



Improving imputation of missing PM_{2.5} speciation data using PMF-informed source-receptor relationships

Wubin Zhu¹, Mingjie Xie², Qili Dai^{1,3}, Xiaohui Bi¹, Yufen Zhang¹, and Yinchang Feng¹

¹State Environmental Protection Key Laboratory of Urban Ambient Air Particulate Matter Pollution Prevention and Control, College of Environmental Science and Engineering, Nankai University, Tianjin 300350, China

²Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Atmospheric Environment Monitoring and Pollution Control, School of Environmental Science and Engineering, Nanjing University of Information Science & Technology, 219 Ningliu Road, Nanjing, 210044, China

³Tianjin Key Laboratory of Software Experience and Human Computer Interaction, Tianjin 300457, China

Correspondence: Qili Dai (daiql@nankai.edu.cn)

Received: 27 January 2026 – Discussion started: 5 March 2026

Revised: 30 May 2026 – Accepted: 10 June 2026 – Published: 26 June 2026

Abstract. Missing values are ubiquitous in atmospheric monitoring due to instrument drift, calibration cycles, operational interruptions, and other random malfunctions. Such gaps can undermine the reliability of subsequent analyses and introduce systematic biases. Conventional imputation methods, such as geometric mean substitution, K-nearest neighbor (KNN), Bayesian principal component analysis (BPCA), and deep learning models often rely primarily on statistical correlations, may require auxiliary inputs, and offer limited physical interpretability. To address this issue, we propose a novel source-receptor-informed Positive Matrix Factorization Reconstruction (PMFr) method that leverages PMF-derived source-receptor relationships, rather than purely statistical interpolation, to impute missing PM_{2.5} speciation data without requiring auxiliary data. Benchmarking on a two-month dataset against commonly used imputation techniques, including KNN, BPCA, and a deep learning predictive model, demonstrates that PMFr achieves superior accuracy and robustness across real-world missing scenarios, with a mean coefficient of determination (R^2) of 0.81, index of agreement (IoA) of 0.92, and mean absolute percentage error (MAPE) of 22.8 %, reducing MAPE by 25.5 %–29.1 %, particularly for key PM_{2.5} species. Further PMF-based validation shows that PMFr better preserves source-profile composition and source-contribution temporal features, indicating that the completed dataset retains more physically meaningful source information and is more suitable for source apportionment. These results highlight PMFr as a robust and

physically interpretable approach for reconstructing reliable PM_{2.5} speciation data.

1 Introduction

Ambient fine particulate matter (PM_{2.5}) remains a pressing global environmental challenge due to its well-documented impacts on climate forcing, atmospheric visibility degradation, and adverse health outcomes (Liu and Matsui, 2021; Peng et al., 2023; Hao et al., 2023; Kim et al., 2025b). These effects are governed by the chemically diverse nature of PM_{2.5}, which comprises inorganic ions, carbonaceous materials, trace metals, and other species. Comprehensive PM_{2.5} speciation measurements are therefore fundamental for tracking source contributions, elucidating atmospheric processes, and evaluating their diverse impacts. However, missing data are ubiquitous in both routine monitoring networks and intensive field campaigns due to instrument drift, calibration cycles, operational interruptions, and other random malfunctions (Yu et al., 2017). Such gaps can undermine the reliability of subsequent analyses and introduce systematic biases. Consequently, accurate and robust imputation of missing values is essential, as inappropriate handling of missing data can lead to distorted interpretations and erroneous scientific conclusions.

A wide range of methods have been developed to address missing values in PM_{2.5} chemical component datasets, gen-

erally falling into listwise deletion, simple substitutions, and advanced statistical models (Alwateer et al., 2024). Basic approaches such as listwise deletion and mean or median substitution, although recommended in the U.S. EPA's guidelines for their simplicity (Norris et al., 2014), often compromise data quality: listwise deletion discards samples containing any missing species and substantially reduces statistical power, whereas median or mean substitution introduces bias that becomes more pronounced as data variability increases (Emmanuel et al., 2021; Polissar et al., 1998; Khan and Hoque, 2020). Linear interpolation is also frequently applied because of its ease of implementation, yet its performance is highly sensitive to the temporal pattern and extent of missingness (Samal et al., 2021; Junninen et al., 2004). To better capture inter-species correlations and nonlinear dependencies, more advanced techniques, including KNN, BPCA, and deep learning models such as deep belief networks (DBN), have been explored and often outperform simpler methods (Lee et al., 2023; Lai and Kuok, 2019; Zaini et al., 2022; Xie, 2017; Shen et al., 2018). Nevertheless, these statistical and machine-learning approaches typically rely on mathematical interpolation, may require auxiliary inputs such as meteorological variables or satellite-derived AOD, and offer limited physical interpretability (van Donkelaar et al., 2019; Lee et al., 2024; Hu et al., 2014; Kim et al., 2024, 2025a). As a result, accurately reconstructing missing PM_{2.5} chemical species remains a methodological challenge.

To address these limitations, we develop a physically interpretable imputation method grounded in air pollution source-receptor principles to reconstruct missing PM_{2.5} chemical species. Source contributions and profiles are first resolved from pre-existing speciation data using Positive Matrix Factorization (PMF), which decomposes the dataset into a source chemical profile matrix and its corresponding contribution matrix. Under the commonly assumed temporal stability of source chemical compositions, the resolved profiles are then used to reproduce new PM_{2.5} speciation datasets containing missing species by multiplying the estimated source-specific PM_{2.5} mass by the resolved source profiles, enabling estimation of missing values based on physically meaningful source signatures. This approach ensures that reconstructed concentrations align with both chemical structure and emission characteristics rather than relying solely on mathematical interpolation. To evaluate performance, we generated artificial missing data in complete-speciation datasets; the proposed method was then compared against geometric mean substitution, linear interpolation, K-nearest neighbors (KNN), deep belief networks (DBN), and Bayesian principal component analysis (BPCA). The datasets completed by each imputation method were subsequently used as PMF inputs to assess how different imputation strategies influence downstream source apportionment results. This study highlights the potential of a source-informed strategy for robust and interpretable imputation, as well as for generating completed datasets suitable for subse-

quent source apportionment and chemical-constrained health and climate impact analyses.

