https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

Atmospheric Measurement Techniques
Discussions

A novel probabilistic source apportionment approach: Bayesian Auto-correlated Matrix Factorization

Anton Rusanen[1,2], Anton Björklund[3], Manousos Manousakas[4], Jianhui Jiang[4,5], Markku T. Kulmala[1,6,7], Kai Puolamäki[1,3], and Kaspar R. Daellenbach[4]

[1]Institute for Atmospheric and Earth System Research (INAR) / Physics, Faculty of Science, University of Helsinki, Finland
[2]Atmospheric Composition Research, Finnish Meteorological Institute, Helsinki, Finland
[3]Department of Computer Science, Faculty of Science, University of Helsinki, Finland
[4]Laboratory of Atmospheric Chemistry, Paul Scherrer Institute (PSI), 5232 Villigen-PSI, Switzerland
[5]Shanghai Key Lab for Urban Ecological Processes and Eco-Restoration, School of Ecological and Environmental Sciences, East China Normal University, 200241, Shanghai, China
[6]Aerosol and Haze Laboratory, Beijing Advanced Innovation Center for Soft Matter Sciences and Engineering, Beijing University of Chemical Technology (BUCT), Beijing, China
[7]Joint International Research Laboratory of Atmospheric and Earth System Sciences, School of Atmospheric Sciences, Nanjing University, Nanjing, China

Correspondence: Anton Rusanen (anton.rusanen@helsinki.fi), Kaspar R. Daellenbach (kaspar.daellenbach@psi.ch)

Abstract.

The concentrations and sources of particulate matter in the atmosphere are temporally auto-correlated. Here, we present a Bayesian matrix factorization model (BAMF) that considers the temporal auto-correlation of the components (sources) and provides a direct error estimation. The performance of BAMF is compared to positive matrix factorization (PMF) using synthetic Time-of-Flight Aerosol Chemical Speciation Monitor data, representing different urban environments from typical European towns to megacities. We find that BAMF resolves sources better than PMF on all datasets with auto-correlated components, but highly cross-correlated components continue to be challenging. However, we demonstrate that adding even partial prior information about the chemical composition of the components to BAMF improves the factorization. Overall, BAMF-type models are promising tools for source apportionment and merit further research.

1 Introduction

Air pollution in the form of particulate matter (PM) has a substantial impact on earth's climate (IPCC, 2021) and severe adverse effects on human health (Lelieveld et al., 2015; Daellenbach et al., 2020). PM's noxiousness could strongly depend on the particles' chemical composition, which is governed by their origin (Bates et al., 2019; Daellenbach et al., 2020). PM is affected by many emission sources and dynamic atmospheric processes making PM a poorly understood complex mixture, especially the organic aerosol (OA) fraction of PM. Typically directly emitted OA (primary OA - POA) is distinguished from OA formed in the atmosphere from emitted vapours (secondary OA - SOA). Identifying and quantifying the sources of PM is, therefore, essential for designing effective and efficient air pollution reduction strategies. Such analyses (called source apportionment) combine chemical characterization data with matrix factorization methods. The idea of these methods is to

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

use the variation in the chemical composition of a set of measurements, such as outputs from mass spectrometers, and to decompose the measurements into "source terms" by using a matrix factorization formalism. The underlying assumption is that the measurement is a linear combination of strictly non-negative source terms.

Multiple methods for weighted non-negative matrix factorization exist (Wang and Zhang, 2012). The most used one in atmospheric sciences is positive matrix factorization (PMF) (Paatero and Tapper, 1994). In earlier studies, chemical mass balance, CMB, was a popular method, but it has the drawback that factor profiles must be defined beforehand, see e.g. Watson et al. (2001). This introduced significant uncertainty since these factor profiles are usually not known beforehand or only with considerable uncertainty. PMF improved on this by optimizing the source profiles (Canonaco et al., 2013). These models produce point solutions with arbitrary rotations (Paatero and Tapper, 1994; Ulbrich et al., 2009), which makes interpreting the results difficult.

Previous studies have revealed that chemical data from the Aerosol Mass Spectrometer family (Aerodyne Aerosol Mass Spectrometer (Canagaratna et al., 2007), Aerosol Chemical Speciation Monitor (Ng et al., 2011; Fröhlich et al., 2013)) retains sufficient information for the resolution of some sources (Zhang et al., 2007, 2011; Daellenbach et al., 2017). However, distinguishing factors with chemical or temporal similarities or accurately resolving low-concentration factors is often challenging (Ulbrich et al., 2009; Canonaco et al., 2013; Zhang et al., 2011). Several studies have shown that utilizing a priori information to constrain POA sources' chemical composition is usually required to accurately estimate their contribution to OA (Canonaco et al., 2013; Crippa et al., 2014; Reyes-Villegas et al., 2016; Schlag et al., 2017; Zhang et al., 2018; Huang et al., 2019; Zhu et al., 2018). In addition, different statistical data reduction methods applied to mass spectrometry data extract different components (Isokääntä et al., 2020). This demonstrates that the problem does not have one unique solution, and the choice of method can emphasize different features of the resolved components.

While developments related to source apportionment focused on different ways to pre- and post-process data (Canonaco et al., 2021; Zhang et al., 2019), in atmospheric science, the underlying solver algorithm mainly remained the same, PMF. Particularly relevant for this study is that the commonly used optimization goals do not include any temporal terms of the resolved components (Wang and Zhang, 2012; Paatero and Tapper, 1994), and thus any time information is ignored. Lack of time information is a drawback because many atmospheric measurements exhibit strong temporal auto-correlation (Figure 1). Here, we present a probabilistic matrix factorization method that accounts for auto-correlation. We evaluate the model's performance in resolving air pollution sources based on realistic synthetic chemical data.

2 Methods

2.1 Notation

In this paper, we describe the data by $\mathbf{X} \in \mathbb{R}^{n \times m}$, where the rows $i \in [n] = \{1, \ldots, n\}$ correspond to measurements taken at consecutive times t_i. The columns $j \in [m]$ correspond to the different dimensions of the measurement. Our objective is to find a lower dimensional non-negative decomposition $\mathbf{X} \approx \mathbf{G}\mathbf{F}$ with p factors such that $\mathbf{G} \in \mathbb{R}_{\geq 0}^{n \times p}$ and $\mathbf{F} \in \mathbb{R}_{\geq 0}^{p \times m}$, where $p \ll \min(n, m)$. In other words, the objective is to present the data as a multiplication of two much smaller matrices. The

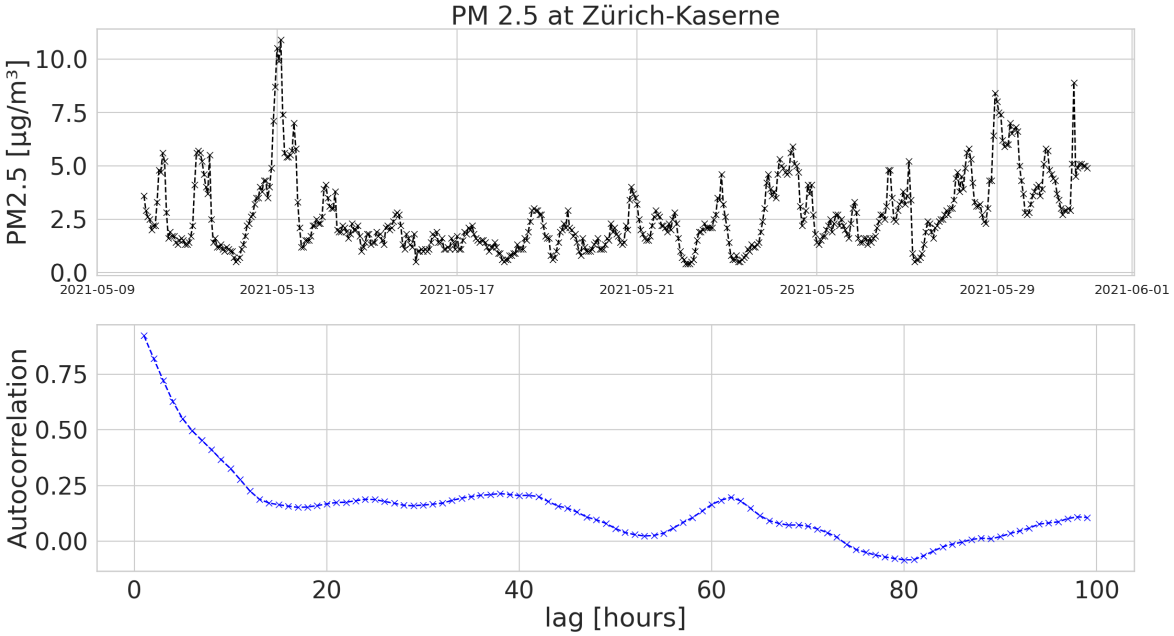

Figure 1. Temporal variability of PM2.5 in Zürich, data from NABEL - the Swiss National Air Pollution Monitoring Network. This illustration display hourly PM2.5 concentrations as well as their auto-correlation (Pearson correlation coefficient as a function of the time lag).

rows of $\mathbf{F}_{i\cdot}$ contain the time-independent components of the decomposition, which we call *factor profiles*. Factor profiles are defined to sum to unity to facilitate comparisons between different datasets and models. The columns of $\mathbf{G}_{\cdot i}$ contain the time dependency of each of the rows of $\mathbf{F}$; we will call these the *factor time series*. Simply put, the factor profiles represent the sources' concentration-independent chemical composition, and the factor time series describes the sources' time-dependent concentration. Note that the ordering of these profiles is arbitrary for the overall solution.

2.2 Bayesian Auto-correlated Matrix Factorization, BAMF

We define a Bayesian probabilistic model that captures our prior assumptions of the process that generated the measurements. The only observed variables in our model are the data matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ and the error estimate $\boldsymbol{\sigma} \in \mathbb{R}^{n\times m}_{\geq 0}$. The error matrix contains the measurement uncertainty determined as standard deviations of the error terms for each data point. In addition to these observed variables, there are several latent variables. These include the matrices $\mathbf{G}$ and $\mathbf{F}$ mentioned above, vectors $\alpha_a \in \mathbb{R}^p_{\geq 0}$ and $\alpha_b \in \mathbb{R}^p_{\geq 0}$ which determine the auto-correlation behaviour of the model, as well the "noise-free data matrix"

$\mathbf{Z} \in \mathbb{R}_{\geq 0}^{n \times m}$. We define the probabilistic model as

$$\begin{aligned}
\boldsymbol{Z} &= \boldsymbol{GF} && (1)\\
\boldsymbol{X}_{ij} &\sim \text{Normal}(\boldsymbol{Z}_{ij},\, \boldsymbol{\sigma}_{ij}) \text{ for all } i \in [n] \text{ and } j \in [m] && (2)\\
\boldsymbol{F}_{i\cdot} &\sim \text{Dirilecht}(\mathbf{1}_m) \text{ for all } i \in [n] && (3)\\
\boldsymbol{G}_{i+1,k} &\sim \text{Cauchy}(\boldsymbol{G}_{ik},\, \alpha_a[k]\Delta t_i + \alpha_b[k]) \text{ for all } k \in [p] \text{ and } i \in [n-1] && (4)
\end{aligned}$$

where Normal corresponds to the normal probability distribution with a given mean and standard deviation, and Dirichlet to the Dirichlet distribution parameterized by a unit vector $\mathbf{1}_m$. The model specification implies that the components of $\mathbf{F}_{i\cdot}$ can have values from $[0,1]$ with equal likelihood, but all rows must sum to unity. Cauchy is the Cauchy probability distribution, with the width depending on the time difference between the ith and $(i+1)$th observation ($\Delta t_i = t_{i+1} - t_i$). Essentially our model describes the data as a non-negative matrix decomposition (NMF) with a lag-1 auto-correlation term and a Gaussian error term for the reconstruction of $\mathbf{X}$.

We chose the Cauchy distribution for the auto-correlation term because the long tails make large jumps between the ith and $(i+1)$th observation more probable than, for example, a Gaussian distribution. Other choices are possible, but the experiments in this paper suggest the Cauchy distribution works as an approximation for real data. Choosing the distribution shape also implicitly influences the weight of $\mathbf{G}$'s auto-correlation. The term $\alpha_a[k]\Delta t_i + \alpha_b[k]$ determines the width of the Cauchy distribution. The α-terms allow the model to deal with time steps of different lengths and missing data. This term has a minimum width $\alpha_b[k]$ and increases linearly as Δt increases (with $\alpha_a[k]$). Thus for arbitrarily large time steps, it approaches a uniform distribution. It is possible to use other formulations for the width, which would be appropriate if one wishes to include a more complex and computationally intensive description of auto-correlation.

