohr, C., Nemitz, E., O'Dowd, C., Ovadnevaite, J., Pandis, S. N., Petaja, T., Poulain, L., Saarikoski, S., Sellegri, K., Swietlicki, E., Tiitta, P., Worsnop, D. R., Baltensperger, U., and Prevot, A. S. H.: Organic aerosol components derived from 25 AMS data sets across Europe using a consistent ME-2 based source apportionment approach, Atmos Chem Phys, 14, 6159-6176, https://doi.org/10.5194/acp-14-6159-2014, 2014.

Cubison, M. J., Ortega, A. M., Hayes, P. L., Farmer, D. K., Day, D., Lechner, M. J., Brune, W. H., Apel, E., Diskin, G. S., Fisher, J. A., Fuelberg, H. E., Hecobian, A., Knapp, D. J., Mikoviny, T., Riemer, D., Sachse, G. W., Sessions, W., Weber, R. J., Weinheimer, A. J., Wisthaler, A., and Jimenez, J. L.: Effects of aging on organic aerosol from open biomass burning smoke in aircraft and laboratory studies, Atmos. Chem. Phys., 11, 12049-12064, https://doi.org/10.5194/acp-11-12049-2011, 2011.

Davison, A. C., and Hinkley, D. V.: Bootstrap Methods and Their Application, Cambridge University Press, 1997.

Efron, B.: 1977 Rietz Lecture - Bootstrap Methods - Another Look at the Jackknife, Ann Stat, 7, 1-26, https://doi.org/10.1214/aos/1176344552, 1979.

Henry, R., Norris, G. A., Vedantham, R., and Turner, J. R.: Source Region Identification Using Kernel Smoothing, Environ Sci Technol, 43, 4090-4097, 10.1021/es8011723, 2009.

Kendrick, E.: A Mass Scale Based on CH2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds, Anal. Chem., 35, 2146-2154, https://10.1021/ac60206a048, 1963.

Miller, M. S., Hidy, G. M., and Friedlander, S. K.: Chemical element balance for Pasadena aerosol, J. Colloid Interface Sci., 39, 165-+, https://doi.org/10.1016/0021-9797(72)90152-x, 1972.

Ng, N. L., Canagaratna, M. R., Jimenez, J. L., Chhabra, P. S., Seinfeld, J. H., and Worsnop, D. R.: Changes in organic aerosol composition with aging inferred from aerosol mass spectra, Atmos Chem Phys, 11, 6465-6474, https://doi.org/10.5194/acp-11-6465-2011, 2011a.

Ng, N. L., Herndon, S. C., Trimborn, A., Canagaratna, M. R., Croteau, P. L., Onasch, T. B., Sueper, D., Worsnop, D. R., Zhang, Q., Sun, Y. L., and Jayne, J. T.: An Aerosol Chemical Speciation Monitor (ACSM) for routine monitoring of the composition and mass concentrations of ambient aerosol, Aerosol Science and Technology, 45, 780-794, https://doi.org/10.1080/02786826.2011.560211, 2011b.

Norris, G., Duvall, R. M., Brown, S. S., and S., B.: EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide, U.S. Environmental Protection Agency, 2014.

Ogulei, D., Hopke, P. K., Zhou, L., Paatero, P., Park, S. S., and Ondov, J. M.: Receptor modeling for multiple time resolved species: The Baltimore supersite, Atmos Environ, 39, 3751-3762, https://doi.org/10.1016/j.atmosenv.2005.03.012, 2005.

Paatero, P., and Tapper, U.: Positive Matrix Factorization - a nonnegative factor model with optimal utilization of error-estimates of data values, Environmetrics, 5, 111-126, http://10.1002/env.3170050203, 1994.

Paatero, P.: Least squares formulation of robust non-negative factor analysis, Chemometr Intell Lab, 37, 23-35, http://10.1016/S0169-7439(96)00044-5, 1997.

Paatero, P.: The multilinear engine - A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model, J Comput Graph Stat, 8, 854-888, http://10.2307/1390831, 1999.

Paatero, P., Hopke, P. K., Song, X. H., and Ramadan, Z.: Understanding and controlling rotations in factor analytic models, Chemometr Intell Lab, 60, 253-264, Doi 10.1016/S0169-7439(01)00200-3, 2002.