2 Material and Methods

2.1 Sample Collection and Data Processing

Hourly PM_{2.5} speciation data were collected on the rooftop of the Nanjing Environmental Protection Building (NEPB, 118.75° E, 32.06° N). Water-soluble inorganic ions including NH₄⁺, SO₄²⁻, NO₃⁻, Cl⁻, Ca²⁺, Mg²⁺, K⁺, and Na⁺ were determined by MARGA (ADI 2080; Applikon Analytical B.V., the Netherlands). Hourly OC and EC concentrations were measured using a semi-continuous OC/EC analyzer (RT-4, Sunset laboratory Inc., USA) with the NIOSH method 5040 (Birch, 2002). An Xact 625 ambient metals monitor (Cooper Environmental, United States) was configured to quantify twenty-three elements (K, Fe, Zn, Ca, Si, Mn, Pb, Cu, Ti, As, V, Ba, Cr, Se, Ag, Cd, Ni, Au, Co, Sn, Sb, Tl, and Hg). Detailed information on the monitoring site, instrument setup and maintenance, and chemical analysis were reported by studies (Yu et al., 2019, 2020; Xie et al., 2022).

The dataset used in this study comprises inorganic ions (NH₄⁺, SO₄²⁻, and NO₃⁻), trace elements (K, Fe, Zn, Ca, Si, Mn, Pb, Cu, Ti, As, V, Ba, Cr, and Se), and carbonaceous materials (OC and EC), which were measured from 1 October 2017, 01:00 am local time (LT, GMT+8) to 30 November 2017, 11:00 pm LT. The summary for the missing of raw data can be seen in Table S1 in the Supplement.

2.2 Missing Data Generation

Four factors are considered to affect the performance of imputation methods: the missing-data generation mechanism, the proportion of missing data, the gap pattern of missing data (Junger and Ponce de Leon, 2015), and whether multiple species are missing simultaneously (MCMS) or independently (MCMI). Specifically, MCMS refers to the simultaneous absence of multiple species at a single timestamp, while MCMI denotes missing values occurring independently at distinct timestamps.

The mechanisms of missing values are typically classified into three categories (Little and Rubin, 2019): (i) missing completely at random (MCAR), where missing values are generated independently, namely independent of both observed and unobserved values. (ii) missing at random (MAR), where missingness is related to observed data, and (iii) missing not at random (MNAR), where missing values are related to unobserved data such as values below the detection limit (BDL). Analysis of the NEPB dataset shows no systematic association between missing occurrences and pollutant concentration levels or temporal patterns, indicating that the missingness does not follow the MNAR mechanism (García-Laencina et al., 2010). Consequently, missing data

were generated randomly to ensure that the artificial missingness remained independent of pollutant concentrations.

The proportion of missing data is a critical factor affecting imputation performance. In this study, missingness rates of 10 %, 15 %, and 20 % were imposed, matching the observed range of 10 %–20 % in the monitored dataset.

Gap pattern refers to the proportion of different gap lengths within the total missing data. Based on the summary of the missing data, the physical meaning and prior research (Plaia and Bondi, 2006; Betancourt et al., 2023; Jing et al., 2022; Richardson and Hollinger, 2007), gap lengths (l) were categorized into three types: (i) short gaps, with l from 1 to 6; (ii) medium gaps, with lengths greater than 6 but less than 23; and (iii) large gaps, l ranging from 23 to 115 consecutive values (1 to 5 d), which represents the longest gap observed in the raw dataset (Table S2).

In summary, missing data were generated according to the scenarios listed in Table 1. These scenarios include: (i) random single-species missing that create short gaps in individual species (Case 1); (ii) instrument-failure-induced missing that produce medium gaps across all species measured by a given instrument, affecting ionic (Case 2) and carbonaceous (Case 3) monitors; and (iii) station-wide instrument malfunctions that result in large gaps spanning multiple species. These scenarios include malfunction of the ionic monitoring instrument (Case 4) and elemental monitoring instrument (Case 5), concurrent malfunction of two instruments (Cases 6–8), and malfunction of all monitoring instruments (Case 9). Potassium (K) was treated as missing in multi-instrument malfunction scenarios (Cases 6, 7 and 9) due to its strong correlations with both ionic and carbonaceous species (Fig. S1). The performance of the imputation methods was evaluated using the coefficient of determination (R^2), the mean absolute percentage error (MAPE), and the index of agreement (IoA) (Sect. S1.2).

2.3 Source-Receptor Informed Positive Matrix Factorization Reconstruction (PMFr) and Validation

A source-tracer for imputation, hereafter referred to as a tracer, is defined as a key species that distinguishes a specific factor (source) from others and reflects how that factor (source) influences the receptor over time. Co-tracers refer to co-varying tracers of the same factor, collectively characterizing the temporal behavior of the corresponding source. As illustrated in Fig. 1, PMF is first applied to resolve factor profiles and their contributions, providing source-receptor relationships constrained by expert knowledge, given that pollution sources imprint distinct temporal patterns on the receptor site. Details of the usage of PMF for SA can be found in the literature, and the uncertainty settings are provided in Sect. S1.3 (Hopke, 2016; Paatero, 1999; Paatero and Hopke, 2003). Based on the SA results with selected source profiles, species requiring imputation are classified as tracers

or non-tracers through a knowledge-driven step (Bi et al., 2019). When imputing tracers, the availability of co-tracers should be checked at each timestamp before reconstruction, because the source contribution vector (\mathbf{g}) needs to be constrained by source-specific tracer information. If all tracers associated with a specific factor are simultaneously missing, the corresponding \mathbf{g} vector is less directly constrained by observed species; in such cases, these missing tracer values are first imputed using another imputation method, with KNN recommended for its simplicity, efficiency, and ability to provide a reasonable estimate of temporal variation. The corresponding uncertainty is set to 10 % of the imputed concentration. For missing tracers with available co-tracers, as well as for non-tracers, missing values are replaced by the geometric mean. The uncertainty calculation is further discussed in Sect. S1.4. The pre-imputed dataset and its associated uncertainty matrix are then fed into the PMF analysis for reconstruction. The PMF run decomposes the dataset into factor profiles (\mathbf{F}) and source contributions (\mathbf{G}), and data reconstruction is achieved by multiplying the \mathbf{G} and \mathbf{F} matrices. Rather than relying directly on covariance in the high-dimensional chemical dataset, PMFr reconstructs missing values within this low-entropy source structure represented by PMF-resolved source profiles and temporal contributions.

The performance of PMFr was evaluated using two complementary validation endpoints: direct reconstruction accuracy and physical source-feature preservation. The reconstructed concentrations were directly compared with observed values and benchmarked against baseline methods, including LI, KNN, DBN, BPCA, and geometric mean imputation (Mean), using R^2 , IoA, and MAPE. The U.S. EPA PMF 5.0 User Guide recommends handling missing values by replacing them with the species median and assigning a high uncertainty to downweight these substituted values. Here, missing values were replaced by the species-specific geometric mean, following the same constant-substitution and downweighting principle. Because the geometric mean is also a robust central value for skewed data and was adopted in the previous PMF analysis using the same hourly PM_{2.5} speciation dataset (Xie et al., 2022), it was used here as a representative conventional PMF missing-value treatment for comparison with PMFr. Physical source-feature preservation was further assessed by comparing the PMF-resolved source profiles and corresponding source contributions obtained from different imputed datasets with those derived from the original complete dataset.