2.2.1 Uncorrelated Bayesian matrix factorization, BAMF-0

For comparison, we created a version of the BAMF model without the lag-1 auto-correlation terms of Equation (4). The model is otherwise identical to the BAMF model. This variation is essentially a probabilistic weighted NMF model, making it possible to assess the impact of the auto-correlation terms on the solution.

2.2.2 Bayesian matrix factorization with additional constraints, BAMF-C

In source apportionment analyses, it is common to utilize reference spectra as boundary conditions for the factor analysis – to find components with, e.g., previously observed chemical compositions. We include this scenario in another model, called BAMF-C, by adding peak intensity ratios to the BAMF model,

$$F[i,j]/F[k,l] \sim \text{Normal}(\text{ratio}, \text{width}) \qquad (5)$$

where i,j and k,l are the indices of the peak pair to constrain. The ratio is the desired intensity ratio, and width is a free parameter describing the width of the distribution, i.e. the uncertainty of the intensity ratio. This approach allows constraining the range of **F** for arbitrarily many m/z pairs, which is currently not done in the other models. A similar concept could constrain **G** to have a similar time behaviour as an external ancillary measurement, e.g., of a source tracer. The constraint is similar to the widely used *a-value* (anchor value) approach in Source Finder coupled to PMF (Canonaco et al., 2013, SoFi/PMF) with two notable differences. Firstly, the intensity ratio of the peaks is constrained in BAMF-C, while the a-value constraint approach in SoFi/PMF operates on the peak intensity. Secondly, BAMF-C punishes any deviation from the target value with a soft boundary (with an increasing penalty). At the same time, SoFi/PMF employs a hard boundary (defined as a relative deviation from the a-value), considering all solutions within the boundary to be of equal quality.

2.3 Solver

We use STAN (Carpenter et al., 2017) to compile and run the probabilistic models. STAN solves the probabilistic inference problem with a Markov Chain Monte Carlo (MCMC) method. Instead of obtaining a single solution for the latent variables, for example, by finding the model with the highest likelihood, we get a distribution of possible solutions from which we can infer, e.g., confidence intervals. STAN takes our model and observed variables as input and outputs samples from the posterior distribution of the latent variables. We run MCMC in multiple parallel chains, starting from different initial conditions and usually extract a few thousand posterior samples per chain.

The standard way to initialize the model in STAN is by randomly sampling from the prior distributions. However, our model has many parameters with fairly strict distributions. Consequently, we found this starting point to be poor, sometimes causing STAN to slow down markedly or even fail. Thus, we initialize the model with a point solution. Here we use STAN's capability to find a single maximum-a-posteriori (MAP) point solution for the parameters, which we use as the initialization. Note, however, that the solutions typically have several local optima, in which case the point solution is only one such local optimum.

2.4 Pre- & post-processing

Before running our model, we normalize the data such that the mean of the data ($\boldsymbol{X}$) is 1. The error estimate is scaled with the same scaling factor. The equations to do this are

$$f_{norm} = \sum_{i,j} \boldsymbol{X}_{ij}/(n \times m)$$

$$\boldsymbol{X}^*_{ij} = \boldsymbol{X}_{ij}/f_{norm}$$

$$\boldsymbol{\sigma}^*_{ij} = \boldsymbol{\sigma}_{ij}/f_{norm}$$

where f_{norm} is a scalar normalization factor, $\boldsymbol{X}^*_{ij}$ and $\boldsymbol{\sigma}^*_{ij}$ are the scaled data and error, respectively, which we use as model inputs. In practice, this only affects the scale of the priors and posteriors, and the user can also choose to use non-scaled values.

STAN outputs posterior samples from the two matrices $\boldsymbol{F}$ and $\boldsymbol{G}$, representing the factors' time-independent chemical composition and their time-dependent concentration. Since all the magnitude information is in **G**, **G** needs to be renormalized by simply multiplying with the normalization factor. The rows of $\boldsymbol{F}$ are constrained to sum to unity, and are, thus, directly comparable to mass spectrometric references, which are normalized similarly (Crippa et al., 2013; Ulbrich et al., 2022).

2.4.1 Sorting the components

The order of the components in $\boldsymbol{F}$ and $\boldsymbol{G}$ is arbitrary in our samples. The problem is not unique to BAMF but inherent to all such matrix decompositions. To be able to compare solutions, we need to be able to sort the components. The contribution of the same component to $\boldsymbol{Z}$ should be similar between two samples. To calculate the contribution for each component, we multiply the row of $\boldsymbol{F}$ with the corresponding column of $\boldsymbol{G}$ and use this to sort the components.

To select the ordering of the components, we take a small number of representative samples, usually the last five, and compute the optimal permutation using the Hungarian algorithm (Kuhn, 1955), minimizing the Manhattan distances between the $\boldsymbol{Z}$ contributions in the samples. We then select the most common permutation as the ordering of the factors for all samples.

We use this approach for sorting the outputs of all models (BAMF-0/C, BAMF, PMF) to ensure the most direct comparability of the results. Finally, median, 25% and 75% percentiles are computed using the sorted samples. In the comparisons, we use medians for all models, but in some figures, we also show 25% and 75% percentiles. The median, or any central estimate, is not guaranteed to be the "best" optimized solution in any metric (log probability or root mean squared sum of residuals). Still, we use it to represent a reasonable solution inferred from the samples.

2.5 Evaluating model performance

The first metric to check is if the model explains the data well (reconstruction performance). If $\boldsymbol{X}$ is not reconstructed appropriately, the solution is not acceptable. This can either mean that the data cannot be factorized this way (the model assumptions are wrong) or that the solver failed to find a solution. In such cases, the number of iterations should be increased, or solver parameters must be adjusted, such as the number of warm-up samples and parameters influencing the step size.

Even at moderate data sizes, assessing if the original data falls within the model's confidence bounds for every variable individually is not practical. Therefore, we summarised this information by computing the model residual (difference between the model input and output, mean of all samples) normalized with the uncertainty of the model output (standard deviation of all samples); essentially observing if the original data is inside the sample standard deviation.

$$\boldsymbol{S}_{ij} = (\boldsymbol{X}_{ij} - E[\boldsymbol{X}_{samples}])/\sigma(\boldsymbol{X}_{samples}) \tag{6}$$

Data in $\boldsymbol{S}$ should be centred at zero and have a standard deviation below 1, which means the data is often less than 1 model standard deviation away from the model mean. In addition, we also use a common evaluation metric in PMF analyses (Q_m/Q_{exp}), where Q_m is defined as (Canonaco et al., 2013, notation adapted):

$$Q_m = \sum_{i,j} \left(\frac{(\boldsymbol{X}-\boldsymbol{Z})_{i,j}}{\boldsymbol{\sigma}_{i,j}} \right)^2$$

Essentially Q_m describes the sum of squared model residuals normalized to the input error. We use the same reduction as Zhang et al. (2011) where Q_{exp} is approximated as data size and denote Q_m/Q_{exp} as Q_m*.

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

For synthetic data—with a known ground truth—it is possible to assess how well the methods resolve the actual components in addition to the *reconstruction performance*. We call this evaluation *factorization performance*. We compare the median solutions with the corresponding actual components by calculating the average distance, Pearson and Spearman (nonlinear) correlations. Optimal matching between median solutions and true components is obtained using the Hungarian algorithm (Kuhn, 1955). The approach is similar to the sorting above but with the true components defining the order. Direct comparison to true components is only possible in cases where the number of true components matches the number of modelled components. Otherwise, the model must combine multiple components or create additional ones.

2.6 PMF

We use PMF, specifically the multilinear engine 2 (ME-2) controlled by the user interface SoFi (Canonaco et al., 2013; Paatero and Tapper, 1994), as a baseline comparison. PMF solves the decomposition in Equation 1 by minimizing the sum of the squared residuals normalized with the input error (object function), given the boundary condition that all values must be positive (Canonaco et al., 2013). Since SoFi finds local optima, we ran it with different random seeds to get multiple solutions for all comparisons. As the runs have varying starting points, they often lead to different local optima, especially in cases with high rotational ambiguity. Thus, PMF provides a collection of local minima, while BAMF tries to sample the model's posterior distribution, including sub-optimal but still plausible answers. For simplicity, we will refer to the group of PMF solution sets as samples, even though PMF is not a sampler.

A priori information in the form of known rows of factor profiles or of known columns of factor time series can be added to the model to reduce the rotational ambiguity. By adding this external information, the user can reduce the space PMF searches for the optimized solution, reducing the rotational ambiguity of the solution. Using external data to run PMF is usually referred to as constraining the solution, and external information is used as constraints. Here we used two approaches, a) entirely unconstrained PMF runs and b) constrained runs using external source profiles. We rely on the commonly used a-value approach to constrain the PMF runs. In the a-value approach, the user inputs one or more factor profiles or factor time series and defines a relative tolerated deviation from the anchor (termed a-value) (Canonaco et al., 2013). Constraint strengths are not directly comparable between the hard cut-off approach used in PMF and the softer Gaussian error term approach used in BAMF-C. For the best possible comparability, we first ran BAMF-C (constraint strength 0.001) and determined the equivalent a-value by taking the maximum deviation from the anchor value on a constrained component in $\boldsymbol{F}$ (a-value of 18%).

3 Datasets

We generated synthetic datasets mimicking the OA sources in different urban environments. These synthetic datasets mimic mass spectral OA analyses of a Time-of-Flight Aerosol Chemical Speciation Monitor (ToF-ACSM, Fröhlich et al. (2013)), a ubiquitous instrument in measuring PM composition with a focus on OA. The instrument measures a signal for several mass-to-charge ratio channels. We model these mass spectra as a sum of 2–5 different mixed sources. The sources are constructed of time-independent chemical fingerprints, in our notation $\mathbf{F}$, from the AMS Spectral Database (Ulbrich et al., 2009, 2022) combined with their time behaviour and magnitude, **G**. The noiseless spectra, **Z**, is then acquired by matrix multiplication of **F** and **G**. We then generate **X**, as Equation 2, by applying random Gaussian noise to each data point. The errors are applied to $\boldsymbol{X}$, which is a sum of all the components, so individual component error in the original $\boldsymbol{G}$ is undefined.

The ToF-ACSM alternates between measuring particles and air together, called open signal (I_{open}), and measuring only air, called closed signal (I_{closed}). The difference signal (I_{diff}=I_{open}-I_{closed}) represents the signal caused by the measured particles. For computing the error of I_{diff}, we use an error function based on the signal strength, according to Allan et al. (2003), and Ulbrich et al. (2009):

$$Error_{open} = \sqrt{\frac{(I_{open} + I_{baseline}) \times t_{open}}{\sqrt{\frac{28}{m/z}}}} \tag{7}$$

$$Error_{closed} = \sqrt{\frac{(I_{closed} + I_{baseline}) \times t_{closed}}{\sqrt{\frac{28}{m/z}}}} \tag{8}$$

$$Error = max(Error_{min}, \frac{1.2 \times \sqrt{Error_{open}^2 + Error_{closed}^2}}{t_{open} \times \sqrt{\frac{28}{m/z}}}) \tag{9}$$

Where t_{open} and t_{closed} are the open and closed signal measurement times, respectively, m/z is the mass charge ratio of the measurement, $Error_{min}$ is a lower limit set on the measurement error, and $I_{baseline}$ is the baseline signal in the mass spectrometer. Since the organic fragments ions at the m/z values 16,17,18 and 28 are computed based on the measurement at m/z 44 and thus contain duplicate information, they are removed before running any of the models and only later reintroduced in the results.

