Supplementary information for: Sparsity introduction in Bayesian models for aerosol source apportionment through the regularised horseshoe prior

Marta Via[1], Jure Demšar[2], Yufang Hao[3], Manousos Manousakas[3,4], Anton Rusanen[5], Jianhui Jiang[6], Stuart K. Grange[7], Jean-Luc Jaffrezo[8], Vy Ngoc Thuy Dinh[8], Gaëlle Uzu[8], Griša Močnik[1], and Kaspar R. Daellenbach[3]

[1]Center for Atmospheric Research, University of Nova Gorica, Ajdovščina 5270, Slovenia
[2]Faculty of Computer and Information Science, Tržaška Cesta 25, 1000 Ljubljana, Slovenia
[3]Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, 5232 Villigen PSI, Switzerland
[4]Environmental Radioactivity Aerosol Tech. for Atmospheric Climate Impacts, INRaSTES, National Centre of Scientific Research "Demokritos", Ag. Paraskevi, 15310, Greece
[5]Atmospheric Composition Research, Finnish Meteorological Institute, 00101 Helsinki, Finland
[6]School of Ecological and Environmental Sciences, East China Normal University, 200241, Shanghai, China
[7]Climate and Environmental Physics, Physics Institute, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland
[8]University of Grenoble Alpes, CNRS, INRAE, IRD, Grenoble INP, IGE, Grenoble 38000, France

Correspondence to: Marta Via (marta.viagonzalez@ung.si)

A) Supplementary figures

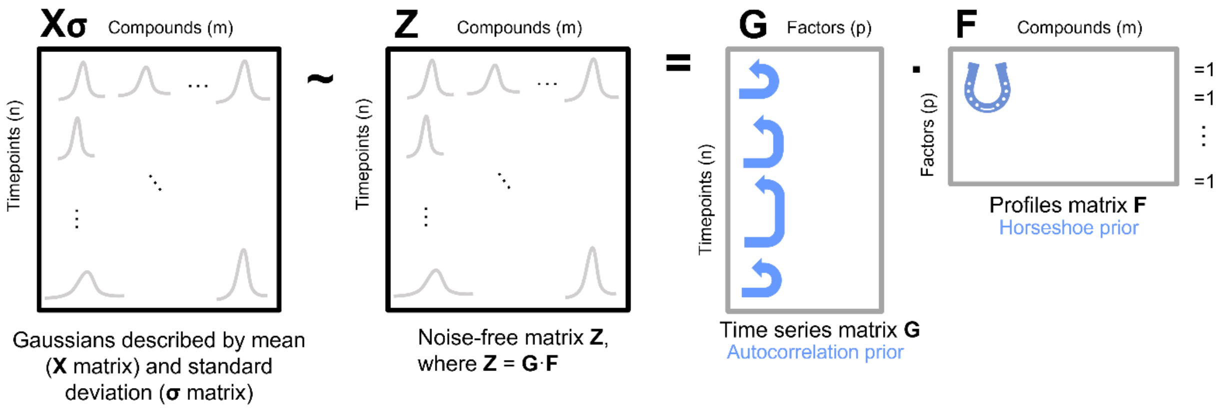


Figure S1. Bayesian matrix factorisation sketch with autocorrelation and horseshoe priors.

Table S1. Hamiltonian MonteCarlo sampling parameters for all the conducted experiments

Experiment	**Total number of samples**	**Number of warm-up samples**	**Number of chains**
Toy dataset	4000	2000	4
European synthetic datasets	12000	6000	4

Filters synthetic dataset	6000	3000	4
Filters real-world dataset	6000	3000	4

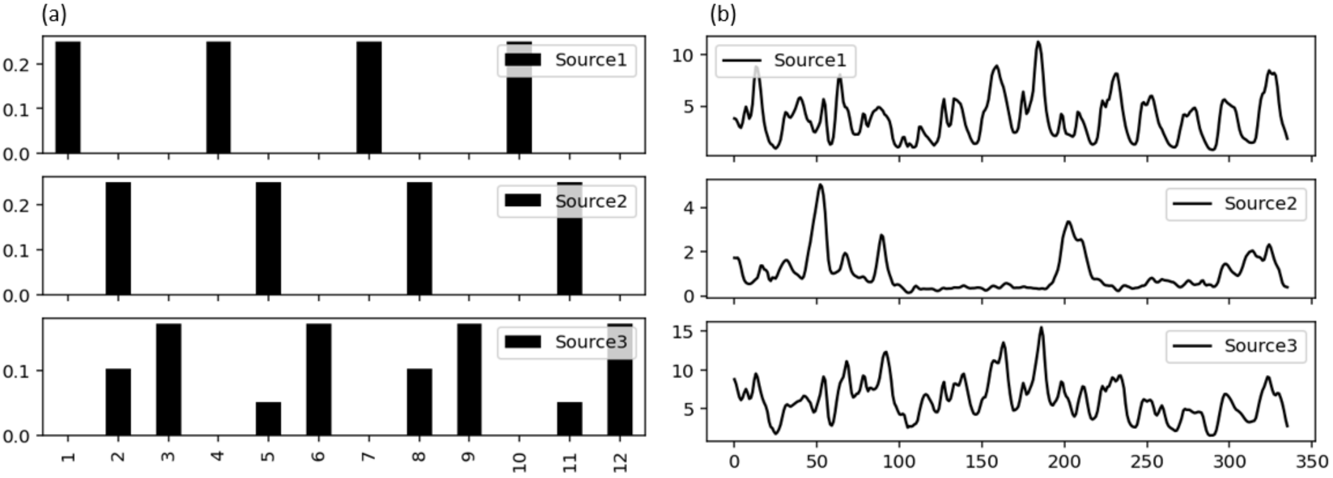


Figure S2. Toy dataset (a) Profiles. (b) Time series.