3 Results and Discussion

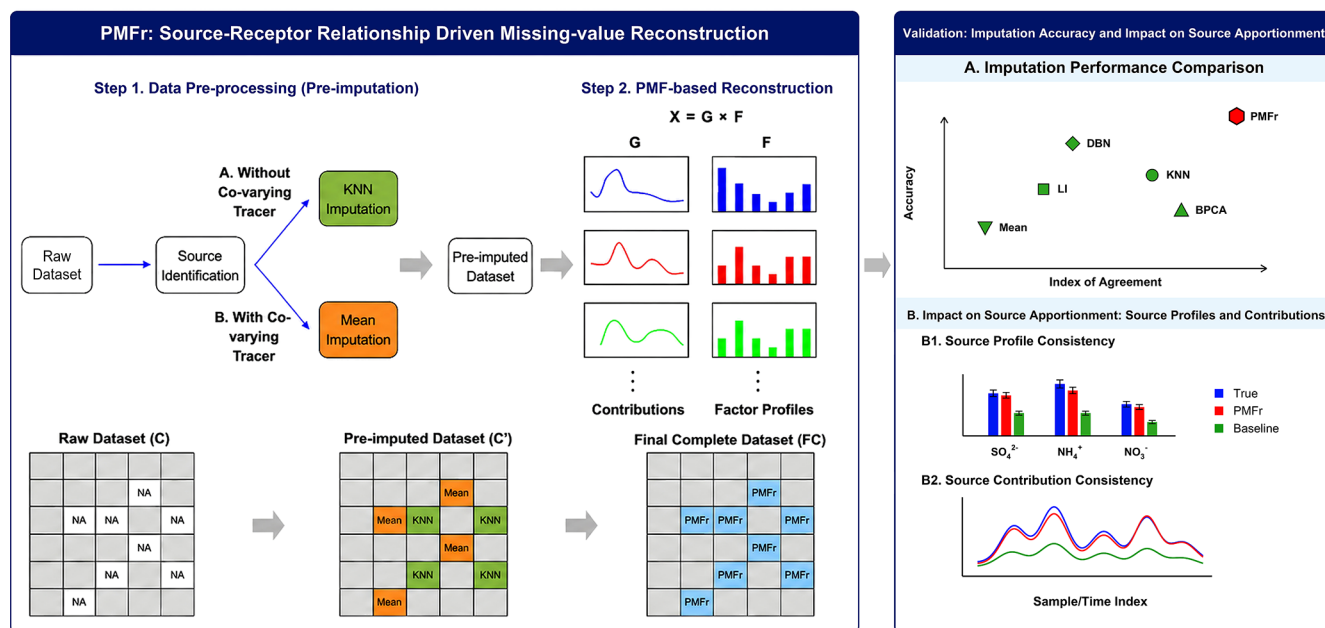
3.1 Source-receptor relationship resolved by PMF

PMF solutions were explored from four to nine factors using datasets containing 10 % missing values. The best-fitting solution was selected by the model performance, including

Table 1. Description of different missing scenarios considered in this study.

Missing Scenario	Case	Missing Compositions	Proportion (%)	Missing Pattern	Gap Length
Scenario no. 1: Random Single Species Missing	1	NH ₄ ⁺ , SO ₄ ²⁻ , NO ₃ ⁻ , Ca, Fe, Si, OC, EC	15	Missing Separately	Short
Scenario no. 2: Instrument Failure Induced Missing	2	NH ₄ ⁺ , SO ₄ ²⁻ , NO ₃ ⁻	10, 20	MCMS	Medium
	3	OC, EC	10, 20, 30	MCMS	Medium
Scenario no. 3: Station-Wide Instrument Malfunctions	4	NH ₄ ⁺ , NO ₃ ⁻	10, 20	MCMS, MCMI	Large
	5	Fe, Ca, Si, Ti	10, 20	MCMS, MCMI	Large
	6	K, NH ₄ ⁺ , NO ₃ ⁻	10, 20	MCMS, MCMI	Large
	7	K, OC, EC	10, 20	MCMS, MCMI	Large
	8	NH ₄ ⁺ , NO ₃ ⁻ , OC, EC	10, 20	MCMS, MCMI	Large
	9	K, NH ₄ ⁺ , NO ₃ ⁻ , OC, EC	10, 20	MCMI	Large

Source-Receptor Informed PMF Reconstruction (PMFr) and Validation

**Figure 1.** Flow chart of source-receptor informed Positive Matrix Factorization Reconstruction (PMFr) and validation.

the interpretability of the factor profiles, which is essential for determining the optimal factor number and imputation, and the distributions of scaled residuals (Reff et al., 2007; Brown et al., 2015; Dai et al., 2020, 2021) (Figs. S2 and S3). Bootstrapping (BS), displacement (DISP), and combined BS-DISP analyses were also performed for these solutions (Paatero et al., 2014; Liu et al., 2017; Wang et al., 2018). Four-to-six factor solutions were statistically insufficient to fully explain the variance in the input data matrix. When the factor number increased from six to seven, the Q/Q_{exp} ratio experienced a decline of 11.2%. This drop indicates that the 6-factor model leaves a substantial amount of residual variance unexplained. Specifically, the 5-factor solu-

tion improperly lumped on-road traffic emissions with metal smelting (Fig. S5). In the 6-factor solution, sulfate and nitrate were mixed together as a single identified secondary inorganic aerosol factor (Fig. S6). Eight and nine factor solutions demonstrated statistical over-resolution with diminishing returns. As the factor number increased from seven to eight, the Q/Q_{exp} ratio dropped less dramatically (8.5%) compared to the previous step. Furthermore, the 8-factor solution exhibited a high unmapped rate during the BS analysis, highlighting statistical instability. From a physical perspective, these higher-factor solutions over-resolved the data into physically meaningless profiles. For instance, the 8-factor solution isolated a Cu-high loading factor that lacks a clear chemical pro-