3.1 Synthetic data representing a polluted megacity

First, we generated a synthetic ToF-ACSM OA mass spectral dataset mimicking a polluted megacity environment affected by multiple OA sources. The synthetic datasets used here are based on observations from Beijing, as it is a relatively well-studied environment. In our case, the modelled sources are traffic exhaust: HOA, cooking: COA, biomass burning: BBOA, coal combustion: CCOA, and secondary OA: OOA. In addition, we also constructed more simple datasets generated with fewer factors (2 factors: HOA + OOA, three factors: HOA+COA+OOA, and four factors: HOA+COA+BBOA+OOA). $\boldsymbol{F}$ were chemical fingerprints from literature (Elser et al., 2016; Ulbrich et al., 2009, 2022). $\boldsymbol{G}$ was created as a mix of Gaussian and Cauchy (BBOA Gaussian, others Cauchy), biased, positive random walks, with added typical diurnal concentration cycles (Kulmala et al., 2021) corresponding to the matching OA sources ($\boldsymbol{F}$). The components' (OA sources') overall order of magnitudes and diurnal concentration variability were estimated based on previous literature on OA sources in Beijing (Kulmala et al., 2021). For each number of factors, we constructed ten different datasets amounting to a total of 40 datasets. For reference, one 5-component dataset in its component form is shown in Figure 2. CCOA and BBOA are very similar in $\boldsymbol{F}$ (Pearson correlation coefficient: 0.94, other factors 0.43 to 0.92), making it a challenging dataset. The time series of CCOA and BBOA are also similar in magnitude and have similar diurnal behaviour. While the short-term auto-correlation is high, the random walks of the megacity dataset are not as highly auto-correlated as the PM data in Figure 1. The added diurnal peaks can be seen as the periodicity of the tail. See Appendix A for an example of the m/z dependence of the measurement errors on this dataset.

3.2 Synthetic dataset representing a typical European urban environment

As another test, we used chemical transport model data from Jiang et al. (2019) representing approximately two weeks of simulated measurements in Zurich, Switzerland. Zurich represents a typical European city with low pollution levels. In this case, the $\boldsymbol{G}$ time-series come from the transport model and $\boldsymbol{F}$ is taken from the literature (HOA and BBOA from an ambient analysis presented by (Elser et al., 2016; Ulbrich et al., 2009, 2022), biological SOA ($\mathrm{SOA_{bio}}$) from an ambient analysis presented by Daellenbach et al. (2017), anthropogenic SOA, $\mathrm{SOA_{anthro}}$, is represented by laboratory Diesel generator SOA presented by Sage et al. (2008)).

This dataset differs from those in Section 3.1 in two important ways. Firstly the concentrations are lower since the environment is less polluted, which affects the error estimation as larger relative measurement errors, median 1.0 % of data magnitude for this dataset and median 0.6% for one of the datasets described in Section 3.1. Secondly, the sources exhibit high cross-correlation in the time series (Pearson correlation for components of $\boldsymbol{G}$ is 0.77 between HOA and BBOA and 0.71 between HOA and $\mathrm{SOA_{anthro}}$, for comparison in the synthetic megacity ToF-ACSM OA datasets (Section 3.1) a high correlation in $\boldsymbol{G}$ would be 0.5), possibly indicating that meteorological conditions and transport of pollutants are important drivers of the concentration of the components. From a source apportionment analysis perspective, this simulates the worst-case situation with the data having poor separability in **G**. The components of this dataset compared to the megacity data can be seen in Figure 2. The auto-correlation behaviour of the two datasets is very similar, indicating that our fully synthetic data behave as realistically as the transport model.

4 Experiments

4.1 Simulated megacity source apportionment

In the first experiment, we assess the performance of BAMF, BAMF-0, and PMF on synthetic data mimicking the conditions in a polluted megacity described in Section 3.1. First, we assess the reconstruction of the input by the different models. In addition to minimizing the residuals, BAMF also includes a penalty for deviations from auto-correlation in **G**. Due to this and BAMF being a sampled model instead of an optimizer as described in Section 2.3, PMF would be expected to give answers with lower absolute measurement error weighted residuals compared to BAMF. In other words, PMF is expected to have a better reconstruction performance.

Figure 3 shows the median solution reconstruction across the ten different datasets as the number of components increases. All models reconstruct the input data well within the error estimate. The similar reconstruction metrics for BAMF and BAMF-0 suggest that the inclusion of the auto-correlation term does not substantially deteriorate the reconstruction accuracy. On the other hand, PMF has, as expected, marginally lower absolute measurement error-weighted residuals than BAMF and BAMF-

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

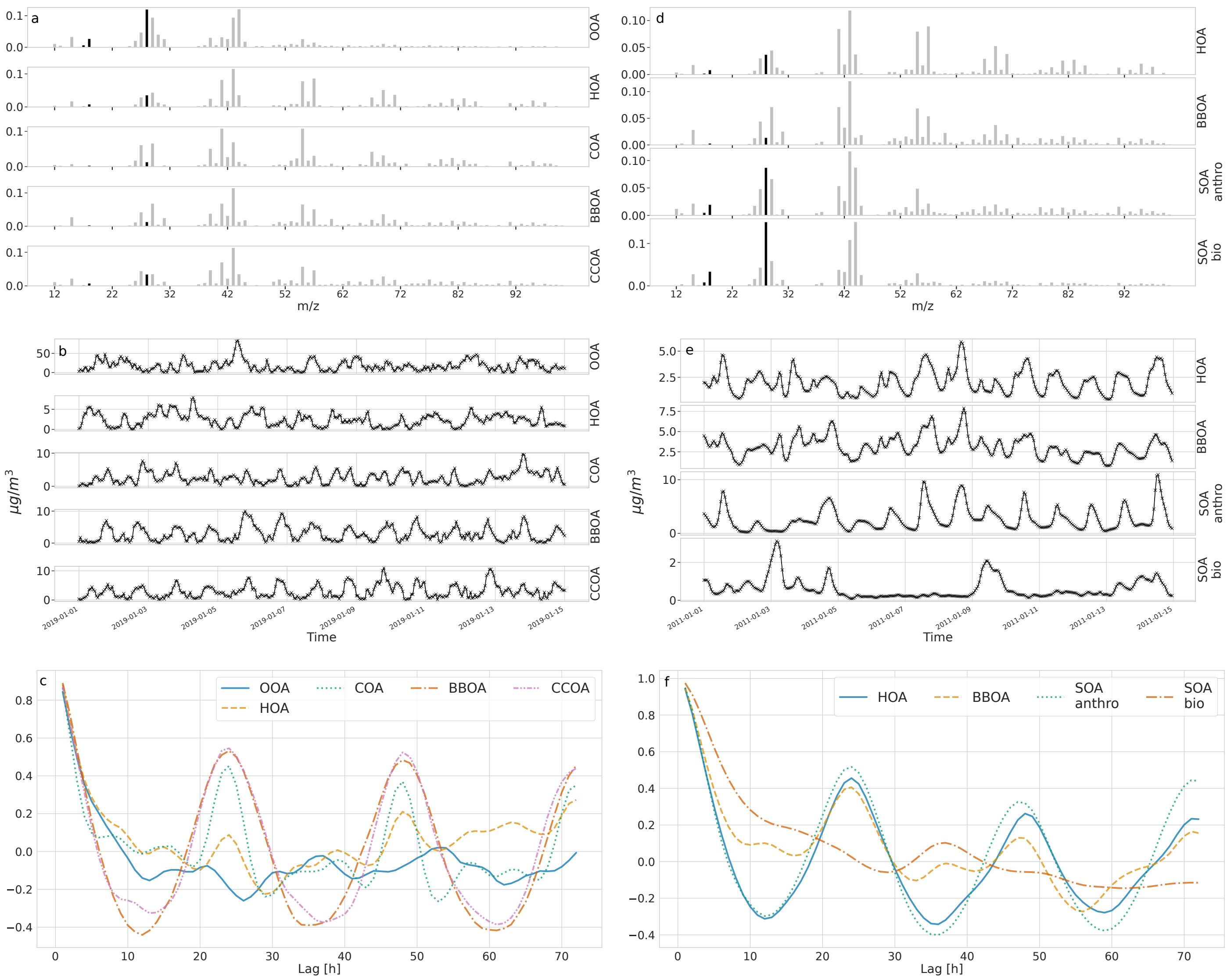

Figure 2. Characteristics of the synthetic ToF-ACSM OA datasets. Panels a and d show the factor profiles ($\boldsymbol{F}$) used to construct the synthetic datasets. The solid black bars are ions derived from m/z 44 and are only used for converting concentrations. Panels b and e show the factor time series ($\boldsymbol{G}$) used to construct each dataset, and the unit for them is $\mu g/m^3$, and panels c and f show each factor's temporal auto-correlation. Auto-correlation refers to the Pearson correlation coefficient of the component with the same component time-shifted by the number of hours. Panels a,b and c are for megacity data and panels d, e, and f are for the European urban environment.

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

0. However, they are below unity and within the error estimate, judging by the normalized residuals. Therefore, PMF likely finds solutions that fit the noise in the data better. All other reconstruction metrics (**S** mean, **S** standard deviation, Q_m*) are comparable for all models.

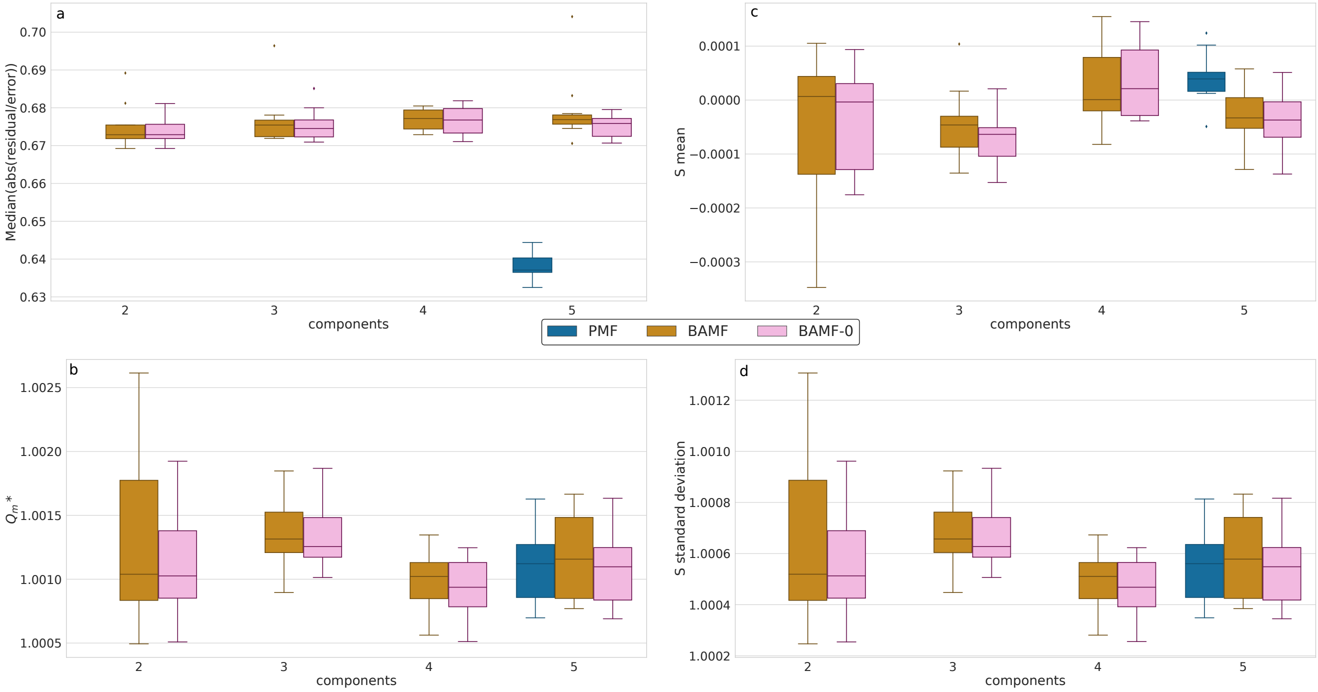

Figure 3. Reconstruction metrics for BAMF, BAMF-0, and PMF for synthetic megacity data. Panel a shows the relative error of **X** of the median solution as a function of the number of components for the synthetic megacity data (BAMF and BAMF-0: results from 10 different datasets for each number of factors, PMF: only for the 5-factor cases). Panel b shows the Q_m* statistic, where all models are very similar. Panel c shows the mean and Panel d shows the standard deviation of **S**, as a function of the number of factors for the synthetic megacity data (the ideal value for the mean is 0 and the ideal value for standard deviation is smaller than 1). **S** refers to the difference between the original data and the samples normalized with the standard deviation of the samples (uncertainty of the model).

Data reconstruction is essential to get within the error limits. However, source apportionment aims to accurately and precisely resolve the actual components in **G** and **F**, i.e. factorization performance. Each model's solution to the example dataset is shown in Figure 4. For this example, we observe that all models resolve all five components. However, BAMF has a better factorization performance both in **F** & **G** than the models not accounting for auto-correlation (BAMF-0, PMF) in Table 1. Similarly, for the diurnal cycles in the example in Figure 4c, the OA sources' diurnal concentration, as identified by BAMF, closely resembles the ground truth components. While all models capture the time behaviour of the diurnals, the absolute magnitude has a bias, with BAMF having a substantially smaller bias than the other models for four out of five components (Table 1).