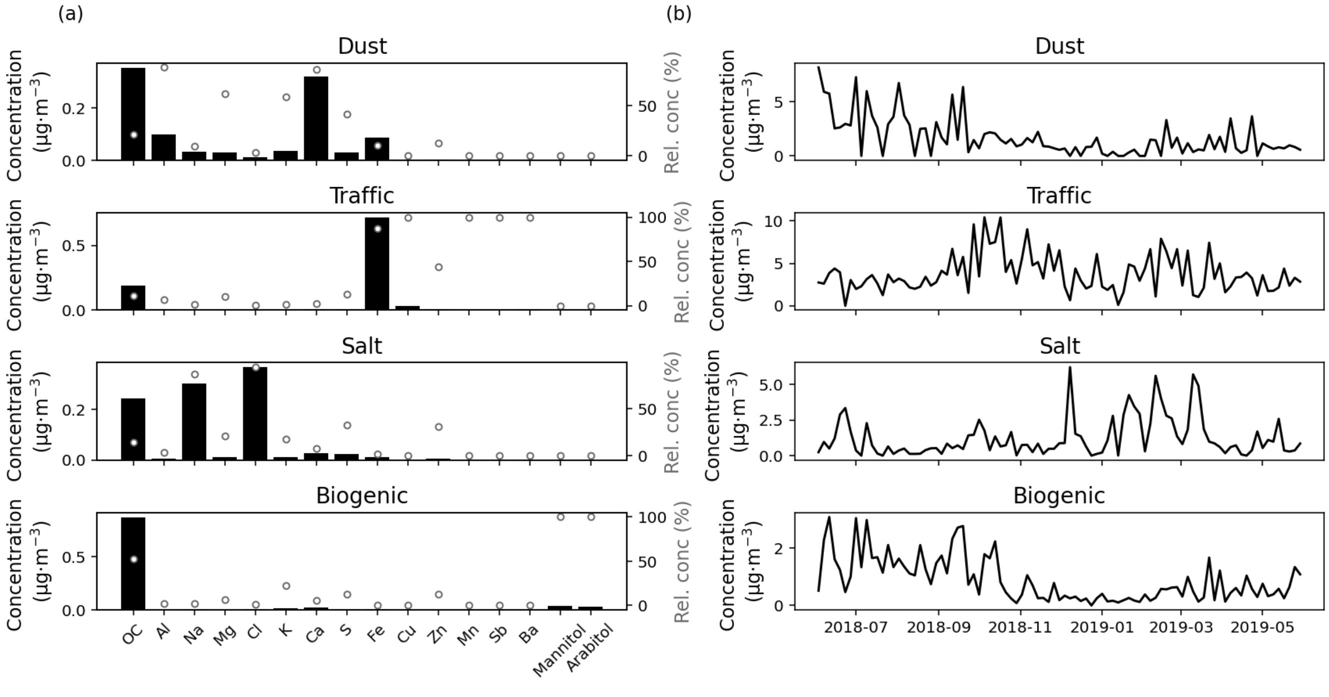


Figure S3. Synthetic offline dataset (a) Profiles. (b) Time series.

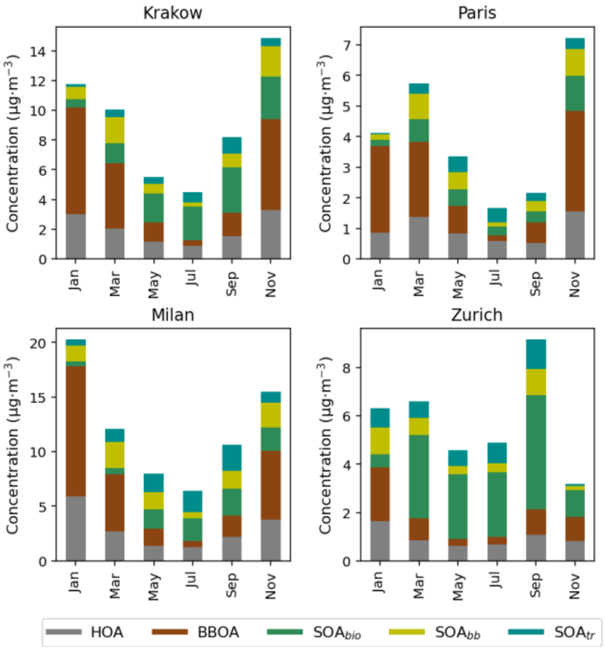


Figure S4. Generated synthetic datasets source mean concentrations for 4 European cities.

Table S2. Profiles employed for the 4-cities synthetic datasets.

City	**Source**	**Citation**
Krakow	HOA	Mohr et al. (2012)
	BBOA	Tobler et al. (2021)
Milan	HOA	Via et al. (2021)
	BBOA	Daellenbach et al. (2021)
Paris	HOA	Crippa et al. (2011)
	BBOA	Zhang et al. (2022)
Zurich	HOA	Elser et al. (2016)
	BBOA	Ulbrich et al. (2002), Ulbrich et al. (2022)

	SOA_{bio}	Daellenbach et al. (2017)
	SOA_{bb}	Ulbrich et al. (2002)
	SOA_{tr}	Sage et al. (2007)

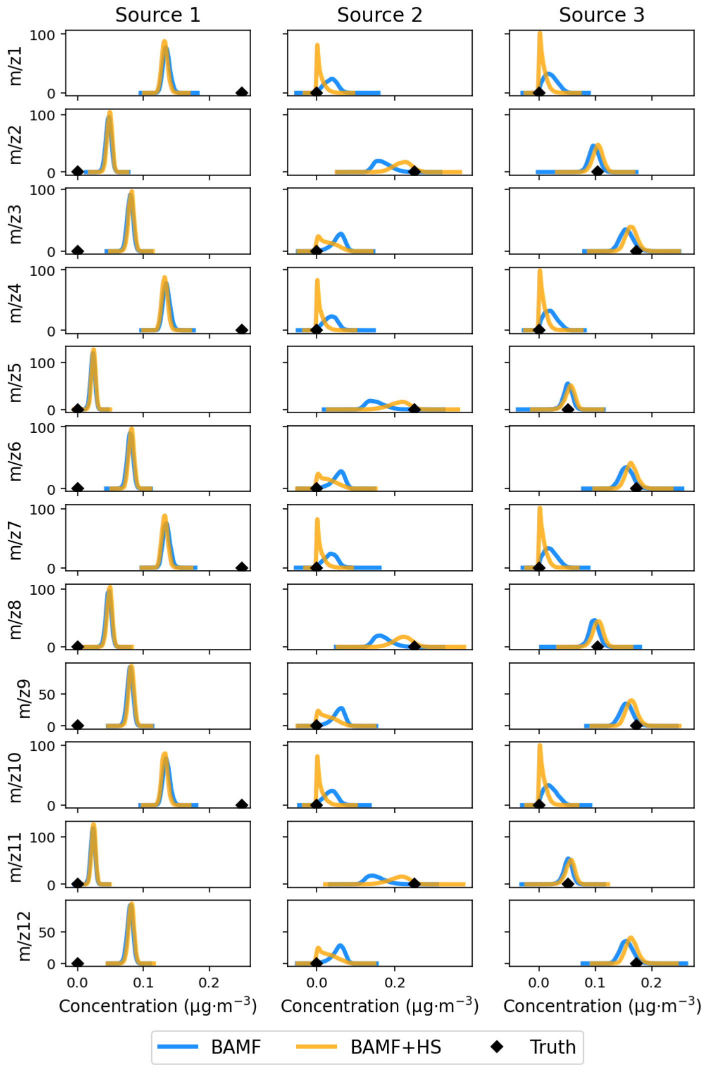
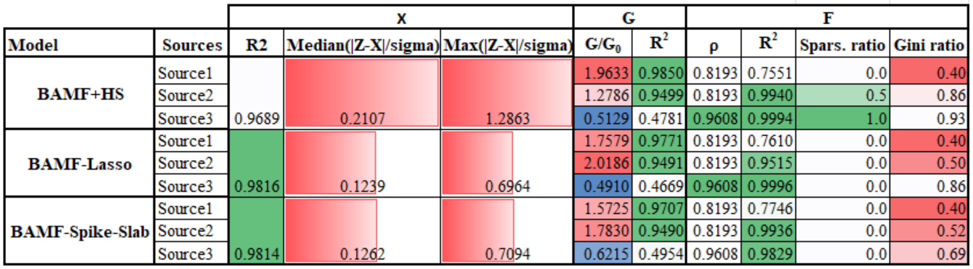


Figure S5. Toy dataset distribution of F components for BAMF, BAMF+HS.

Table S3. Statistics of the toy dataset reconstruction and comparison to truth for the different shrinkage alternatives.