Paatero, P., and Hopke, P. K.: Discarding or downweighting high-noise variables in factor analytic models, Anal. Chim. Acta, 490, 277-289, https://doi.org/10.1016/S0003-2670(02)01643-4, 2003.

Paatero, P.: User's guide for positive matrix factorization programs PMF2 and PMF3, University of Helsinki, Helsinki, Finland, 2004.

Paatero, P., and Hopke, P. K.: Rotational tools for factor analytic models, J. Chemom., 23, 91-100, https://doi.org/10.1002/cem.1197, 2009.

Paatero, P.: User's guide for positive matrix factorization programs PMF2 and PMF3. In: University of Helsinki, Helsinki, Finland, 2010.

Paatero, P., Eberly, S., Brown, S. G., and Norris, G. A.: Methods for estimating uncertainty in factor analytic solutions, Atmos. Meas. Tech., 7, 781-797, https://doi.org/10.5194/amt-7-781-2014, 2014.

Parworth, C., Fast, J., Mei, F., Shippert, T., Sivaraman, C., Tilp, A., Watson, T., and Zhang, Q.: Long-term measurements of submicrometer aerosol chemistry at the Southern Great Plains (SGP) using an Aerosol Chemical Speciation Monitor (ACSM), Atmos Environ, 106, 43-55, 10.1016/j.atmosenv.2015.01.060, 2015.

Petit, J. E., Favez, O., Albinet, A., and Canonaco, F.: A user-friendly tool for comprehensive evaluation of the geographical origins of atmospheric pollution: Wind and trajectory analyses, Environmental Modelling & Software, 88, 183-187, https://doi.org/10.1016/j.envsoft.2016.11.022, 2017.

Polissar, A. V., Hopke, P. K., Paatero, P., Malm, W. C., and Sisler, J. F.: Atmospheric aerosol over Alaska: 2. Elemental composition and sources, Journal of Geophysical Research: Atmospheres, 103, 19045-19057, https://doi.org/10.1029/98JD01212, 1998.

Sandradewi, J., Prévôt, A. S. H., Weingartner, E., Schmidhauser, R., Gysel, M., and Baltensperger, U.: A study of wood burning and traffic aerosols in an Alpine valley using a multi-wavelength Aethalometer, Atmos Environ, 42, 101-112, http://dx.doi.org/10.1016/j.atmosenv.2007.09.034, 2008.

Silverman, B. W.: Density Estimation for Statistics and Data Analysis, London: Chapman & Hall/CRC, 1986.

Slowik, J. G., Vlasenko, A., McGuire, M., Evans, G. J., and Abbatt, J. P. D.: Simultaneous factor analysis of organic particle and gas mass spectra: AMS and PTR-MS measurements at an urban site, Atmos Chem Phys, 10, 1969-1988, https://doi.org/10.5194/acp-10-1969-2010, 2010.

Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., and Jimenez, J. L.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmos Chem Phys, 9, 2891-2918, http://10.5194/acp-9-2891-2009, 2009.

Van Krevelen, D. W.: Graphical-statistical method for the study of structure and reaction processes of coal, Fuel, 24, 269-284, 1950.

Visser, S., Slowik, J. G., Furger, M., Zotter, P., Bukowiecki, N., Canonaco, F., Flechsig, U., Appel, K., Green, D. C., Tremper, A. H., Young, D. E., Williams, P. I., Allan, J. D., Coe, H., Williams, L. R., Mohr, C., Xu, L., Ng, N. L., Nemitz, E., Barlow, J. F., Halios, C. H., Fleming, Z. L., Baltensperger, U., and Prévôt, A. S. H.: Advanced source apportionment of size-resolved trace elements at multiple sites in London during winter, Atmos Chem Phys, 15, 11291-11309, https://doi.org/10.5194/acp-15-11291-2015, 2015.

Wold, S., Esbensen, K., and Geladi, P.: Principal component analysis, Chemometr Intell Lab, 2, 37-52, https://doi.org/10.1016/0169-7439(87)80084-9, 1987.

Zotter, P., Herich, H., Gysel, M., El-Haddad, I., Zhang, Y., Močnik, G., Hüglin, C., Baltensperger, U., Szidat, S., and Prévôt, A. S. H.: Evaluation of the absorption Ångström exponents for traffic and wood burning in the Aethalometer-based source apportionment using radiocarbon measurements of ambient aerosol, Atmos. Chem. Phys., 17, 4229-4249, https://doi.org/10.5194/acp-17-4229-2017, 2017.