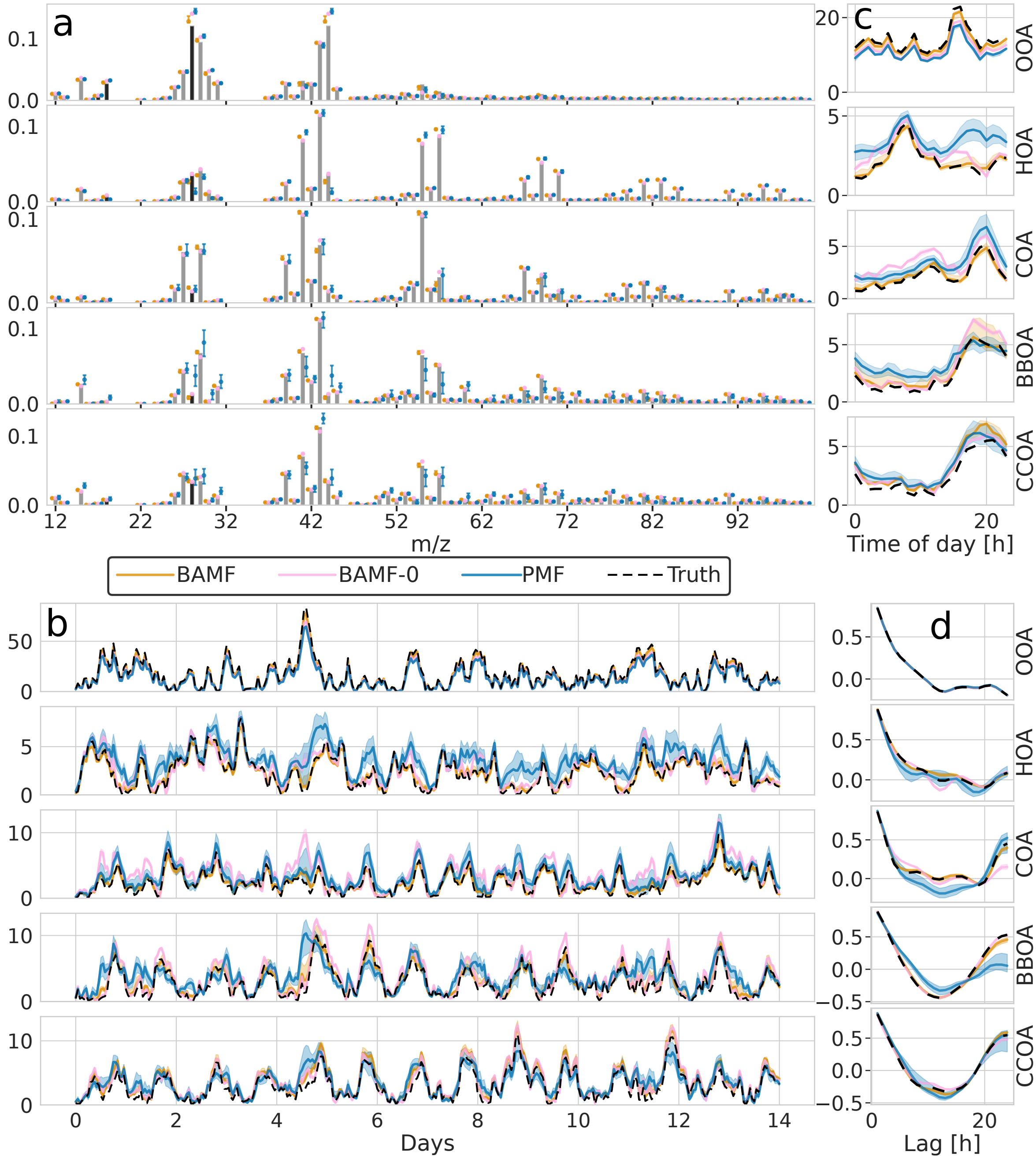

Figure 4. Illustration of **F** and **G** reconstruction of all models for one of the synthetic megacity ToF-ACSM OA datasets with five components. Panel a) is **F**, b) is **G**, c) is the diurnal concentration variation and d) is autocorrelation measured as Pearson correlation coefficient. Here, we display the median and the 0.25 and 0.75 quantiles. For all BAMF-type models, these quantities are computed based on the samples and for PMF, based on the 100 solutions.

		BAMF	BAMF-0	PMF
Reconstruction performance:	Median($\|\boldsymbol{Z}-\boldsymbol{X}\|/\boldsymbol{\sigma}$)	0.68	0.68	**0.64**
	Max($\|\boldsymbol{Z}-\boldsymbol{X}\|/\boldsymbol{\sigma}$)	5.85	5.07	**3.72**
Factorization performance:	**G** / Truth OOA	**0.94**	0.83	0.78
	G r OOA	**1.00**	**1.00**	**1.00**
	F ρ OOA	**0.99**	0.90	0.93
	diurnal **G** / Truth OOA	**0.94**	0.83	0.78
	G / Truth HOA	**0.99**	1.16	1.51
	G r HOA	**0.99**	0.92	0.86
	F ρ HOA	0.99	**1.00**	0.97
	diurnal **G** / Truth HOA	**1.01**	1.21	1.57
	G / Truth COA	**0.99**	1.44	1.39
	G r COA	**0.98**	0.78	0.95
	F ρ COA	0.99	**1.00**	**1.00**
	diurnal **G** / Truth COA	**1.03**	1.53	1.47
	G / Truth BBOA	**1.08**	1.26	1.28
	G r BBOA	0.98	**1.00**	0.66
	F ρ BBOA	0.99	**1.00**	0.94
	diurnal **G** / Truth BBOA	**1.10**	1.26	1.37
	G / Truth CCOA	1.24	1.20	**1.17**
	G r CCOA	**0.99**	0.98	0.86
	F ρ CCOA	**1.00**	0.99	0.96
	diurnal **G** / Truth CCOA	1.27	**1.19**	1.27

Table 1. Reconstruction and factorization performance of all three models for the synthetic megacity ToF-ACSM OA dataset in Figure 4. The reconstruction metrics measure the residuals divided by the error estimate. A value closer to 0 is better and a value below 1 is smaller than the error estimate given to the model. The factorization performance is assessed via three metrics: **G / Truth** is the average ratio of each factor time series, r is the Pearson correlation coefficient between the factor time series, and ρ is the Spearman correlation coefficient between the factor profiles, For the factorization performance, a value closer to 1 is better. For each metric, the best value is highlighted in bold.

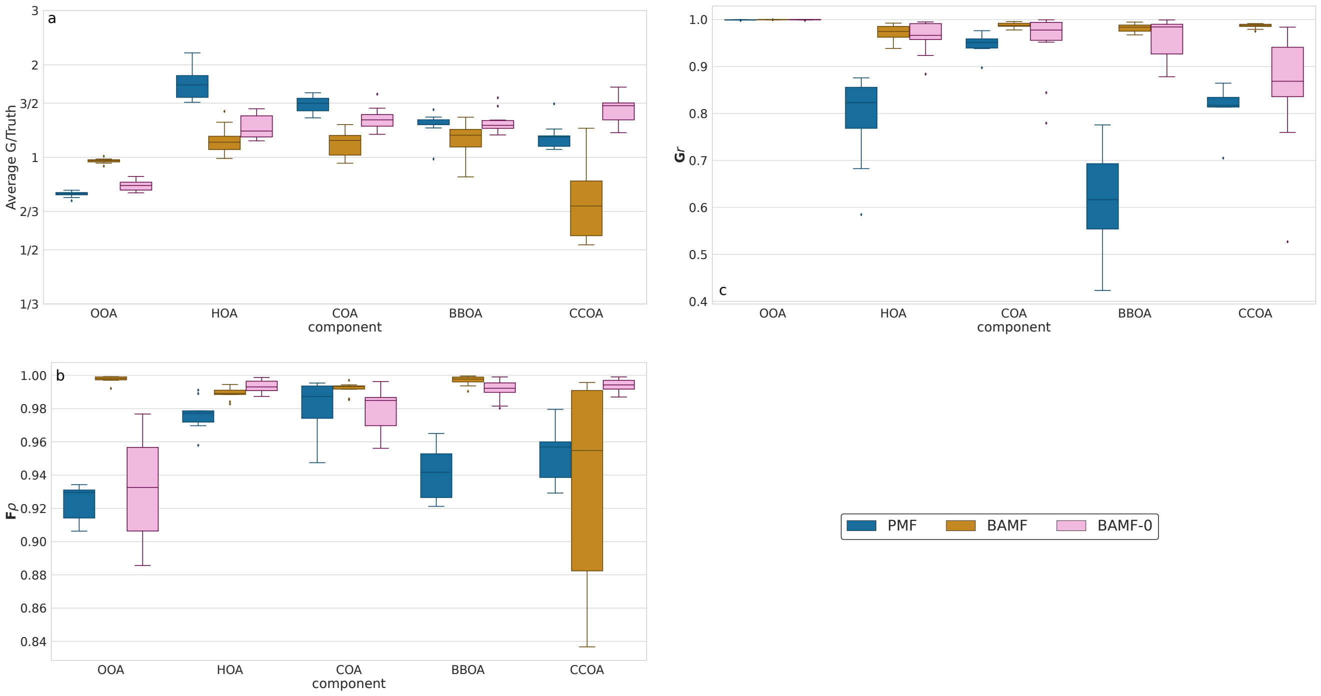

Figure 5. Summary of factorization performance of the three models for all synthetic megacity ToF-ACSM OA datasets with five components (10 datasets): Panel a shows the mean of the median value of the components of **G** divided by the true value (ideal value is 1). Panel b shows the Spearman correlation coefficient median solution components of **F** compared to the true value (a value of 1 refers to a perfect correlation). Panel c shows the Pearson correlation coefficient of the median solution components of **G** compared to the true value (a value of 1 refers to a perfect linear correlation).

When considering all ten synthetic datasets with five components mimicking a polluted megacity, BAMF consistently produces factors closer in magnitude to the truth and correlate better with the actual factors than the other models (Figure 5). Thus, BAMF is better correlated with the time behaviour of the components, while one of the components (CCOA, strongly cross-correlated with BBOA in terms of G) is difficult for all models. Overall, using auto-correlation in source apportionment markedly improves the quality of the resolved factors while keeping the overall reconstruction metrics similar.

4.2 Simulated European low pollution city source apportionment

In a second exercise, we assessed the performance of BAMF, BAMF-0, and PMF on a synthetic dataset mimicking the conditions in a typical European city (Section 3.2). In contrast to the fully synthetic dataset in Section 3.1, here, the true components **G** are OA source components computed by an air quality model (Jiang et al., 2019). This provides **G** time series close to the atmosphere while still knowing the ground truth. While the three models show somewhat different components (Figure 6), the reconstruction metrics indicate that all models have acceptable solutions (Table 2). In fact, the metrics also show that the

European dataset is reconstructed almost within the error limits with already only three components, i.e. 1 component less than is present in the synthetic dataset (HOA, BBOA, $\mathrm{SOA_{anthro}}$, $\mathrm{SOA_{bio}}$). This could explain why there is significant freedom in acceptable 4-component solutions and variation between them. All models show signs of mixing between the components, likely due to the cross-correlation of the true **G** components (time behaviour is very similar) as well as similar chemical signatures **F**. However, it is worth noting that BAMF has a bias on several components in **G** as seen in Table 2, but otherwise reflects their time behaviour well. PMF, on the other hand, produces two almost identical components (both in **F** and **G**) for two of the four components, and PMF cannot thus distinguish the components present in the dataset. Overall, all models are challenged by the European dataset, with BAMF having the most consistent performance.

4.3 Resolving an unknown amount of sources

For real-world source apportionment analyses, the true amount of components, i.e. sources, to be resolved via matrix factorization is unknown yet crucial. Despite the importance, accurately determining and specifying the correct number of modelled components is not trivial, see, e.g. Isokääntä et al. (2020); Ulbrich et al. (2009); Zhang et al. (2011). Typical strategies rely on reconstructing **X** within the measurement error, the absence of structure in the measurement error weighted residuals and the resolved components' environmental interpretability. Here, we assess the behaviour of the BAMF, BAMF-0, and PMF models as the number of components are changed on the 4-component chemical transport model dataset from Section 3.2. The model runs were performed both with an underspecified setting using three components and overspecified setting using five components (Figures 7 and 8). While underspecified, all models extract an $\mathrm{SOA_{anthro}}$ component and merge the remaining three components into two. At the same time, BAMF extracts a component similar to $\mathrm{SOA_{bio}}$, BAMF-0 and PMF extract components similar to HOA and BBOA but lack $\mathrm{SOA_{bio}}$.