		X			G		F			
Model	**Sources**	**R2**	**Median(|Z-X|/sigma)**	**Max(|Z-X|/sigma)**	**G/G_0**	**R^2**	**ρ**	**R^2**	**Spars. ratio**	**Gini ratio**
BAMF+HS	Source1	0.9689	0.2107	1.2863	1.9633	0.9850	0.8193	0.7551	0.0	0.40
	Source2				1.2786	0.9499	0.8193	0.9940	0.5	0.86
	Source3				0.5129	0.4781	0.9608	0.9994	1.0	0.93
BAMF-Lasso	Source1	0.9816	0.1239	0.6964	1.7579	0.9771	0.8193	0.7610	0.0	0.40
	Source2				2.0186	0.9491	0.8193	0.9515	0.0	0.50
	Source3				0.4910	0.4669	0.9608	0.9996	0.0	0.86
BAMF-Spike-Slab	Source1	0.9814	0.1262	0.7094	1.5725	0.9707	0.8193	0.7746	0.0	0.40
	Source2				1.7830	0.9490	0.8193	0.9936	0.0	0.52
	Source3				0.6215	0.4954	0.9608	0.9829	0.0	0.69

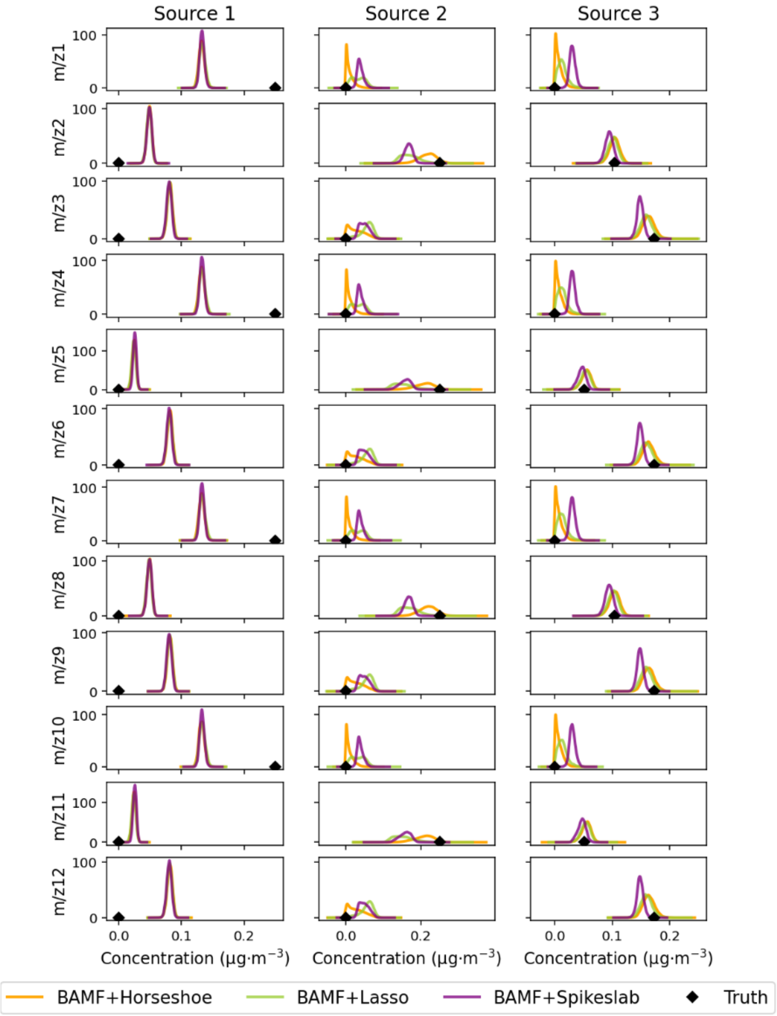


Figure S6. Comparison of toy dataset distribution of F components for BAMF+HS, BAMF+Lasso, BAMF+Spike-and-slab.

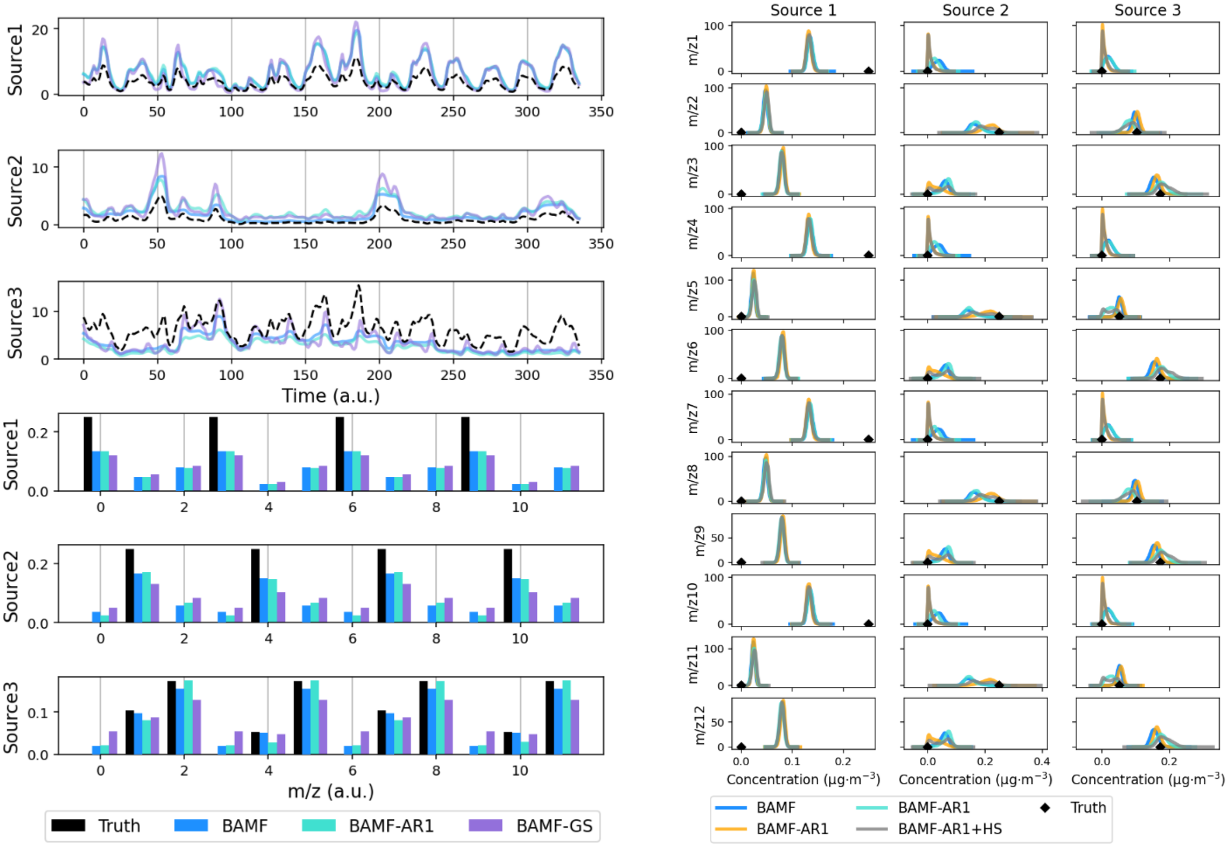


Figure S7. Left:Toy dataset time series (top) and profiles (bottom) results with BAMF, BAMF-AR1 and BAMF-GS. Right:Toy dataset distribution of F components for BAMF, BAMF+HS, BAMF-AR1, BAMF-AR1+HS.