file (Fig. S7), while the 9-factor solution further fragmented the coal combustion into two unidentifiable sources (Fig. S8). For the 7-factor solution, the model predicted concentrations of tracers such as Ca, V, NH₄⁺, and NO₃⁻ correlated with the observed values with coefficients of determination (R^2) of 0.92, 0.91, 0.98, and 0.88, respectively (Table S4). The high R^2 values of bulk species indicate that the 7-factor model fits well for the data. For tracers like Si, Mn, Se, and Cu, the scaled residuals follow a normal distribution with a mean of 0 and a variance of 1. For bulk species like NH₄⁺, NO₃⁻, and SO₄²⁻, the scaled residuals exhibit a small-tailed distribution, with the highest frequency concentrated near 0 and ranging from -2 to 2. Additionally, the scaled residuals of the bulk species OC and EC follow a normal distribution with a mean of 0 and a variance of 1. The distribution of scaled residuals demonstrates the validity of our solution. Physically, the 7-factor solution successfully decouples all distinct emission sources without redundant splitting. The first factor was interpreted as Coal Combustion (CC), characterized by high explained variances of Pb, As, and Se (Cheng et al., 2015) and higher daytime concentrations (Fig. S10a) (Li et al., 2020). These species have relatively low DISP intervals. The Heavy Oil Combustion (HOC) was characterized by V and Ni, which are tracers of HOC (Becagli et al., 2012). The presence of HOC is consistent with the fact that Nanjing is the biggest container port on the Yangtze River. The Metal Smelting (MS) factor was identified with high Cr, Fe, Mn, Zn and Ni explained variances. Cr, Mn, Zn and Fe are typically emitted from iron and steel production (Chang et al., 2018; Pekney et al., 2006), and they exhibit relatively low DISP intervals. Cu and Ba, along with high loadings of OC and EC, serve as tracers for On-road Traffic (OT), reflecting vehicle exhaust and non-exhaust emissions such as brake and tire wear (McKenzie et al., 2009; Becagli et al., 2012). OT is also identified by high loadings of OC and EC, with the increase of concentration during the rush hour (Fig. S10d). The Crustal Dust (CD) factor is composed of crustal elements Ca, Si, Ba, Fe and Ti (Wang et al., 2018). The remaining two factors are Secondary Sulfate (SS) and Secondary Nitrate (SN), whose tracers are sulfate for SS, and ammonium and nitrate for SN, respectively. SS and SN exhibit enhanced formation around midday and nighttime, respectively (Fig. S10f and g). The following reconstruction process will be based on the 7-factor solution.

3.2 Comparison of Imputation Methods under Different Missing Scenarios

3.2.1 Overall Performance under All Missing Scenarios

As shown in Fig. S30, the PMFr method achieves the overall R^2 of 0.81 and MAPE of 22.8% under the three evaluated missing scenarios. In comparison, DBN results in an R^2 of 0.73 and a MAPE of 32.2%, BPCA yields an R^2 of 0.72 and a MAPE of 30.6%, and KNN achieves an R^2 of 0.72 and a

MAPE of 31.2%. For simple baseline methods, LI produces an R^2 of 0.35 and a high MAPE of 61.7%, while the geometric mean imputation method (Mean) results in a higher MAPE of 66.75%. Given that mean imputation produces a constant value without temporal variation and consistently fails to provide effective reconstruction across individual scenarios (Figs. S11–S29), its performance is solely quantified by MAPE here and is excluded from further detailed comparisons in subsequent sections. Furthermore, the Taylor diagram (Fig. S31) illustrates that the PMFr reconstructed data yield a normalized standard deviation (σ) of 0.93, closely matching the observational variance ($\sigma = 1.0$), suggesting its capability to capture the amplitude of data variations.

3.2.2 Scenario no. 1: Random Single Species Missing

As shown in Fig. 2, PMFr achieves the highest mean IoA (0.96) and lowest mean MAPE (16.88%), both with low standard deviations. Both PMFr and DBN attain the highest mean R^2 of 0.86, with DBN exhibiting a lower standard deviation.

For inorganic ions, PMFr performs best when imputing NH₄⁺ and NO₃⁻ with R^2 values of 0.96 and 0.91, respectively. PMFr shows the highest agreement with the observed values when imputing NH₄⁺ and NO₃⁻, especially for both low and high concentrations (Figs. S11 and S12). The performance of PMFr declines when imputing SO₄²⁻, with $R^2 = 0.79$, IoA = 0.92 and MAPE = 15.09%. Nevertheless, PMFr still outperforms LI, KNN, and BPCA. PMFr achieves higher R^2 and IoA (0.83 and 0.96, respectively), but it attains a lower MAPE (15.09%) compared to DBN (19.81%). As shown in Fig. S12, values imputed by PMFr show better agreement with true observations when the missing data correspond to low SO₄²⁻ concentrations. All methods except LI struggle to accurately impute high SO₄²⁻ concentrations. The absence of other cations like Na⁺ and Mg²⁺ may impact the imputation efficiency when the missing SO₄²⁻ concentrations are high. This difference is likely because the formation of NH₄NO₃ typically dominates the nitrate fraction, while (NH₄)₂SO₄ accounts for only a portion of the total sulfate.

For elements, PMFr performs well, with R^2 values of 0.82–0.93, IoA values of 0.95–0.98, and MAPE of 13.21%–17.17%, all accompanied by low standard deviations. Compared with PMFr, DBN performs better when imputing Fe, yielding higher R^2 and IoA, but also a higher MAPE. Conversely, PMFr shows better performance when imputing Ca, particularly for high concentration values (Fig. S14). The proposed method underestimates Fe, whereas DBN shows better consistency for missing observations that correspond to high Fe concentrations. Nevertheless, all methods fail to accurately reconstruct those high Fe concentrations (Fig. S16). LI performs better when imputing elements than ions, indicating that element concentrations fluctuate more steadily.