For the overspecified models (5 instead of 4 components), the results differ (Figure 8). While the models without the auto-correlation assumption (PMF, BAMF-0) split the true components into multiple sub-components (mostly the POA components, HOA and BBOA), BAMF produces an extra component that is easily identifiable as unnecessary in addition to the four components resolved with four factors. This unnecessary component is characterized by an extremely high auto-correlation and low magnitude, as seen in Figure 8b, c, and d. The extra component doesn't affect the factorization performance of BAMF's other components substantially. At the same time, BAMF-0 and PMF have a reduced factorization performance with too many components (Figure 8, Table 2). It should be noted that the behaviour of BAMF can be changed by adding new model terms, such as constraints on **F**. For example, the model minimizes the constrained component when it is run equally overspecified (5 instead of 4 components) but with a priori information on **F** as can be seen in the Annex (Appendix B).

4.4 Using ancillary information to improve resolving sources

As highlighted above in Section 4.2, all models are challenged by the European dataset. Imperfect matrix factorization results are likewise often observed when using PMF for real-world chemical datasets, see, e.g., Canonaco et al. (2013); Daellenbach et al. (2017). Often, information is available that could help resolve the sources, such as chemical fingerprints of specific components associated with different sources. In current practice, previously observed **F** profiles are often used as boundary

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

Atmospheric Measurement Techniques
Discussions

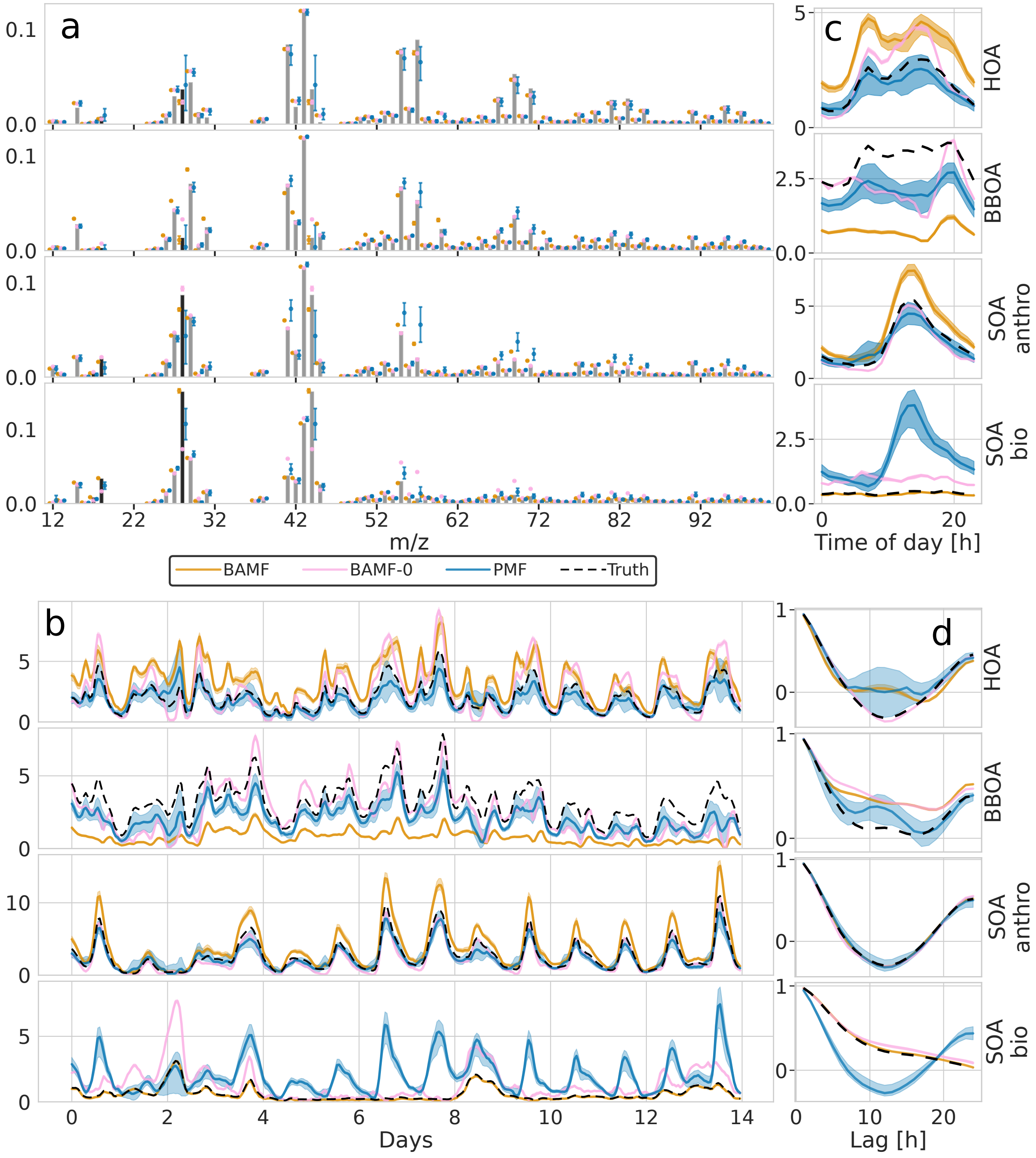

Figure 6. Factorization performance of all three models for the synthetic European city ToF-ACSM OA dataset. The shaded area is the interquartile range (0.25 to 0.75 quantile). Panel a) is **F**, b) is **G**, c) is the diurnal and d) is autocorrelation measured as Pearson correlation coefficient.

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

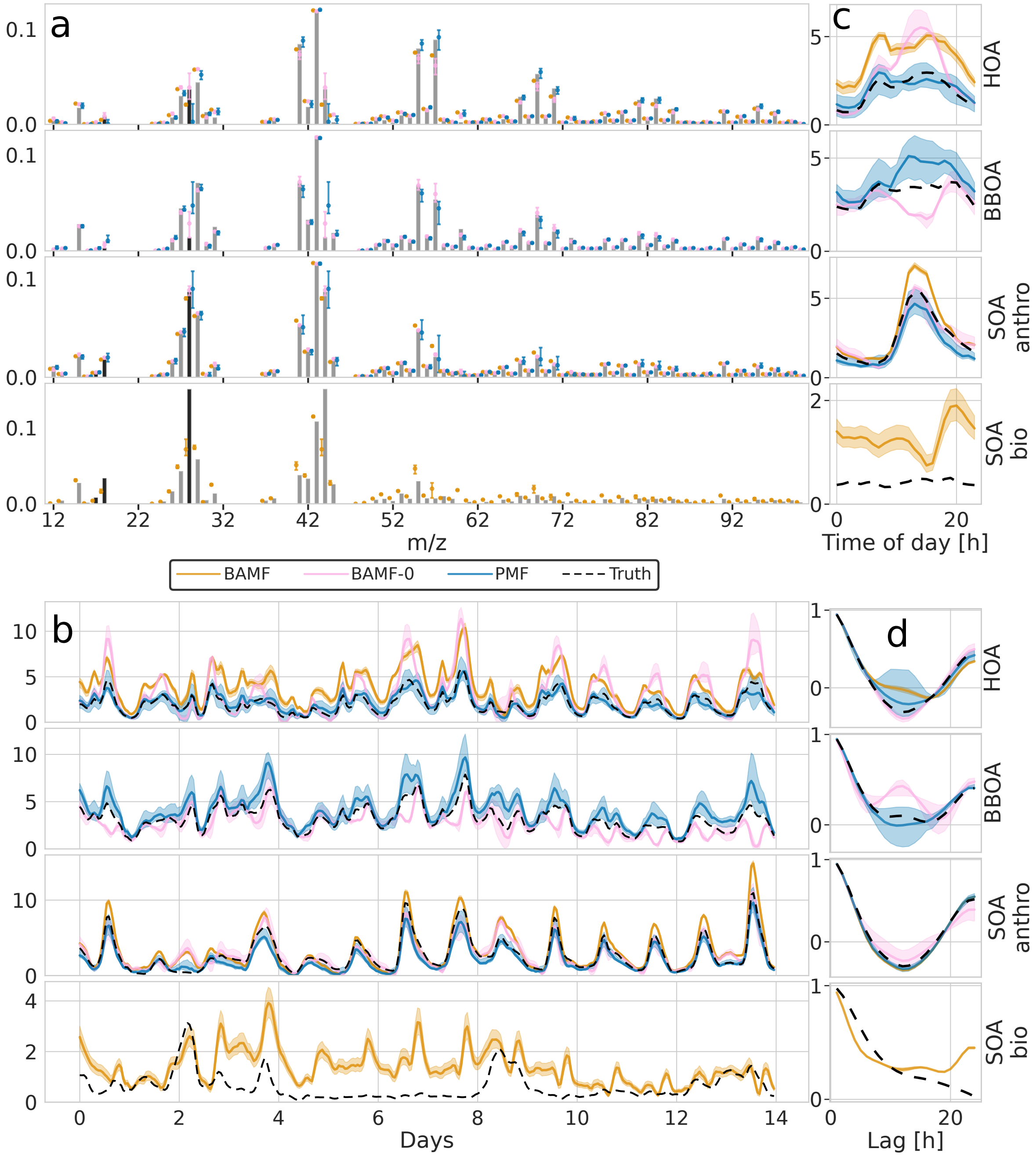

Figure 7. Factorization performance of underspecified (3 components) models for the synthetic European city ToF-ACSM OA dataset. Panel a) is **F**, b) is **G**, c) is the diurnal and d) is autocorrelation measured as Pearson correlation coefficient. The models extract different components and they are shown next to the closest original component. This is why there are 4 components shown even though the models extract only 3 components each.

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

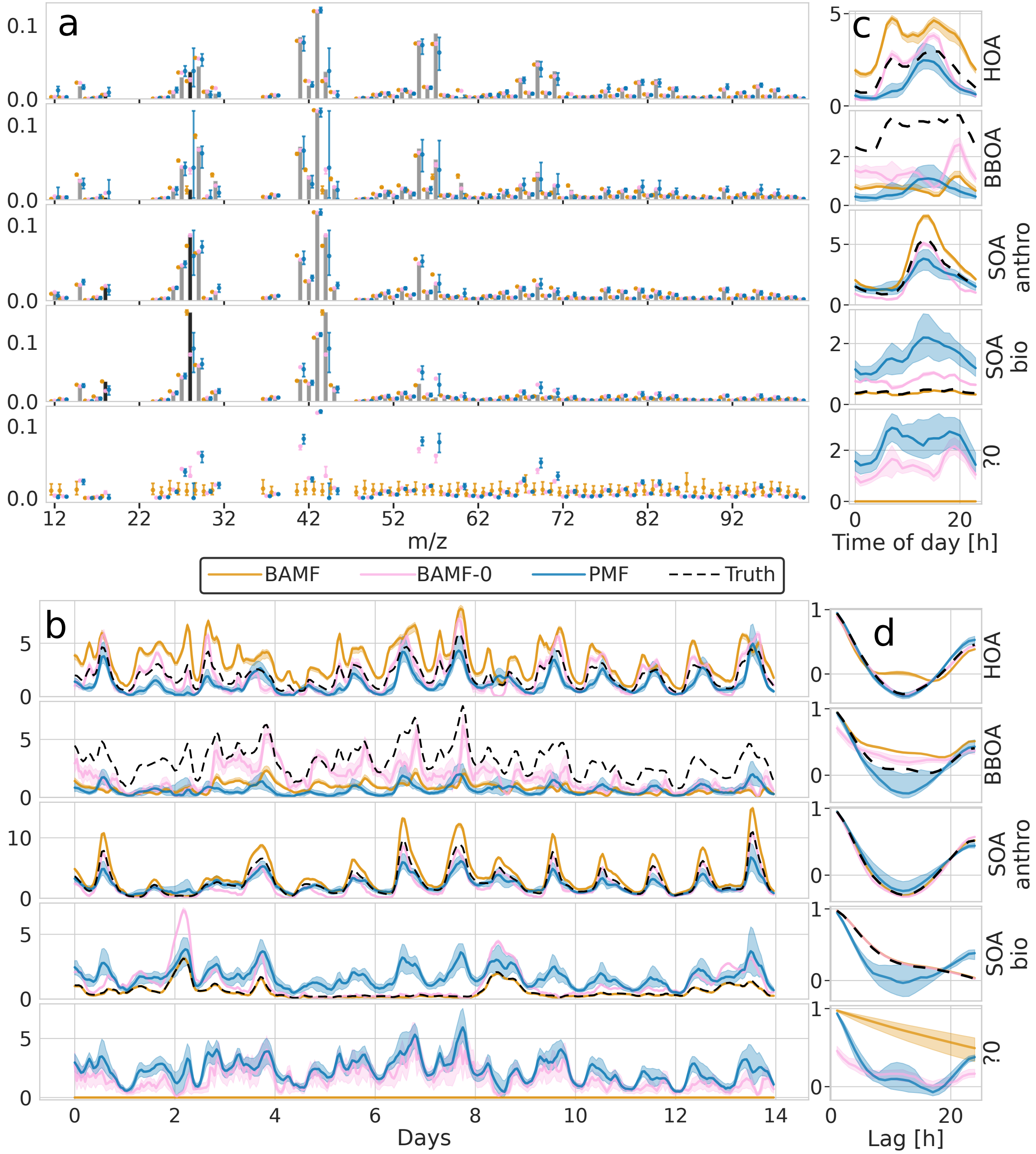

Figure 8. Factorization performance of overspecified (5 components) models for the synthetic European city ToF-ACSM OA dataset. Panel a) is **F**, b) is **G**, c) is the diurnal and d) is autocorrelation measured as Pearson correlation coefficient. "?0" denotes an unidentified component.