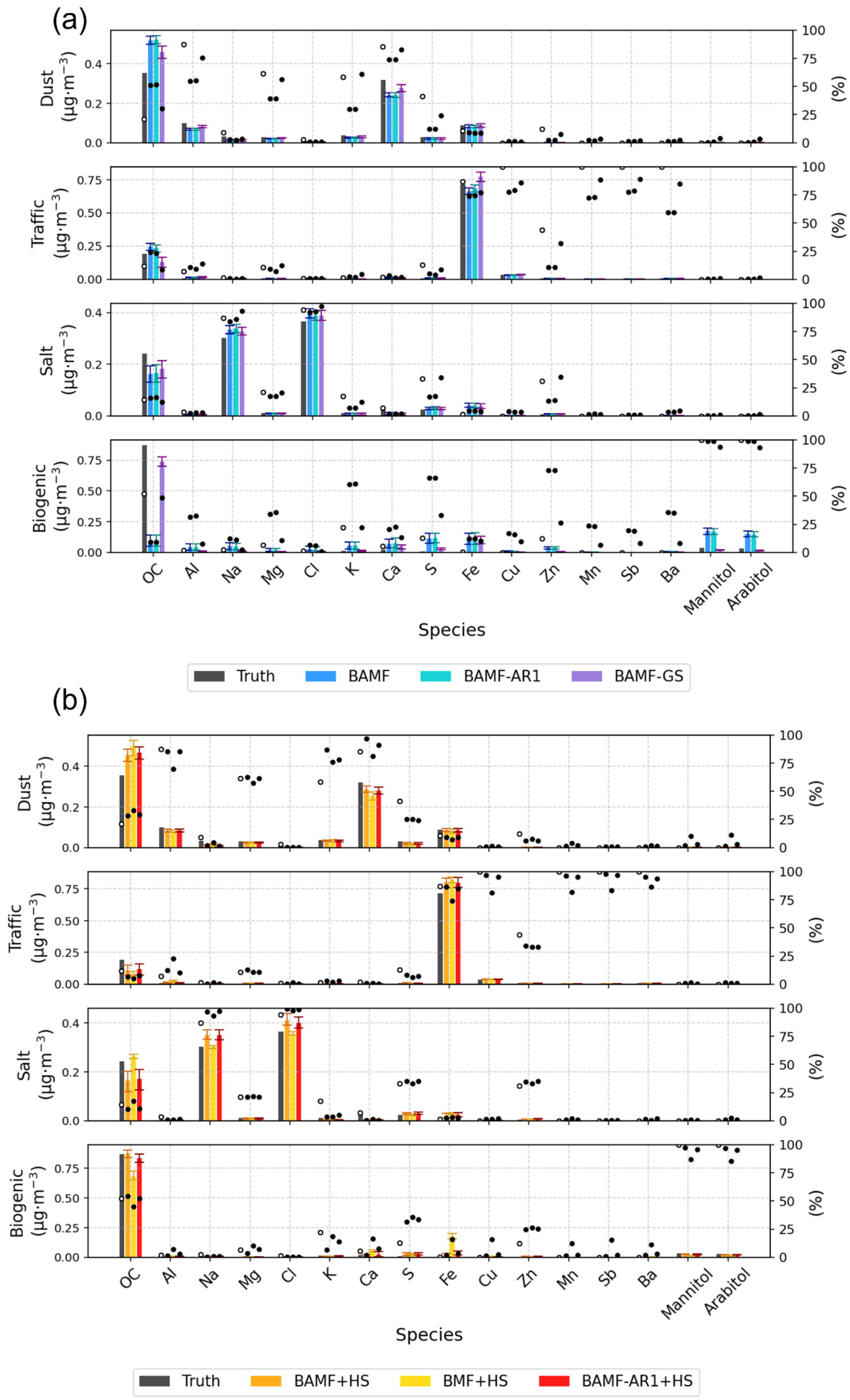


Figure S8. Synthetic offline dataset comparison between (a) BAMF, BAMF-AR1, BAMF-GS (b) BAMF+HS, BMF+HS, BAMF-AR1+HS.

Table S4. Purely-measurement-based offline synthetic dataset factorisation statistics. The last three rows correspond to the sum of the metrics for all all factors, except for the G/G0 metric column, which shows the sum of the factors G/G_0 deviation to 1.

Model	**Factor**	**F R^2**	**F ρ**	**F contr. R^2**	**F contr. ρ**	**F Gini**	**G R^2**	**G/G_0 (*)**
Dust	PMF	0.98	0.90	0.93	0.90	0.77	0.95	1.00
	BAMF	0.99	0.97	0.78	0.93	0.74	0.92	0.98
	BAMF+HS	0.96	0.96	0.96	0.86	0.77	0.93	1.02
Traffic	PMF	0.99	0.67	0.87	0.88	0.89	0.94	0.83
	BAMF	0.97	0.77	0.46	0.75	0.81	0.93	1.02
	BAMF+HS	0.96	0.85	0.85	0.94	0.88	0.94	0.75
Salt	PMF	0.90	0.60	0.95	0.63	0.83	0.77	0.92
	BAMF	0.90	0.92	0.71	0.63	0.63	0.30	0.11
	BAMF+HS	0.78	0.82	0.92	0.55	0.86	0.52	0.86
Biogenic	PMF	0.96	0.27	0.77	0.16	0.88	0.96	0.68
	BAMF	0.03	0.85	0.56	0.21	0.51	0.96	0.14
	BAMF+HS	0.999	0.63	0.85	0.41	0.89	0.96	0.65
$\sum_k$ (*)	PMF	3.85	2.44	3.52	2.57	3.37	3.64	**0.61**
	BAMF	2.89	3.51	2.51	2.52	2.69	3.11	1.79
	BAMF+HS	3.70	3.26	3.58	2.76	3.40	3.35	0.78