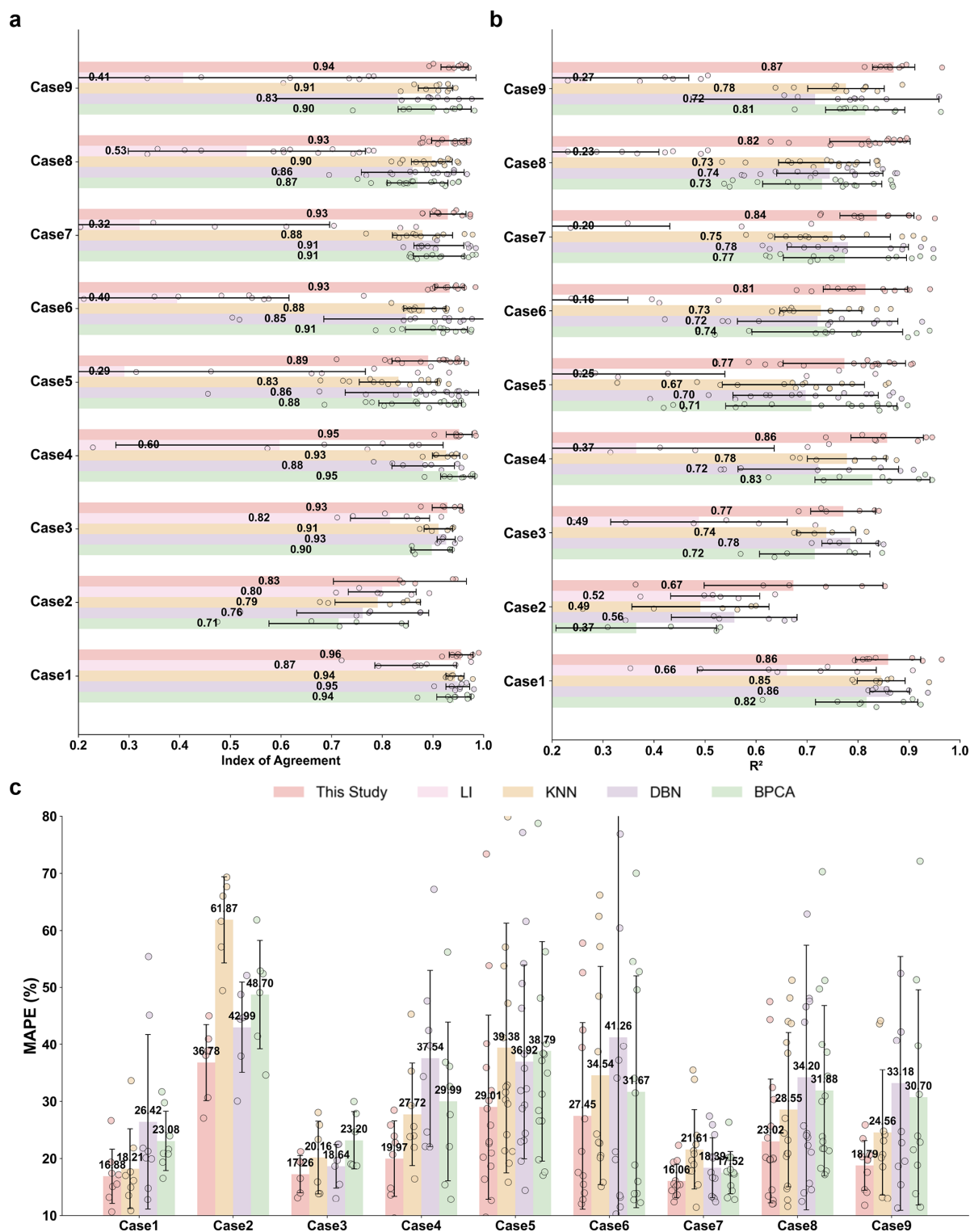


Figure 2. Performance of five imputation methods across nine Cases. Asymmetric error bars indicate the standard deviation. Points show the performance for individual species. (a) R^2 , (b) IoA, and (c) MAPE, where LI method is excluded due to poor performance under system failure conditions.

For carbonaceous materials, PMFr attains the highest IoA (0.94) for OC and the second-highest IoA (0.95) for EC, with low MAPE values of 17.42 % and 15.53 %, respectively. KNN achieves the highest R^2 (0.86 for OC and 0.87 for EC), although with lower IoA values compared to PMFr. DBN performs worse for OC, especially for low concentrations (Fig. S17). Although LI performs reasonably well for OC, it exhibits weak correlations with the true observations for EC, a trend also observed when imputing NO_3^- and SO_4^{2-} , as its performance is easily affected by the distribution pattern of missing data (Junninen et al., 2004). EC is primarily emitted from motor vehicles, whereas OC encompasses both directly emitted primary organic carbon (POC) and secondary organic carbon (SOC) formed through atmospheric processes. The behavior of POC is consistent with partial origins from vehicular emissions, while the variations of SOC are likely associated with secondary sources such as SS and SN (Liao et al., 2023). The proposed method effectively captures SOC by utilizing reasonable factor profiles, whereas other imputation methods fail to reveal the formation of SOC due to limited data. Therefore, PMFr is recommended for imputing missing components caused by random missingness.

3.2.3 Scenario no. 2: Instrument Failure Induced Missing

As shown in Fig. 2, PMFr achieves the highest mean R^2 (0.67) and IoA (0.83), and the lowest mean MAPE (36.78 %) in Case 2. Although the R^2 error bar is relatively broad, indicating species-dependent variability in correlation performance, the smaller MAPE error bar suggests that PMFr maintains more stable magnitude accuracy across species. The R^2 , IoA, and MAPE of PMFr range from 0.54–0.81, 0.59–0.95, and 27.09 %–52.01 %. Performance declines for SO_4^{2-} , with IoA values of 0.58 and 0.64 for 10 % and 20 % missingness, respectively. PMFr yields lower MAPE (34.09 % and 52.01 %) when the missing percentages are 10 % and 20 %, respectively. When imputing NH_4^+ and NO_3^- , PMFr shows the best agreement with true observed values among all methods, particularly for both low and high concentrations (Figs. S18–S21), owing to the constructed source-receptor relationships, which effectively address the difficulties that machine-learning methods face in capturing extreme values.

In Case 3, the proposed method achieves the highest mean IoA (0.93) and the lowest mean MAPE (17.26 %), both with low standard deviations. Although performance declines relative to Case 1, PMFr remains comparable to DBN, which leverages inter-variable correlations for imputation. The decline is likely attributable to the absence of key tracers, consistent with the tracer-dependent variability observed at the NEPB site – where the strong OC-EC correlation may reflect their common origin in motor-vehicle emissions (Yu et al., 2020). The performance of PMFr may be impacted because PMF tends to overestimate the loading of OC and EC

in the OT factor, thereby obscuring their contributions from other sources. Nevertheless, this highlights the interdependence between OC and EC, and the greater decline observed in KNN and BPCA compared with the proposed method.

3.2.4 Scenario no. 3: Station-Wide Instrument Malfunctions

As illustrated in Fig. 2, PMFr achieves the highest mean R^2 , IoA, and the lowest mean MAPE with low standard deviations in Case 4 (0.86 %, 0.95 %, and 19.97 %, respectively) and Case 5 (0.77 %, 0.89 %, and 29.01 %, respectively). In Case 4, PMFr captures the temporal variability of the imputed species more effectively, yielding higher R^2 and IoA values, particularly for both low and high concentrations of NH_4^+ and NO_3^- , indicating the stability of SN even under extreme missing cases. In Case 5, the imputation results show that all elemental species are well reconstructed, with Ti being the only exception. For Ti, PMFr demonstrates the highest accuracy, achieving IoA values between 0.82 and 0.95 and outperforming other methods, especially at low concentration levels (Figs. S26–S29). This improvement is likely associated with the predominant emission of Ti from dust sources, enabling PMFr to estimate missing values by leveraging the characteristic Ti-Ca-Si ratios in source profiles once the CD factor is identified (Wang et al., 2018).