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

conditions in source apportionment analyses. This approach has significant uncertainty in the general case because true **F** is unknown. Still, in our specific test case, the a priori information is precisely correct—the known information on **F** of the true components.

We tested the models' performance when using a priori information on **F** using three different approaches on the European dataset (from Section 3.2):

1. Full constraint: for BAMF-C and PMF, the two POA components (HOA & BBOA) were fully (for all m/zs) constrained with a roughly 18% allowed deviation from the anchor (see Section 2.2.2 for the determination of this).

2. Incomplete constraint: For the BAMF model, a priori information was only used in a limited m/z range (12–60) instead of for all m/zs (m/z 12–100), same allowed deviation from the anchor for the constrained components.

3. Partial constraint: For the BAMF model, a priori information was only used for 4 out of 74 m/z peaks (m/zs 45, 57 and 60 compared to m/z 43) defined with the same allowed deviation from the anchor for constrained components.

The reconstruction and factorization performance of the different models are compared in Table 2. The fully constrained BAMF model (BAMF-C) performs substantially better in extracting BBOA both in **F** and **G** compared to BAMF. In fact, the extracted components are very similar to the true components with very similar temporal behaviour and reduced biases in **G** (Figure 9). Fully constrained PMF performs marginally better than unconstrained PMF but still mixes $\mathrm{SOA_{bio}}$ with $\mathrm{SOA_{anthro}}$, Figure 9. This can be expected since the a priori information is applied to HOA and BBOA, not to the SOA components. The incompletely constrained BAMF model performs slightly worse than the fully constrained BAMF model. Partially constrained BAMF performs worse but is still on par with the fully constrained PMF (Table 2). The partially constrained BAMF model reduces bias in BBOA but starts mixing SOA bio with the other components (Figure 10, Table 2). Using a priori information on **F** improves the factorization performance of both PMF and especially BAMF, with more information leading to solutions closer to the ground truth. This is especially helpful when the additional information is on the components the model mixes when unconstrained.

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

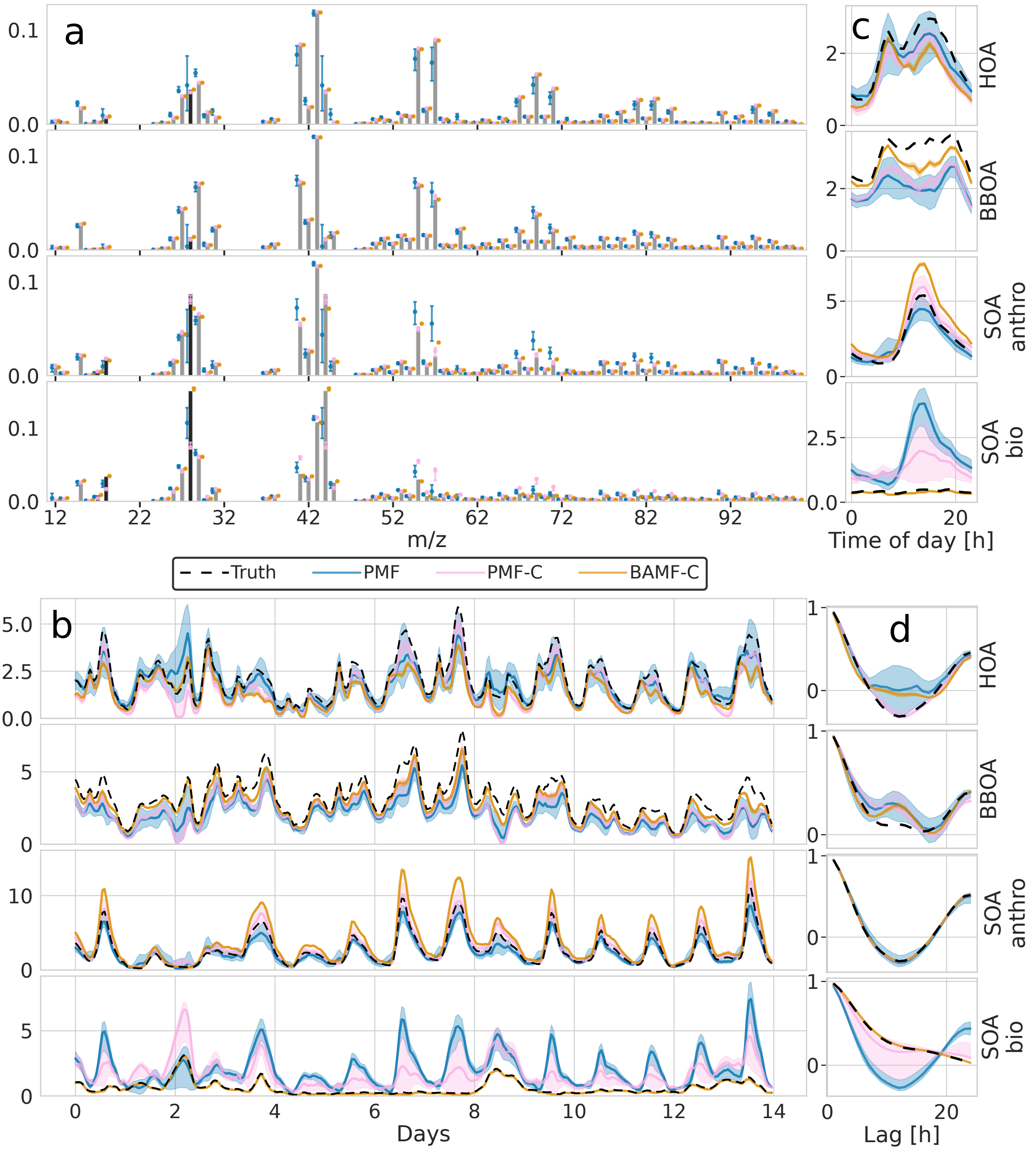

Figure 9. Factorization performance of models using a priori information on **F** for the synthetic European city ToF-ACSM OA dataset, fully constrained BAMF and fully constrained PMF results compared to unconstrained PMF. Panel a is **F**, b is **G**, c is the diurnal concentration, and d is the auto-correlation behaviour.

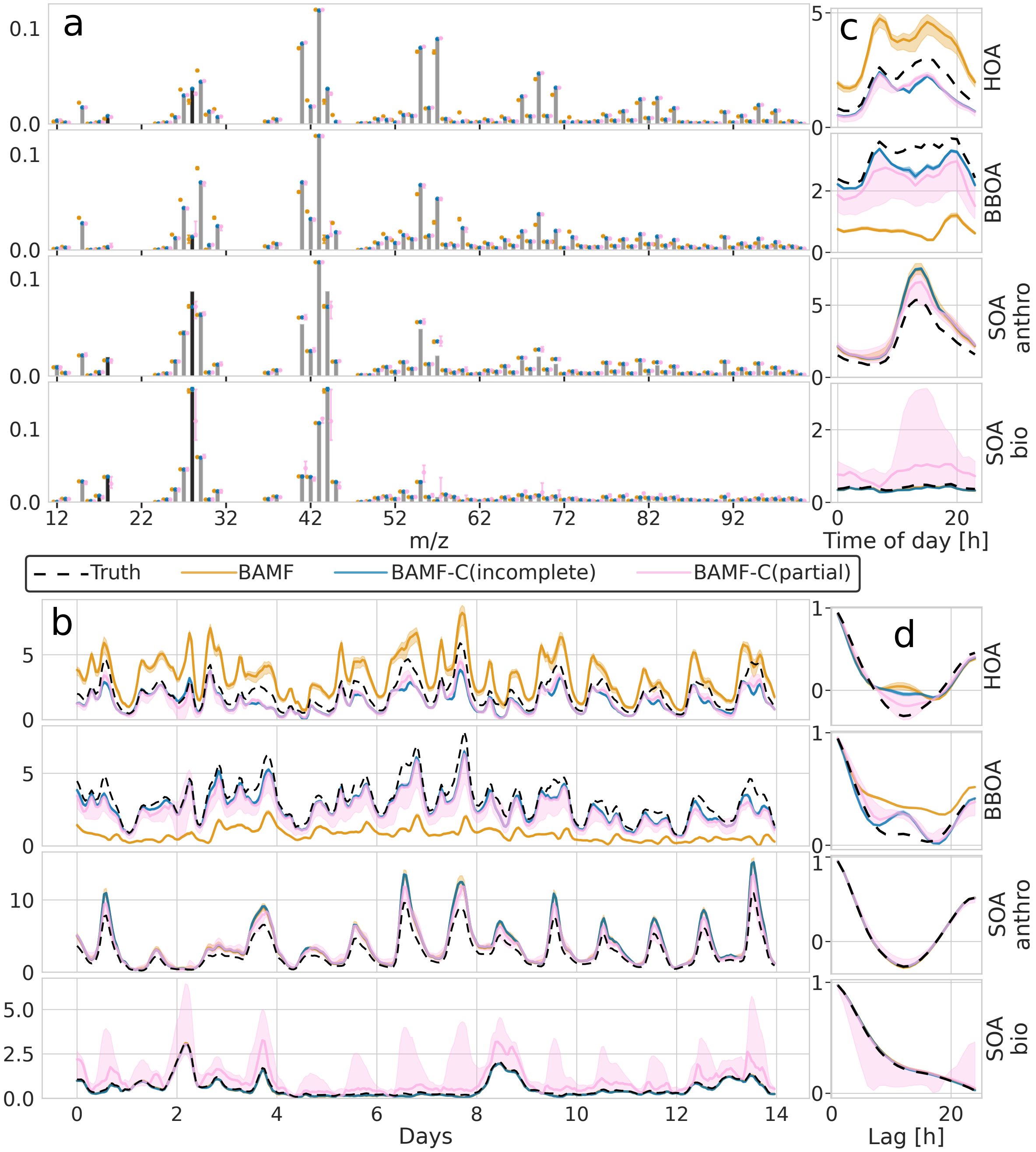

Figure 10. Factorization performance of models using a priori information on **F** for the synthetic European city ToF-ACSM OA dataset, partial and incomplete constraints in BAMF compared to unconstrained BAMF. Panel a is **F**, b is **G**, c is the diurnal concentration and d is the auto-correlation behaviour. The results for SOA components overlap between BAMF and BAMF-C(incomplete) such that the BAMF results are not visible in panels b, c, and d.