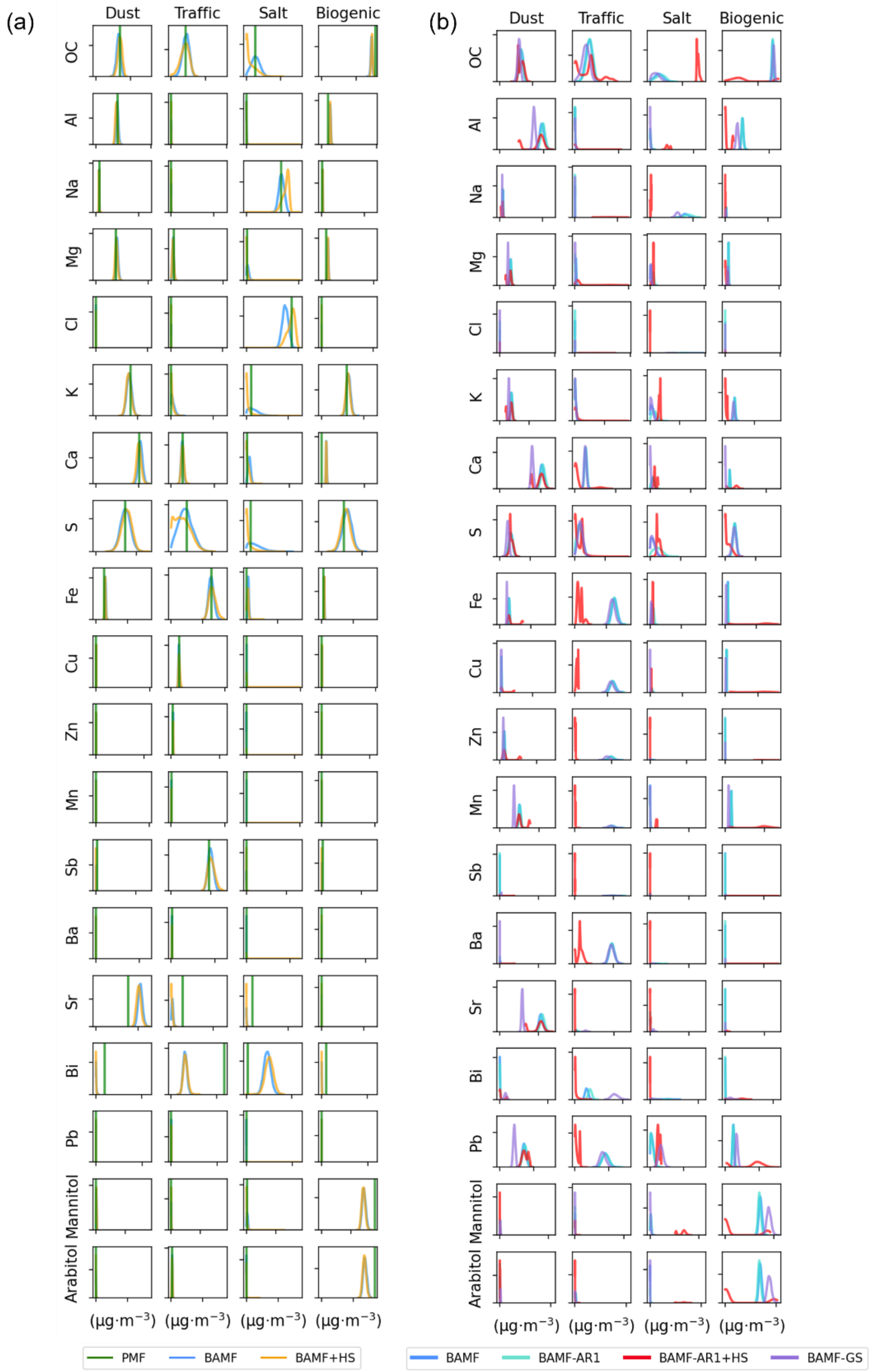

(a)
Dust
Traffic
Salt
Biogenic
OC
Al
Na
Mg
Cl
K
Ca
S
Fe
Cu
Zn
Mn
Sb
Ba
Sr
Bi
Pb
Mannitol
Arabitol
(μg·m^{-3})
PMF
BAMF
BAMF+HS
(b)
Dust
Traffic
Salt
Biogenic
OC
Al
Na
Mg
Cl
K
Ca
S
Fe
Cu
Zn
Mn
Sb
Ba
Sr
Bi
Pb
Mannitol
Arabitol
(μg·m^{-3})
BAMF
BAMF-AR1
BAMF-AR1+HS
BAMF-GS

Figure S9. Filters real-world filters dataset profiles components distributions for (a) PMF, BAMF, BAMF+HS; (b) BAMF, BAMF-AR1, BAMF-AR1+Hs, BAMF+GS.

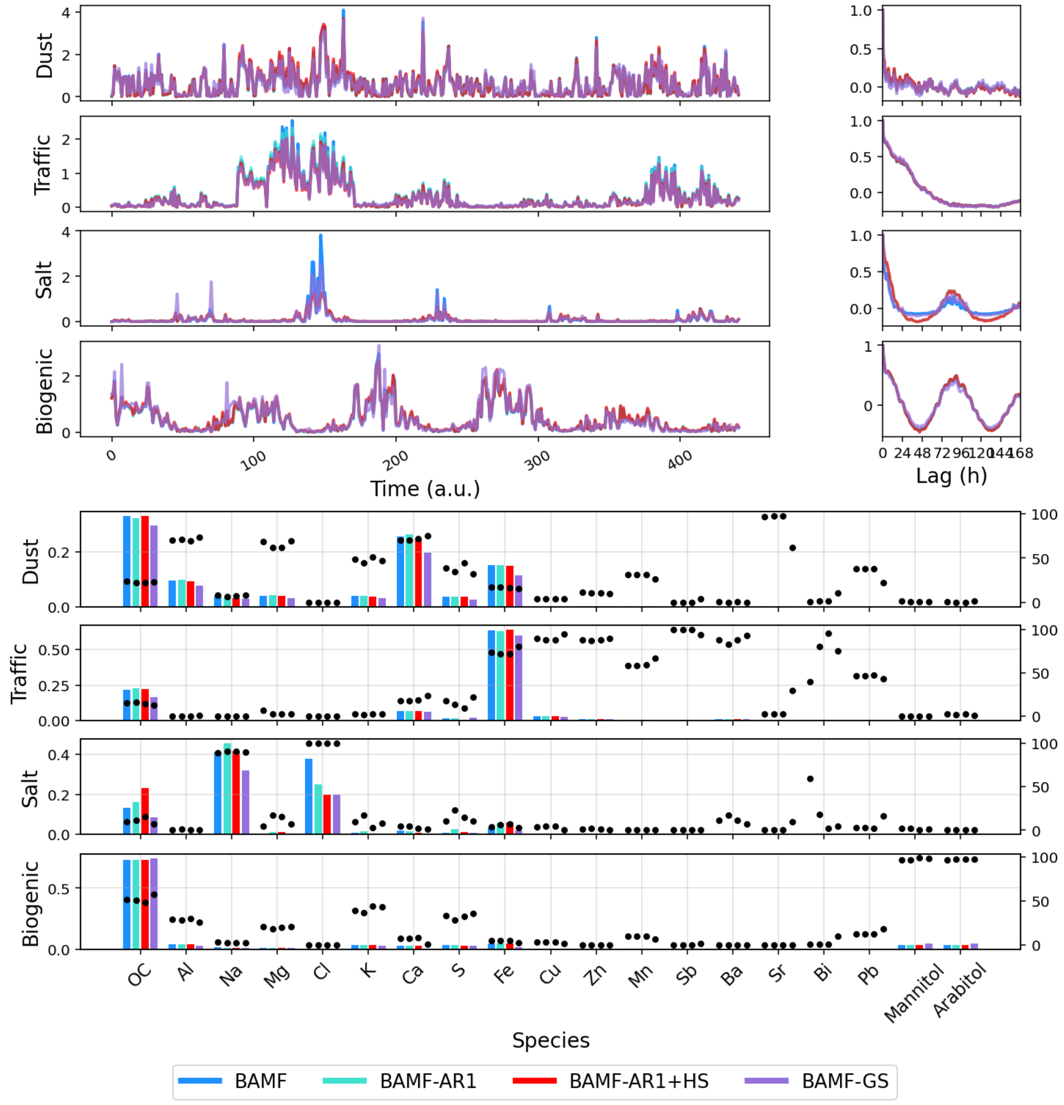


Figure S10. Comparison of BAMF, BAMF-AR1, BAMF-AR1+HS, BAMF-GS on the real-world filters dataset. Plots from left to right and from the top to the bottom are: Time series, autocorrelation, profiles.