PMFr consistently achieves the highest mean R^2 (0.81–0.84), IoA (0.93), and the lowest MAPE (16.06 %–27.45 %) under Cases 6–8, all accompanied by low standard deviations. Compared with Case 4, the performance of KNN and BPCA declines in Cases 6 and 8. For KNN, IoA falls from 0.93 to 0.88 (Case 6) and 0.89 (Case 8); for BPCA, IoA declines from 0.95 to 0.89 (Case 6) and 0.90 (Case 8), with both methods showing increased standard deviations. These results indicate that KNN and BPCA become unstable when additional species correlated with NH_4^+ and NO_3^- are missing, with the degradation being substantial in Case 6. In contrast, PMFr remains stable with low standard deviations because NH_4^+ and NO_3^- are estimated from source-receptor relationships – specifically the SN profile – rather than from correlations with species such as K, which are estimated via the CC and CD profiles. In Case 9, PMFr achieves the highest mean R^2 , IoA, and the lowest MAPE (0.87 %, 0.94 %, and 18.79 %, respectively). The strong performance of PMFr, KNN, and BPCA in this MCMI setting is attributable to the abundant co-occurring information, even as the number of missing species increases.

As shown in Fig. 3a and c, PMFr achieves the lowest MAPE (16.45 % and 27.93 %) and the highest R^2 (0.81 and 0.86) and IoA (both 0.94) when imputing ionic and carbonaceous species. The performance of DBN declines for ionic species, which may be attributed to insufficient valid training samples and variables caused by long missing gaps and an increasing number of missing species (Liu et al., 2022) (Figs. S22–S25). Furthermore, NH_4^+ and NO_3^- are

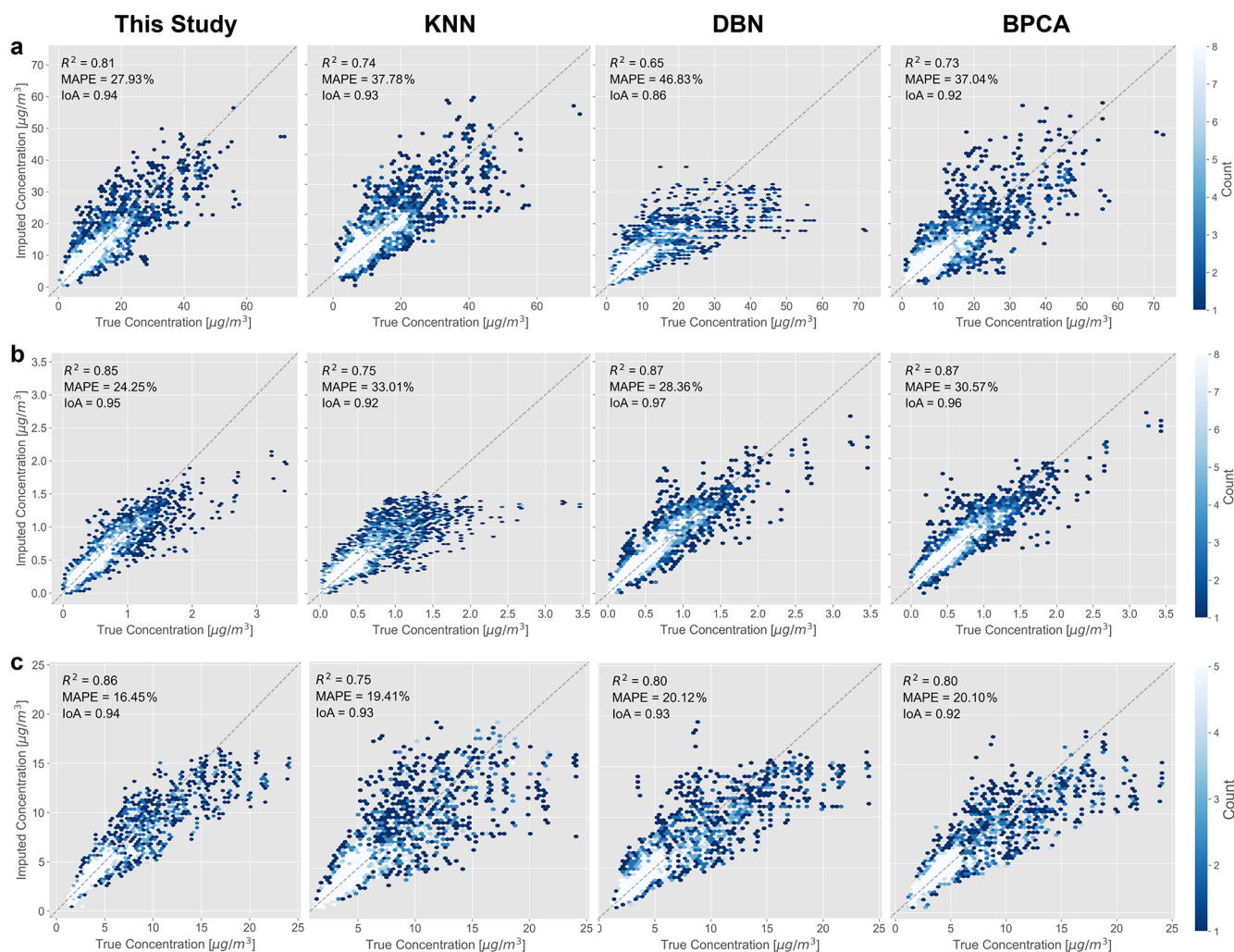


Figure 3. Comparison of observed and imputed values derived from different imputation methods under Scenario no. 3 (Cases 4–9) stratified by chemical species: (a) Inorganic ions; (b) Trace elements; and (c) Carbonaceous materials.

strongly correlated, a pattern consistent with the predominance of NH_4NO_3 during the fall at the NEPB site (Yu et al., 2020). The absence of either species therefore degrades the performance of machine learning methods, whereas PMFr can reconstruct the $\text{NH}_4^+ - \text{NO}_3^-$ relationship using the existing source profiles. When imputing OC and EC, PMFr performs best at low concentration ranges ($0\text{--}10 \mu\text{g m}^{-3}$), likely due to their relatively stable emission patterns. The limitations of machine-learning methods for imputing ionic and carbonaceous species have also been reported by Lee et al. (2023), particularly when the number of missing species increases. PMFr achieves the lowest MAPE (24.25%) for elemental species while still maintaining a high R^2 (0.85) and IoA (0.95), remaining comparable to DBN, which attains the highest R^2 (0.87) and IoA (0.97). Machine-learning methods can effectively capture correlations between a target element and co-varying elements (Li et al., 2023), and elemental species are generally emitted directly without undergo-

ing chemical reactions (Choi et al., 2022), which contributes to the strong performance of DBN when imputing elemental species.