			$\lvert \boldsymbol{X} - \boldsymbol{Z} \rvert / \boldsymbol{\sigma}$		HOA				SOA anthro				SOA bio				BBOA				POA
			Median	Max	**G** / True	**G** r	**F** ρ	Diurn	**G** / True	**G** r	**F** ρ	Diurn	**G** / True	**G** r	**F** ρ	Diurn	**G** / True	**G** r	**F** ρ	Diurn	/ SOA
BAMF	Full	3	1.18	54.56	0.78	0.98	0.97	0.80	1.55	0.95	0.99	1.53					0.85	0.97	0.99	0.86	1.01
		4	0.67	5.13	0.72	0.91	1.00	0.74	1.39	1.00	0.99	1.39	0.91	1.00	0.96	0.88	0.86	0.97	1.00	0.87	0.96
		5	0.67	5.14	0.04	0.16	1.00	0.05	1.41	1.00	0.99	1.42	0.96	1.00	0.95	0.93	0.50	0.77	1.00	0.48	0.38
	Incomplete	4	0.67	5.06	0.71	0.91	1.00	0.73	1.41	1.00	0.98	1.41	0.90	1.00	0.96	0.87	0.85	0.97	1.00	0.87	0.95
	Partial	3	1.18	53.08	0.96	0.90	0.96	1.00	1.10	0.93	0.98	1.11					0.96	0.91	0.97	0.95	1.66
		4	0.67	4.98	0.73	0.96	1.00	0.74	1.32	0.99	0.99	1.33	1.59	0.91	0.98	1.96	0.80	0.97	1.00	0.75	0.88
		5	0.67	5.03	0.28	0.93	1.00	0.30	1.40	1.00	0.99	1.40	0.90	1.00	0.96	0.91	0.10	0.17	0.99	0.12	0.21
	None	3	1.18	52.80	1.91	0.91	0.95	2.00	1.21	0.96	0.99	1.22	2.24	0.35	0.88	3.12					0.83
		4	0.67	4.96	1.70	0.91	0.95	1.77	1.39	1.00	0.99	1.39	0.91	1.00	0.96	0.89	0.24	0.74	0.95	0.24	0.97
		5	0.67	5.07	1.70	0.91	0.95	1.77	1.36	1.00	0.99	1.36	0.92	1.00	0.96	0.91	0.24	0.75	0.96	0.23	0.99
BAMF-0	None	3	1.18	52.60	1.48	0.96	0.95	1.48	1.01	0.93	0.99	1.01					0.85	0.78	0.98	0.86	2.07
		4	0.67	5.17	1.23	0.97	0.95	1.22	0.82	0.99	1.00	0.83	2.24	0.98	0.91	2.25	0.77	0.74	0.98	0.72	1.37
		5	0.67	5.24	0.96	0.96	0.96	0.97	0.76	0.98	1.00	0.78	1.95	1.00	0.92	1.90	0.53	0.70	0.97	0.44	1.11
PMF	Full	3	1.10	82.93	0.82	0.97	1.00	0.85	1.79	0.97	0.98	1.78					0.62	0.80	1.00	0.65	0.74
		4	0.63	4.01	0.73	0.97	1.00	0.74	1.12	1.00	1.00	1.13	2.78	0.93	0.92	3.15	0.71	0.90	1.00	0.70	0.78
		5	0.62	4.29	0.70	0.97	1.00	0.71	0.57	1.00	0.99	0.58	1.79	0.83	0.96	2.10	0.54	0.88	1.00	0.52	1.18
	None	3	1.12	84.37	1.05	0.93	0.89	1.06	0.78	0.98	0.99	0.80					1.28	0.94	0.97	1.25	2.88
		4	0.63	4.01	0.90	0.92	0.95	0.88	0.89	0.98	0.95	0.93	3.29	0.33	0.93	4.33	0.65	0.93	0.97	0.66	0.87
		5	0.62	4.47	0.62	0.84	0.95	0.61	0.82	0.98	0.98	0.85	2.81	0.63	0.94	3.76	0.21	0.68	0.88	0.20	0.48

Table 2. Reconstruction and factorization performance for the synthetic European city ToF-ACSM OA dataset. **G / True** is the average ratio of the component to the true one, r is the Pearson correlation coefficient, and ρ is the Spearman correlation coefficient. Diurn is a factor's average ratio of the diurnal of G to the true diurnal. POA / SOA is the ratio of primary (HOA, BBOA) and secondary organic aerosol (anthropogenic SOA, biogenic SOA); for the ground truth, this ratio is 1.55. Unidentified components were not included in this ratio. For the reconstruction metrics, values closer to 0 are better, and values below 1 are within the error given to the model. For factorization metrics, the ideal value is 1.

5 Conclusions

Most observations in the natural world are auto-correlated. Air pollutants, as well as their sources, are no different. We present a Bayesian matrix factorization model that accounts for components' temporal auto-correlation (BAMF) and provides direct error estimation. BAMF is built on top of STAN, a freely available, robust, actively developed, open-source framework for statistical modelling with the ability of full Bayesian statistical inference with Markov-Chain-Monte-Carlo sampling. Here, we characterize BAMF's performance on synthetic Time-of-Flight Aerosol Chemical Speciation Monitor mass spectral OA data compared to PMF. This approach allows for assessing the model's performance based on input data reconstruction and the ability to accurately model components' chemical composition and concentration time series. Without strongly temporally cross-correlated components, BAMF resolves auto-correlated components well (synthetic megacity dataset). Both BAMF and PMF are challenged by strongly temporally cross-correlated components (European data). Further, we show that using a priori information on the components' chemical composition improves BAMF factorization performance such that all components are well represented. Overall, we believe BAMF-type models are promising tools for source apportionment and deserve further research.

Code and data availability. The datasets will be available upon publication. The code will be available under an open source license upon publication.

Author contributions. AR, AB, KD, and KP participated in the model, dataset, and experiment development. KD and MM contributed to the PMF model solutions. JJ did the transport model runs. All authors contributed to the writing of the manuscript.

340 *Competing interests.* The authors declare to have no conflicts of interests.

Acknowledgements. We thank the Academy of Finland for its support (decisions 337549, 345704, and 346376). KRD acknowledges support by SNSF Ambizione grant PZPGP2_201992.

Appendix A: Concentration and uncertainty at selected m/zs for synthetic megacity ToF-ACSM OA data

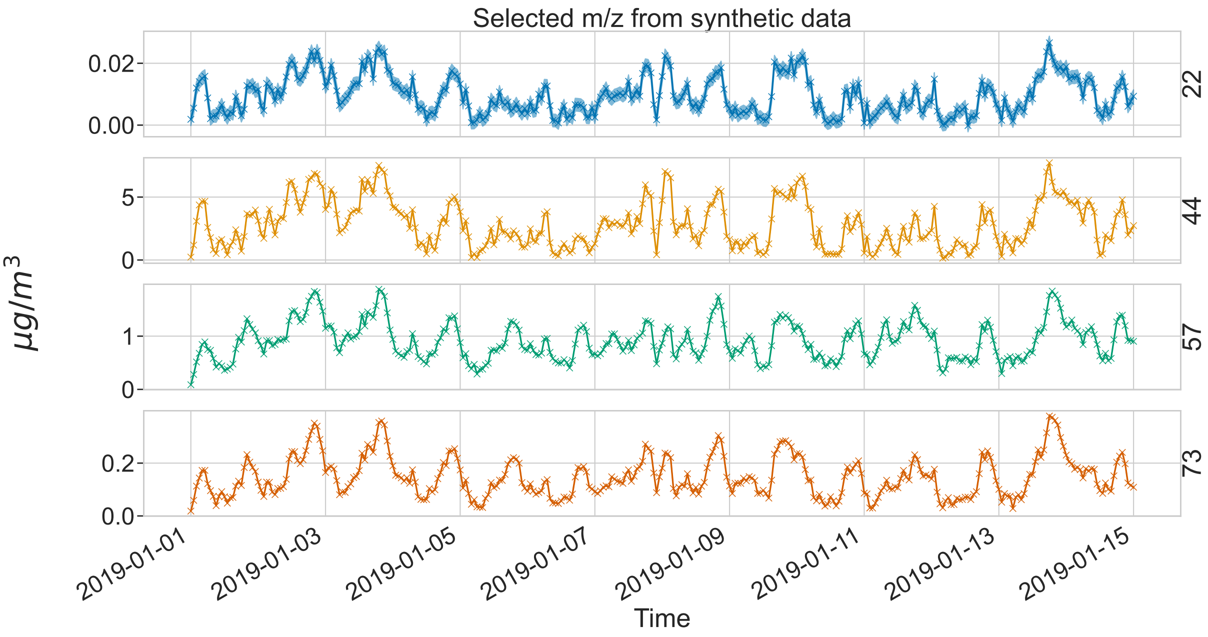

Figure A1. Concentration and uncertainty time series at selected m/zs for synthetic megacity ToF-ACSM OA data. The shaded area contains 95% of the probability mass of the Gaussian distribution of the error.

Appendix B: Overspecified BAMF-C results

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

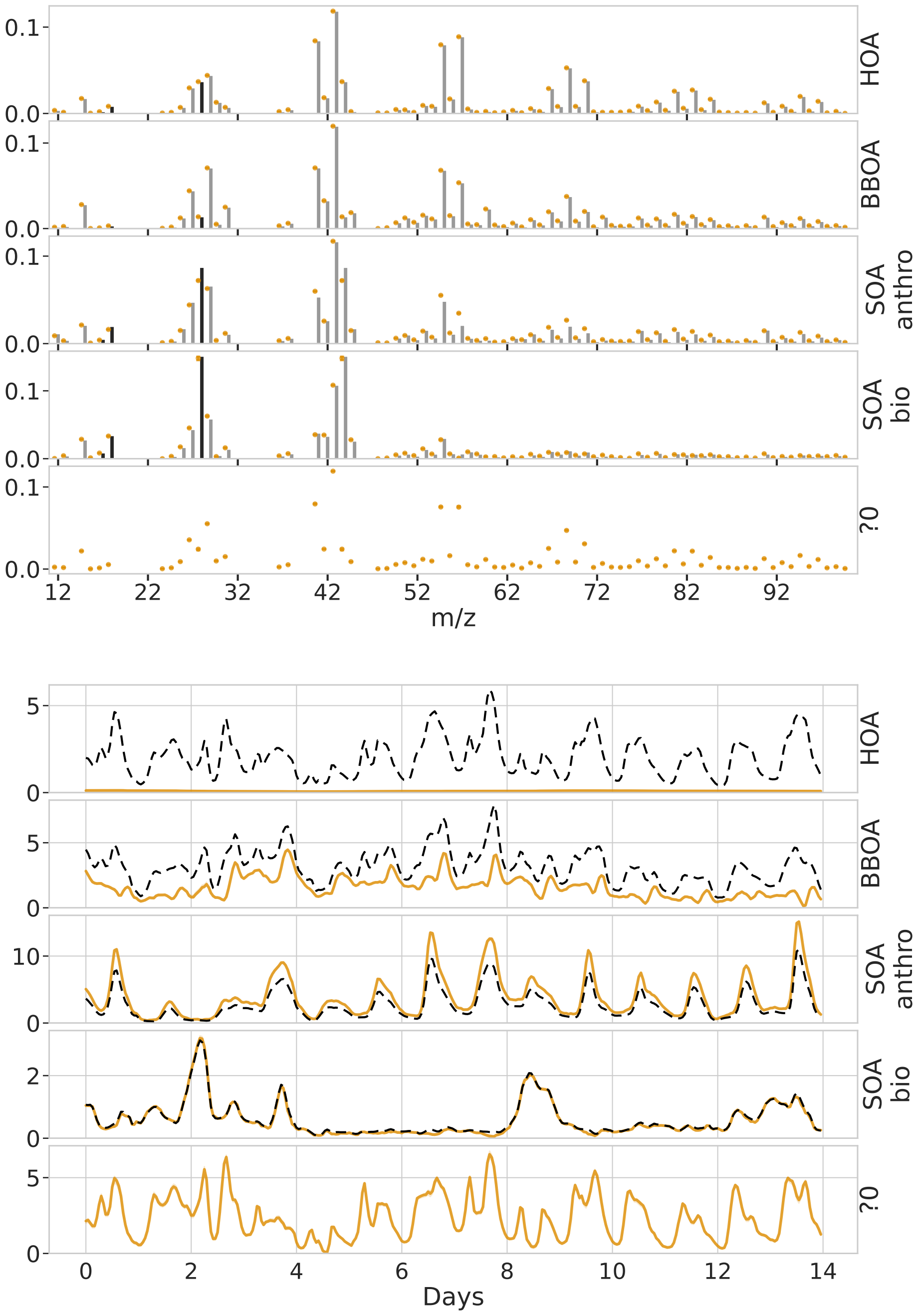

Figure B1. Results from overspecified BAMF-C model for the synthetic European city ToF-ACSM OA dataset with 5 modelled components instead of 4 and HOA & BBOA fully constrained.

https://doi.org/10.5194/amt-2023-70
Preprint. Discussion started: 12 May 2023

References

Allan, J. D., Jimenez, J. L., Williams, P. I., Alfarra, M. R., Bower, K. N., Jayne, J. T., Coe, H., and Worsnop, D. R.: Quantitative sampling using an Aerodyne aerosol mass spectrometer 1. Techniques of data interpretation and error analysis, Journal of Geophysical Research - Atmospheres, 108, 4090–, 2003.