Table S5. Online European synthetic datasets reconstruction and factorisation statistics. Statistics contain data from all sites, sources, and datasets.

Statistic	Model	Mean	Median	Min	Max
|X - Z|/σ	**PMF**	**4.24**	**4.34**	**1.45**	**7.17**
	BAMF	**0.64**	**0.66**	**0.04**	**2.04**
	BAMF+HS	**2.92**	**2.26**	**0.66**	**8.45**
G/G_0	**PMF**	**2.82**	**1.10**	**0.39**	**18.45**
	BAMF	**1.07**	**0.92**	**0.02**	**7.21**
	BAMF+HS	**2.00**	**1.11**	**0.19**	**46.86**
G R^2	**PMF**	**0.78**	**0.91**	**0.01**	**1.00**
	BAMF	**0.84**	**0.95**	**0.04**	**1.00**
	BAMF+HS	**0.83**	**0.92**	**0.05**	**1.00**
F ρ	**PMF**	**0.85**	**0.87**	**0.53**	**1.00**
	BAMF	**0.88**	**0.91**	**0.27**	**0.99**
	BAMF+HS	**0.87**	**0.90**	**0.53**	**1.00**
F R^2	**PMF**	**0.84**	**0.92**	**0.04**	**1.00**
	BAMF	**0.90**	**0.96**	**0.02**	**1.00**
	BAMF+HS	**0.88**	**0.95**	**0.16**	**1.00**
F sparsity	**PMF**	**2.84**	**2.69**	**1.69**	**5.14**
	BAMF	**2.61**	**2.45**	**1.53**	**4.53**
	BAMF+HS	**2.63**	**2.45**	**1.65**	**4.53**

F Gini	**PMF**	**0.59**	**0.59**	**0.45**	**0.72**
	BAMF	**0.57**	**0.56**	**0.05**	**0.75**
	BAMF+HS	**0.57**	**0.58**	**0.41**	**0.83**
F Gini ratio	**PMF**	**0.99**	**0.96**	**0.75**	**1.46**
	BAMF	**0.97**	**0.99**	**0.09**	**1.15**
	BAMF+HS	**0.98**	**0.97**	**0.72**	**1.40**

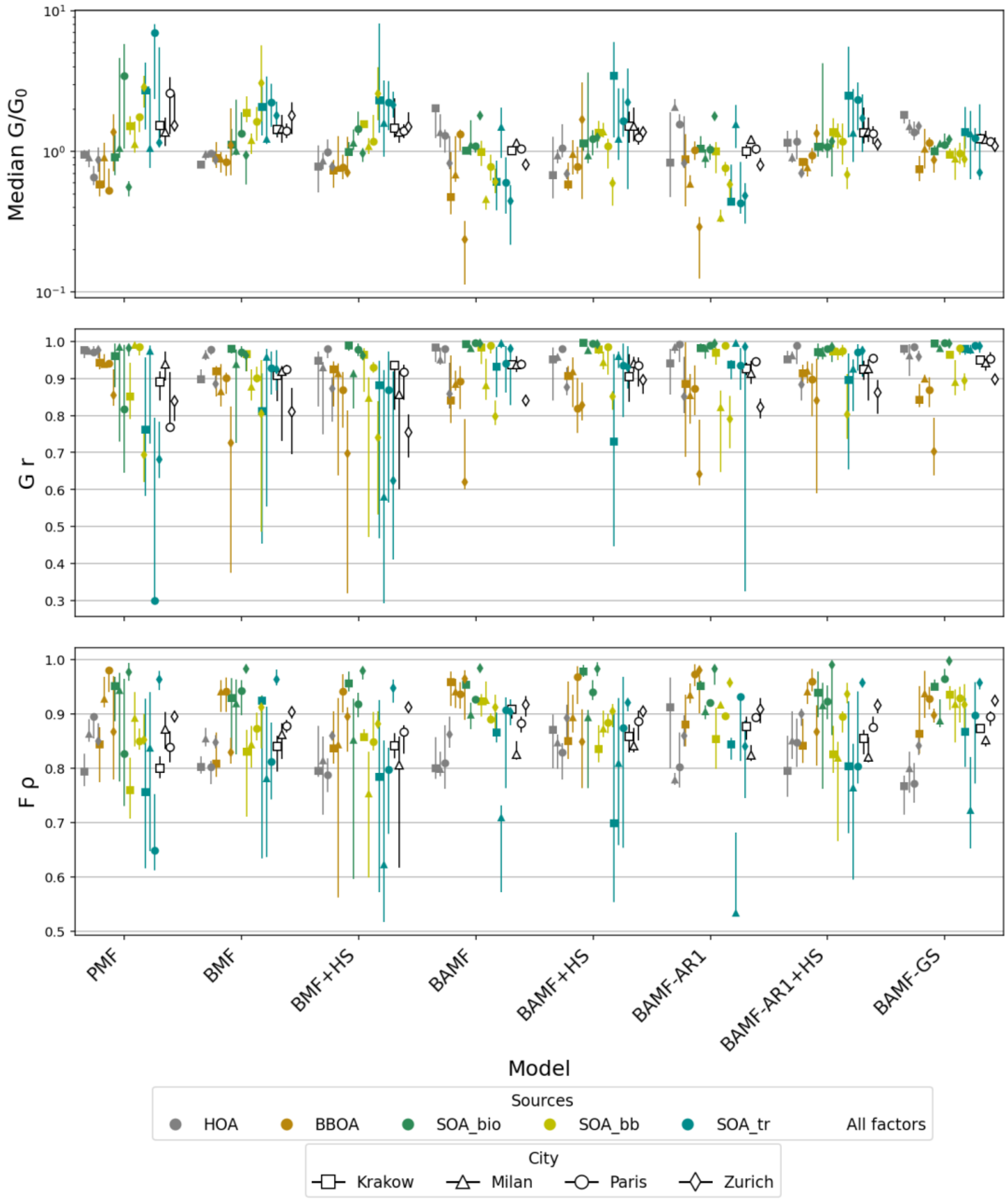
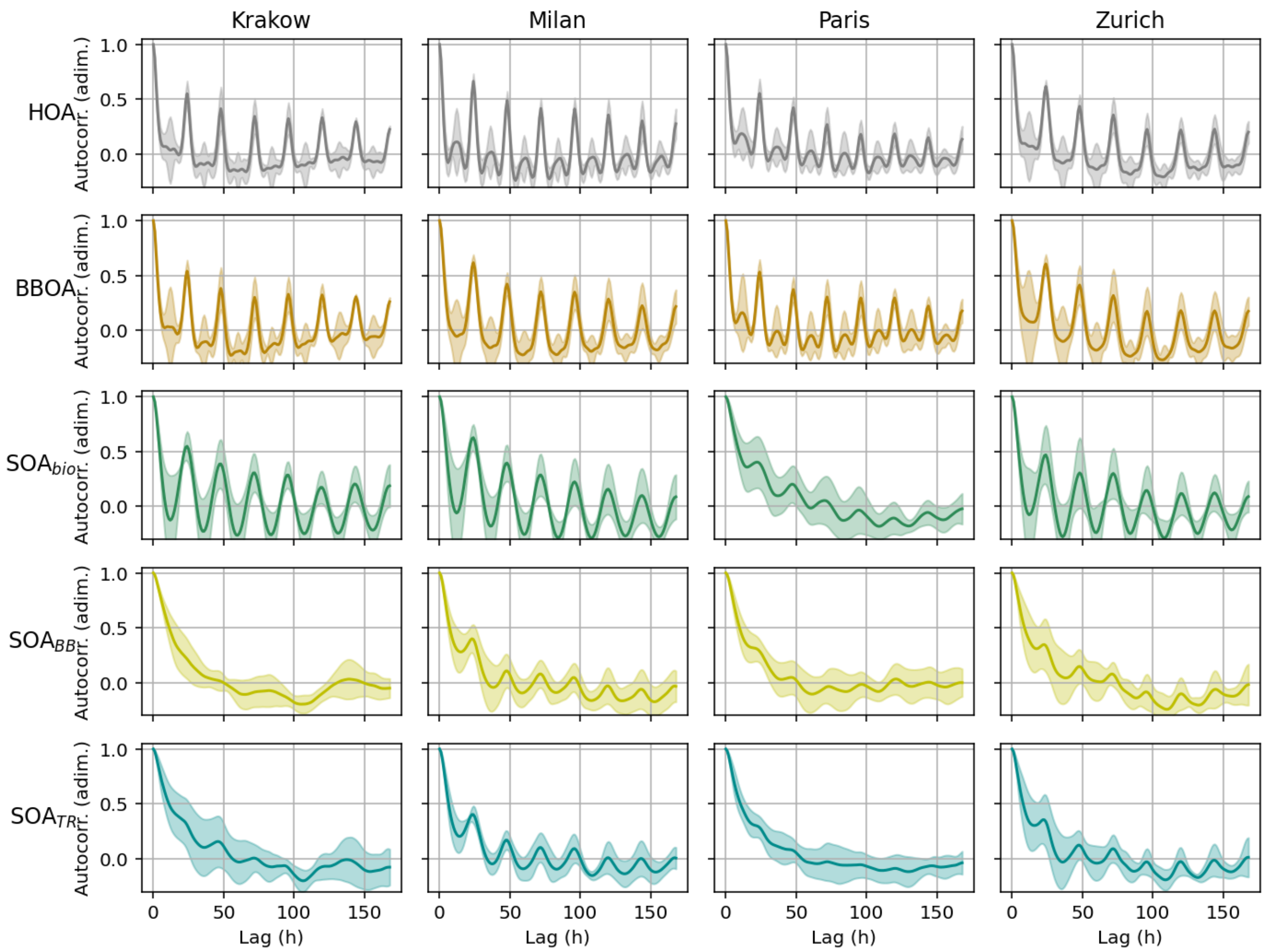