3.3 Assessing the Impact of Imputation on PMF Source Apportionment

Results showed that the numerical advantage of PMFr over baseline methods narrowed mainly under two challenging conditions: instrument-failure-type missingness and missingness of specific species such as SO_4^{2-} and crustal elements such as Fe. Accordingly, two representative cases were selected for downstream PMF evaluation: SO_4^{2-} missingness in Case 2 at a 10% missing rate and high-concentration Fe missingness in Case 5 at a 20% missing rate. The SS and CD factors were used to assess whether these imputation differences propagated into PMF-resolved source profiles and source contributions. For the SS factor (Fig. S32), the

SO₄²⁻ / NH₄⁺ mass ratio derived from the PMFr-completed dataset was 3.39 (NH₄⁺ associated with NH₄NO₃ removed), close to that from the complete observed dataset (3.44). BPCA also produced a comparable ratio of 3.33, whereas LI (3.93), KNN (2.91), DBN (2.55), and Mean (3.83) showed larger deviations. This indicates that PMFr better preserved the SO₄²⁻–NH₄⁺ relationship in the SS profile, which is critical for maintaining the chemical interpretability of the SS factor. For the CD factor (Fig. S33), PMFr also reproduced the crustal elemental ratios consistently. The Fe / Ca and Ca / Si ratios from the complete observed dataset were 0.93 and 1.64, respectively, while PMFr yielded corresponding values of 0.95 and 1.62. In contrast, larger deviations were observed for several baseline methods, such as DBN for Fe / Ca (1.21) and BPCA or LI for Ca / Si (1.43 and 1.44, respectively). These results suggest that inappropriate imputation can alter the resolved source-profile composition, whereas PMFr maintains the physical consistency of source profiles.

For the SS factor contributions, PMFr achieved the highest Pearson's correlation coefficient (*r*) of 0.943, followed by KNN (0.914), BPCA (0.913), LI (0.900), Mean (0.802), and DBN (0.743). For the CD factor contributions, PMFr also showed the highest temporal agreement, with an *r* of 0.954, followed by Mean (0.948), KNN (0.926), BPCA (0.925), LI (0.796), and DBN (0.706). These results indicate that competitive concentration-level imputation does not necessarily guarantee equivalent preservation of PMF-resolved source-contribution patterns. As shown in Fig. S34a, b, the PMFr-derived SS contribution closely reproduced the diurnal pattern from the original complete dataset, particularly during daytime periods when secondary sulfate formation is expected to be enhanced. Similarly, PMFr captured the diurnal variation of CD more consistently than baseline methods, especially around the daytime peak likely associated with dust resuspension and other daytime dust-related activities. The selected time-series episodes showed the same behavior (Fig. S35a, b). For Case 2 at a 20 % missing rate, PMFr achieved the highest *r* of 0.985 for SS, compared with BPCA (0.981), KNN (0.980), DBN (0.954), LI (0.936), and Mean (0.705). For the high-Fe missing case, PMFr also showed the highest agreement for CD, with an *r* of 0.951, followed by Mean (0.928), BPCA (0.888), KNN (0.881), DBN (0.713), and LI (0.683). Therefore, the advantage of PMFr is not limited to pointwise concentration accuracy; it also better preserves the chemical and temporal source structures needed for physically interpretable PMF source apportionment. These results indicate that inaccurate imputation may propagate into PMF analysis and introduce source-apportionment biases, potentially making the imputed dataset less reliable than one processed using conventional PMF missing-value treatments (Kim et al., 2024).

3.4 Applicability and Limitations of PMFr

PMFr is applicable when the chemical profile of each source factor can be constrained by at least one representative tracer, ensuring that the completed remains suitable subsequent source apportionment analysis. One limitation of PMFr is related to missing patterns in which source-related constraints become insufficient. As shown in Table S7, the performance of PMFr declines when the missing pattern shifts from MCMI to MCMS. For NH₄⁺ at a 10 % missing rate, the MAPE increases from 9.57 % under MCMI to 20.67 % under MCMS, and the IoA decreases from 0.98 to 0.95. For NO₃⁻ at a 10 % missing rate, the MAPE increases from 14.82 % under MCMI to 23.92 % under MCMS. At a 20 % missing rate, the MAPE increases from 13.63 % to 25.87 % for NH₄⁺, and from 22.81 % to 28.46 % for NO₃⁻. As shown in Table S6, when OC and EC are simultaneously missing, the performance of PMFr becomes comparable to that of baseline methods. For instance, at a 10 % missing rate, the *R*² values for OC are 0.73 for PMFr, 0.74 for DBN, 0.68 for KNN, and 0.66 for BPCA. For EC, *R*² values are 0.84 for PMFr, 0.85 for BPCA, 0.80 for DBN, and 0.79 for KNN. Fundamentally, PMFr assumes that the source contribution vector (*g*) can be sufficiently constrained by observed species, which requires at least one key tracer for each factor. The key tracers used for imputation and source identification are shown in Table S13. If all key tracers for a specific source are simultaneously missing, the corresponding source contribution vector (*g*) is less directly constrained by observed species and should be interpreted with caution. Nevertheless, sensitivity analysis indicates that PMFr can still outperform baseline methods when the pre-imputation step provides a reasonable estimate of the general temporal variation of the missing species (Sect. S1.5 and Table S14). The numerical advantage of PMFr is less substantial for certain species such as SO₄²⁻ and crustal elements. For crustal elements, baseline methods can become competitive because these species are primarily emitted directly and usually exhibit relatively stable inter-variable correlations. As shown in Table S5, when imputing Ca, Si, and Fe at a 15 % missing rate, several statistical or machine-learning methods perform comparably to PMFr. For Ca, the *R*² values are 0.93 for PMFr, 0.91 for BPCA, and 0.90 for DBN. For Si, PMFr achieves an *R*² of 0.82, which is matched by DBN and closely followed by KNN (0.79). For Fe, the *R*² values are 0.83 for PMFr, 0.86 for DBN, 0.84 for KNN, and 0.84 for BPCA, with DBN and KNN achieving slightly higher IoA values than PMFr. This reduced separation suggests that statistical or machine-learning methods can capture stable co-variation patterns among some primary species, thereby reducing the relative advantage of the source-constrained PMFr method for these specific cases. However, comparable concentration-level performance does not necessarily imply that baseline methods are equally reliable for source apportionment. The PMF evaluation results showed that PMFr better preserved source-profile composi-

tion and source-contribution temporal patterns, even in representative cases where direct imputation metrics became comparable among methods.