Bates, J. T., Fang, T., Verma, V., Zeng, L., Weber, R. J., Tolbert, P. E., Abrams, J. Y., Sarnat, S. E., Klein, M., Mulholland, J. A., and Russell, A. G.: Review of Acellular Assays of Ambient Particulate Matter Oxidative Potential: Methods and Relationships with Composition, Sources, and Health Effects, Environmental Science & Technology, 53, 4003–4019, https://doi.org/10.1021/acs.est.8b03430, 2019.

Canagaratna, M., Jayne, J., Jimenez, J., Allan, J., Alfarra, M., Zhang, Q., Onasch, T., Drewnick, F., Coe, H., Middlebrook, A., Delia, A., Williams, L., Trimborn, A., Northway, M., DeCarlo, P., Kolb, C., Davidovits, P., and Worsnop, D.: Chemical and microphysical characterization of ambient aerosols with the aerodyne aerosol mass spectrometer, Mass spectrometry reviews, 26, 185–222, 2007.

Canonaco, F., Crippa, M., Slowik, J. G., Baltensperger, U., and Prévôt, 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, Atmospheric Measurement Techniques, 6, 3649–3661, https://doi.org/10.5194/amt-6-3649-2013, 2013.

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, Atmospheric Measurement Techniques, 14, 923–943, https://doi.org/10.5194/amt-14-923-2021, 2021.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A.: Stan: A Probabilistic Programming Language, Journal of Statistical Software, Articles, 76, 1–32, https://doi.org/10.18637/jss.v076.i01, 2017.

Crippa, M., DeCarlo, P. F., Slowik, J. G., Mohr, C., Heringa, M. F., Chirico, R., Poulain, L., Freutel, F., Sciare, J., Cozic, J., Di Marco, C. F., Elsasser, M., Nicolas, J. B., Marchand, N., Abidi, E., Wiedensohler, A., Drewnick, F., Schneider, J., Borrmann, S., Nemitz, E., Zimmermann, R., Jaffrezo, J.-L., Prévôt, A. S. H., and Baltensperger, U.: Wintertime aerosol chemical composition and source apportionment of the organic fraction in the metropolitan area of Paris, Atmospheric Chemistry and Physics, 13, 961–981, https://doi.org/10.5194/acp-13-961-2013, 2013.

Crippa, M., Canonaco, F., Lanz, V. A., Äijälä, 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. A., Mohr, C., Nemitz, E., O'Dowd, C., Ovadnevaite, J., Pandis, S. N., Petäjä, T., Poulain, L., Saarikoski, S., Sellegri, K., Swietlicki, E., Tiitta, P., Worsnop, D. R., Baltensperger, U., and Prévôt, A. S. H.: Organic aerosol components derived from 25 AMS data sets across Europe using a consistent ME-2 based source apportionment approach, Atmospheric Chemistry and Physics, 14, 6159–6176, https://doi.org/10.5194/acp-14-6159-2014, 2014.

Daellenbach, K. R., Stefenelli, G., Bozzetti, C., Vlachou, A., Fermo, P., Gonzalez, R., Piazzalunga, A., Colombi, C., Canonaco, F., Hueglin, C., Kasper-Giebl, A., Jaffrezo, J.-L., Bianchi, F., Slowik, J. G., Baltensperger, U., El-Haddad, I., and Prévôt, A. S. H.: Long-term chemical analysis and organic aerosol source apportionment at nine sites in central Europe: source identification and uncertainty assessment, Atmospheric Chemistry and Physics, 17, 13 265–13 282, https://doi.org/10.5194/acp-17-13265-2017, 2017.

Daellenbach, K. R., Uzu, G., Jiang, J., Cassagnes, L.-E., Leni, Z., Vlachou, A., Stefenelli, G., Canonaco, F., Weber, S., Segers, A., Kuenen, J. J. P., Schaap, M., Favez, O., Albinet, A., Aksoyoglu, S., Dommen, J., Baltensperger, U., Geiser, M., El Haddad, I., Jaffrezo, J.-L., and Prévôt, A. S. H.: Sources of particulate-matter air pollution and its oxidative potential in Europe, Nature, 587, 414–419, 2020.

Elser, M., Huang, R.-J., Wolf, R., Slowik, J. G., Wang, Q., Canonaco, F., Li, G., Bozzetti, C., Daellenbach, K. R., Huang, Y., Zhang, R., Li, Z., Cao, J., Baltensperger, U., El-Haddad, I., and Prévôt, A. S. H.: New insights into $PM_{2.5}$ chemical composition and sources in two major cities in China during extreme haze events using aerosol mass spectrometry, Atmospheric Chemistry and Physics, 16, 3207–3225, https://doi.org/10.5194/acp-16-3207-2016, 2016.

Fröhlich, R., Cubison, M. J., Slowik, J. G., Bukowiecki, N., Prévôt, A. S. H., Baltensperger, U., Schneider, J., Kimmel, J. R., Gonin, M., Rohner, U., Worsnop, D. R., and Jayne, J. T.: The ToF-ACSM: a portable aerosol chemical speciation monitor with TOFMS detection, Atmospheric Measurement Techniques, 6, 3225–3241, https://doi.org/10.5194/amt-6-3225-2013, 2013.

Huang, R.-J., Wang, Y., Cao, J., Lin, C., Duan, J., Chen, Q., Li, Y., Gu, Y., Yan, J., Xu, W., Fröhlich, R., Canonaco, F., Bozzetti, C., Ovadnevaite, J., Ceburnis, D., Canagaratna, M. R., Jayne, J., Worsnop, D. R., El-Haddad, I., Prévôt, A. S. H., and O'Dowd, C. D.: Primary emissions versus secondary formation of fine particulate matter in the most polluted city (Shijiazhuang) in North China, Atmospheric Chemistry and Physics, 19, 2283–2298, https://doi.org/10.5194/acp-19-2283-2019, 2019.

IPCC: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, vol. In Press, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896, 2021.

Isokääntä, S., Kari, E., Buchholz, A., Hao, L., Schobesberger, S., Virtanen, A., and Mikkonen, S.: Comparison of dimension reduction techniques in the analysis of mass spectrometry data, Atmospheric Measurement Techniques, 13, 2995–3022, https://doi.org/10.5194/amt-13-2995-2020, 2020.

Jiang, J., Aksoyoglu, S., El-Haddad, I., Ciarelli, G., Denier van der Gon, H. A. C., Canonaco, F., Gilardoni, S., Paglione, M., Minguillón, M. C., Favez, O., Zhang, Y., Marchand, N., Hao, L., Virtanen, A., Florou, K., O'Dowd, C., Ovadnevaite, J., Baltensperger, U., and Prévôt, A. S. H.: Sources of organic aerosols in Europe: a modeling study using CAMx with modified volatility basis set scheme, Atmospheric Chemistry and Physics, 19, 15 247–15 270, https://doi.org/10.5194/acp-19-15247-2019, 2019.

Kuhn, H. W.: The Hungarian method for the assignment problem, Naval research logistics quarterly, 2, 83–97, 1955.

Kulmala, M., Dada, L., Daellenbach, K. R., Yan, C., Stolzenburg, D., Kontkanen, J., Ezhova, E., Hakala, S., Tuovinen, S., Kokkonen, T. V., Kurppa, M., Cai, R., Zhou, Y., Yin, R., Baalbaki, R., Chan, T., Chu, B., Deng, C., Fu, Y., Ge, M., He, H., Heikkinen, L., Junninen, H., Liu, Y., Lu, Y., Nie, W., Rusanen, A., Vakkari, V., Wang, Y., Yang, G., Yao, L., Zheng, J., Kujansuu, J., Kangasluoma, J., Petäjä, T., Paasonen, P., Järvi, L., Worsnop, D., Ding, A., Liu, Y., Wang, L., Jiang, J., Bianchi, F., and Kerminen, V.-M.: Is reducing new particle formation a plausible solution to mitigate particulate air pollution in Beijing and other Chinese megacities?, Faraday discussions, 226, 334–347, 2021.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature, 525, 367–371, 2015.

NABEL: Swiss National Air Pollution Monitoring Network, https://www.bafu.admin.ch/bafu/en/home/topics/air/state/data/data-query-nabel.html, accessed: 2021-06-29.

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, 2011.

Paatero, P. and Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5, 111–126, 1994.

Reyes-Villegas, E., Green, D. C., Priestman, M., Canonaco, F., Coe, H., Prévôt, A. S. H., and Allan, J. D.: Organic aerosol source apportionment in London 2013 with ME-2: exploring the solution space with annual and seasonal analysis, Atmospheric Chemistry and Physics, 16, 15 545–15 559, https://doi.org/10.5194/acp-16-15545-2016, 2016.

Sage, A. M., Weitkamp, E. A., Robinson, A. L., and Donahue, N. M.: Evolving mass spectra of the oxidized component of organic aerosol: results from aerosol mass spectrometer analyses of aged diesel emissions, Atmospheric Chemistry and Physics, 8, 1139–1152, https://doi.org/10.5194/acp-8-1139-2008, 2008.

Schlag, P., Rubach, F., Mentel, T. F., Reimer, D., Canonaco, F., Henzing, J. S., Moerman, M., Otjes, R., Prévôt, A. S. H., Rohrer, F., Rosati, B., Tillmann, R., Weingartner, E., and Kiendler-Scharr, A.: Ambient and laboratory observations of organic ammonium salts in PM1, Faraday discussions, 200, 331–351, 2017.

Ulbrich, I., Handschy, A., Lechner, M., and Jimenez, J.: AMS Spectral Database, http://cires.colorado.edu/jimenez-group/AMSsd/, last access: 25.4.2022, 2022.

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, Atmospheric Chemistry and Physics, 9, 2891–2918, https://doi.org/10.5194/acp-9-2891-2009, 2009.

Wang, Y.-X. and Zhang, Y.-J.: Nonnegative matrix factorization: A comprehensive review, IEEE Transactions on Knowledge and Data Engineering, 25, 1336–1353, 2012.

Watson, J. G., Chow, J. C., and Fujita, E. M.: Review of volatile organic compound source apportionment by chemical mass balance, Atmospheric Environment, 35, 1567–1584, https://doi.org/10.1016/S1352-2310(00)00461-1, 2001.

Zhang, J., Li, R., Zhang, X., Bai, Y., Cao, P., and Hua, P.: Vehicular contribution of PAHs in size dependent road dust: A source apportionment by PCA-MLR, PMF, and Unmix receptor models, The Science of the total environment, 649, 1314–1322, 2019.

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Allan, J. D., Coe, H., Ulbrich, I., Alfarra, M. R., Takami, A., Middlebrook, A. M., Sun, Y. L., Dzepina, K., Dunlea, E., Docherty, K., DeCarlo, P. F., Salcedo, D., Onasch, T., Jayne, J. T., Miyoshi, T., Shimono, A., Hatakeyama, S., Takegawa, N., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Williams, P., Bower, K., Bahreini, R., Cottrell, L., Griffin, R. J., Rautiainen, J., Sun, J. Y., Zhang, Y. M., and Worsnop, D. R.: Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes, Geophysical research letters, 34, 2007.

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Ulbrich, I. M., Ng, N. L., Worsnop, D. R., and Sun, Y.: Understanding atmospheric organic aerosols via factor analysis of aerosol mass spectrometry: a review, Analytical and bioanalytical chemistry, 401, 3045–3067, 2011.

Zhang, Y., Favez, O., Canonaco, F., Liu, D., Močnik, G., Amodeo, T., Sciare, J., Prévôt, A. S. H., Gros, V., and Albinet, A.: Evidence of major secondary organic aerosol contribution to lensing effect black carbon absorption enhancement, NPJ climate and atmospheric science, 1, 2018.

Zhu, Q., Huang, X.-F., Cao, L.-M., Wei, L.-T., Zhang, B., He, L.-Y., Elser, M., Canonaco, F., Slowik, J. G., Bozzetti, C., El-Haddad, I., and Prévôt, A. S. H.: Improved source apportionment of organic aerosols in complex urban air pollution using the multilinear engine (ME-2), Atmospheric Measurement Techniques, 11, 1049–1060, https://doi.org/10.5194/amt-11-1049-2018, 2018.