Figure S11. Summary metrics for all synthetic datasets.

69



70

71 **Figure S12. Autocorrelation means (solid lines) and standard deviations (shaded areas) of the 6 datasets per city and source.**

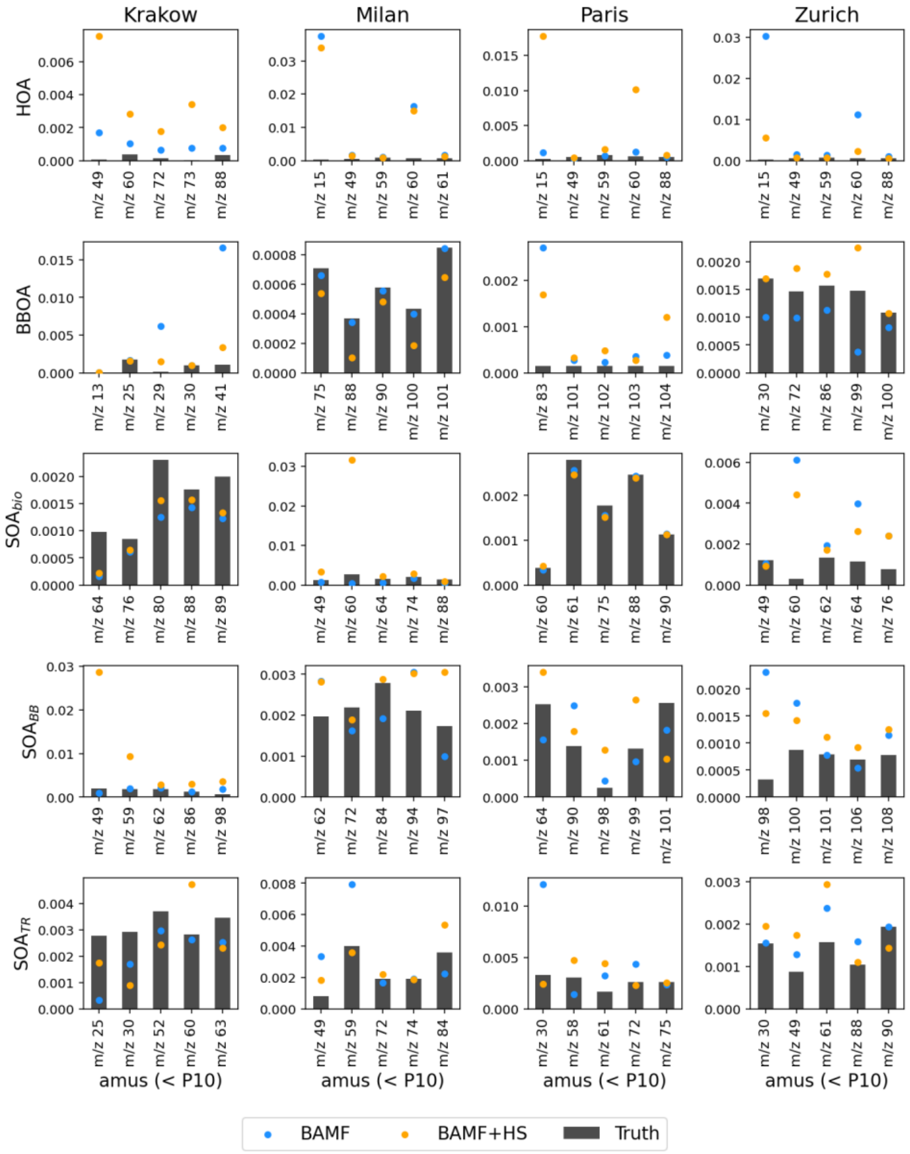


Figure S13. Sparsity introduction results for BAMF and BAMF+HS models in the European synthetic datasets for the 5 less massive m/zs, below the percentile 10. The bars show the truth concentration for the given m/z, and BAMF, BAMF+HS results are shown in markers.

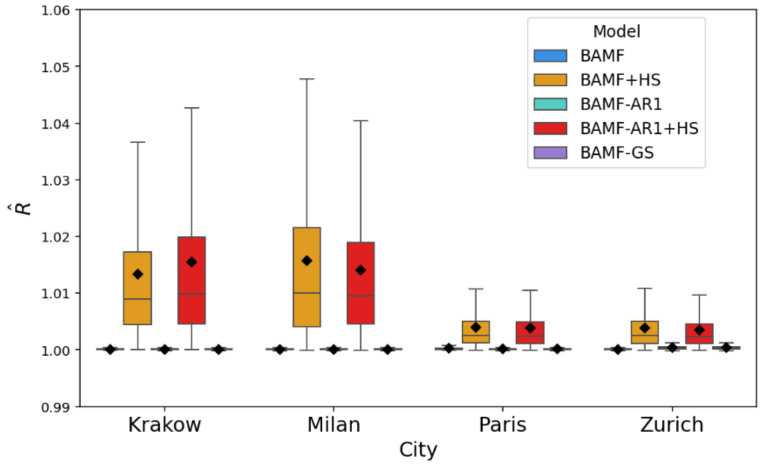


Figure S14. $\widehat{R}$ for the different cities and models. Model boxplots are in the same order as the legend.

B) Supplementary model formulation

Lasso

The Lasso (Least Absolute Shrinkage and Selection Operator, Rasmussen et al. 2012) is a regularization method that adds an L1 penalty to linear regression, shrinking some coefficients exactly to zero. The L1 penalty is a way to discourage large coefficients in a regression model by adding up the absolute values of all the model's coefficients and including that total in the cost the model tries to minimize. This encourages sparsity in the model, making it useful for variable selection. However, it can over-shrink large coefficients and it treats all coefficients equally, regardless of their signal strength, hence it tends to over-shrink even large signals.

Spike-and-slab

The spike-and-slab prior (Andersen et al. 2014) is a Bayesian approach to variable selection that models each coefficient as coming from a mixture of two distributions: a spike at zero (forcing sparsity) and a slab (a wider distribution allowing nonzero values). It provides strong interpretability and explicit variable inclusion, but is computationally intensive due to its discrete model space. This working scheme provides a binary shrinkage behaviour since the signals are considered either zero or slabs, which enforces 0-like signals, however, the selection of non-zero elements can be too harsh.

Horseshoe

The regularized horseshoe prior is an extension of the standard horseshoe prior designed to improve robustness and regularization in sparse Bayesian models. Like the original horseshoe, it combines global shrinkage (for overall sparsity) and local shrinkage (for individual coefficients), using heavy tails to allow large signals while aggressively shrinking noise. What sets the regularized version apart is the addition of a slab component that limits the influence of extremely large coefficients. This slab acts like a soft ceiling, preventing the model from over-trusting outlier variables while maintaining flexibility. This makes the prior more stable in practice, especially in finite data settings or when outliers are present. In comparison to the two above priors, the regularized horseshoe provides a middle ground: it allows for adaptive shrinkage, handles strong signals well, and includes a slab to avoid the horseshoe's instability. It's more robust and scalable than spike-and-slab and more flexible and statistically principled than Lasso.

C) Supplementary model evaluation

C.I. Chemically sparse toy dataset

Some alternative autocorrelation formulations to the regular BAMF autocorrelation priors were tried for the toy dataset. BMF, BMF+HS, and BMF-GS – analogs to BAMF, BAMF+HS, BAMF-GS, respectively, but without the autocorrelation term – provided unstable results, with the highest |**Z**-**X**|/**σ** median and maxima and the poorest **X** vs. **Z** correlation coefficients, hence, these were discarded for this discussion. Figure S7 shows the time series, profiles and **F** components distributions for BAMF, BAMF-AR1, BAMG+GS, and BAMF+HS. Both regarding time series and profiles, BAMF seems to be resembling the truth equally or better than BAMF-AR1 and BAMF-GS with the only exception of the time series of Source 3, which is visually better captured by BAMF-GS. However, BAMF-GS profiles are further from the truth and m/z mass closure for this model is much worse than regular BAMF, as also seen by poorer Spearman and Pearson correlation coefficients. Profiles as captured by BAMF-AR1 are not consistently better than those from BAMF either. BAMF appears as the most balanced and accurate model of the three for the presented toy dataset. Regarding **F** components distributions, BAMF and BAMF-AR1 and BAMF+HS and BAMF-AR1+HS BAMF-AR1, BAMF-AR1+HS are compared in the right panel of Figure S7. BAMF and BAMF-AR1 show very similar distributions, with the slight differences mentioned before. BAMF+HS and BAMF-AR1+HS present some differences, and even if the shrinkage is also present in BAMF-AR1+HS (although with weaker Gini than

BAMF+HS), some components present distributions with a more pronounced multimodal behaviour than the BAMF+HS. Therefore, the BAMF-AR1+HS is shown to be less precise and accurate than BAMF+HS. The BAMF-GS+HS was not tried out since the high concentrations of the components of the non-normalised F matrix on this model hinder their shrinkage to zero.

C.II. Chemically sparse offline synthetic dataset

Figure S8 shows the profiles of the **F** matrix for BAMF, BAMF-AR1, BAMF-GS (a) and BAMF+HS, BMF+HS, BAMF-AR1+HS (b) models. BAMF, BAMF-AR1, and BAMF-GS were run, initialising **F** as a normal distribution to make the sampling more sturdy to avoid initialization failure. The outcomes of these models show general good agreement with the truth, with BAMF+HS being the most accurate model followed by BAMF-AR1+HS. The sum across factors of profile Pearson R^2 with truth for BAMF+HS, BMF+HS, BAMF-AR1+HS, and BAMF-GS were 3.90, 3.80, and 3.90, proving the beneficial effect of the horseshoe prior in the description of sparse profiles. The BAMF-AR1+HS model is performing slightly worse than the BAMF+HS in time series (G/G_0 deviation sum of 0.85, 0.91, respectively). Regarding the non-horseshoed models, the sum of the correlation with truth **G** for all factors was very similar for BAMF, BAMF-AR1, GS (3.93, 3.92, 3.93, respectively) but the Spearman correlation coefficient with truth F was better for BAMF (3.75, 3.71, 3.67, respectively), highlighting a slight better performance for BAMF. Hence, with all, the BAMF+HS seems the most accurate model, benefitting both from its autocorrelation and sparsity properties.

C.III. Chemically sparse offline real-world dataset

Figure S10 shows the performance of the other models discussed in the Toy dataset and European city datasets sections, BAMF-AR1, BAMF-AR1+HS, BAMF-GS, in comparison to BAMF, used as the base case. All models employed seem to agree, providing overlapping time series and autocorrelations, and similar profiles, with the exception of BAMF-AR1+HS, which is slightly differentiated from the others. However, it does not provide sparsified profiles in the species in which the shrinkage effect was found for BAMF+HS or elsewhere. This is more evidently depicted in Figure S9 (b), in which the distributions of **F** components are shown. The BAMF-AR1+HS does not present the expected sparsifying shape except for Ca, K, and some elements in the traffic and salt profiles. In general, the BAMF-AR1+HS distributions are more multimodal, reflecting higher divergence across HMC chains which makes results much less sturdy. Regarding the rest of models, the distributions are very similar ensuring the robustness of all three and stability of the solution. Therefore, all Bayesian models and PMF point to a similar, robust solution for the filters dataset. The horseshoe prior addition to BAMF, though, provides here a useful sparsity introduction in the current dataset which helps purify the profiles from unwanted profile entanglement. However, the implementation of this prior in the BAMF-AR1 model detriments the solution due to HMC chain divergence.