Another limitation of the PMFr lies in the assumption of relatively stable source profiles. In PMFr, source profiles are assumed to remain stable so that the source-receptor relationships resolved by PMF can be used to guide missing-value reconstruction. This assumption is generally more reasonable for short-term datasets, but it may become weaker for long-term datasets, especially those spanning multiple years, during which emission patterns and atmospheric processes can change substantially. Therefore, source-profile stability should be evaluated before applying PMFr in extended applications. Moving-window PMF approaches provide a promising way to examine source-profile stability in long-term applications by resolving time-dependent factor profiles within short moving windows and screening accepted PMF solutions using source-specific criteria, such as factor-tracer correlations, diurnal patterns, and PMF diagnostics and non-modeled time points (Song et al., 2021; Canonaco et al., 2021). Improvements of PMFr could incorporate time-dependent source profiles to address this limitation and better support reconstruction under changing atmospheric conditions.

4 Conclusion

We developed a physically interpretable imputation method (PMFr) for reconstructing missing PM_{2.5} speciation data by leveraging source-receptor relationships encoded in key chemical species. Benchmarking against commonly used imputation techniques, including Mean, LI, KNN, BPCA, and a deep learning predictive model, demonstrates that PMFr achieves improved accuracy and robustness while preserving physical and chemical interpretability, especially for key marker species. Crucially, the PMFr-completed dataset is better suited for subsequent PMF source apportionment because it preserves source-profile composition and source-contribution temporal features. Nevertheless, the advantage of PMFr may become less substantial when source-related constraints are weakened, such as when all key tracers for a specific source factor are simultaneously missing, or when baseline methods can already capture stable co-variation patterns for certain species. These chemically consistent and physically meaningful estimates also rely on the temporal stability of source chemical compositions. Recognizing the limitations of such static assumptions for long-term datasets, we highlight the necessity of systematically verifying source stability in extended applications. Therefore, this work offers a simple and generalizable solution that strengthens the reliability of real-world speciation datasets and enhances their suitability for source apportionment and policy-relevant analyses.

Code and data availability. The PM_{2.5} speciation dataset utilized in this research is derived from previous studies (Yu et al., 2019, 2020; Xie et al., 2022). LI, KNN, and BPCA were implemented in R version 4.3.1, and DBN was applied in python 3.6.13. For the geometric mean substitution method, the geometric mean was used as the input. LI was performed using the R package “imputeTS” (Moritz and Bartz-Beielstein, 2017) (<https://CRAN.R-project.org/package=imputeTS>, last access: 19 June 2026). KNN was implemented by the R package “VIM”, which is a package designed to impute numerical, semi-continuous, and categorical variables (Kowarik and Templ, 2016) (<https://CRAN.R-project.org/package=VIM>, last access: 19 June 2026). DBN was implemented using the third-party open-source Python package deep-belief-network, available at: <https://github.com/albertbup/deep-belief-network> (last access: 19 June 2026; albertbup, 2026). BPCA was selected as it is an advanced factor-based imputation method, which is mathematically similar to the proposed approach. By comparing the imputation efficiency of the proposed method with that of BPCA method, the improvement achieved by incorporating physical information can be better demonstrated. The R package “pcaMethods” was used to implement the BPCA method (Stacklies et al., 2007) (<https://bioconductor.org/packages/pcaMethods/>, last access: 19 June 2026).

Supplement. The supplement related to this article is available online at <https://doi.org/10.5194/amt-19-4219-2026-supplement>.

Author contributions. WZ: Writing – original draft, Writing – review and editing, Visualization, Methodology, Formal analysis, Data curation. MX: Data curation, Resources. QD: Conceptualization, Supervision, Writing – review and editing. XB: Writing – review and editing. YZ: Writing – review and editing. YF: Supervision, Writing – review and editing.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher’s note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Financial support. This research has been supported by the National Natural Science Foundation of China (grant no. 42577117), the project of the Young Scientific and Technological Talents in Tianjin (grant no. QN20230350), the Tianjin Natural Science Foundation Project (grant no. 24JCYBJC01870), and the robotic AI-Scientist platform of Chinese Academy of Sciences.

Review statement. This paper was edited by Haichao Wang and reviewed by Cheng Wu and two anonymous referees.

References

- albertbup: deep-belief-network: A Python implementation of Deep Belief Networks built upon NumPy and TensorFlow with scikit-learn compatibility, GitHub repository [code], <https://github.com/albertbup/deep-belief-network>, last access: 19 June 2026.
- Alwateer, M., Atlam, E.-S., Abd El-Raouf, M. M., Ghoneim, O. A., and Gad, I.: Missing data imputation: A comprehensive review, *Journal of Computer and Communications*, 12, 53–75, <https://doi.org/10.4236/jcc.2024.1211004>, 2024.
- Becagli, S., Sferlazzo, D. M., Pace, G., di Sarra, A., Bommarito, C., Calzolari, G., Ghedini, C., Lucarelli, F., Meloni, D., Monteleone, F., Severi, M., Traversi, R., and Udisti, R.: Evidence for heavy fuel oil combustion aerosols from chemical analyses at the island of Lampedusa: a possible large role of ships emissions in the Mediterranean, *Atmos. Chem. Phys.*, 12, 3479–3492, <https://doi.org/10.5194/acp-12-3479-2012>, 2012.
- Betancourt, C., Li, C. W. Y., Kleinert, F., and Schultz, M. G.: Graph Machine Learning for Improved Imputation of Missing Tropospheric Ozone Data, *Environ. Sci. Technol.*, 57, 18246–18258, <https://doi.org/10.1021/acs.est.3c05104>, 2023.
- Bi, X., Dai, Q., Wu, J., Zhang, Q., Zhang, W., Luo, R., Cheng, Y., Zhang, J., Wang, L., Yu, Z., Zhang, Y., Tian, Y., and Feng, Y.: Characteristics of the main primary source profiles of particulate matter across China from 1987 to 2017, *Atmos. Chem. Phys.*, 19, 3223–3243, <https://doi.org/10.5194/acp-19-3223-2019>, 2019.
- Birch, M. E.: Occupational monitoring of particulate diesel exhaust by NIOSH method 5040, *Applied Occupational and Environmental Hygiene*, 17, 400–405, <https://doi.org/10.1080/10473220290035390>, 2002.
- 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, *Sci. Total Environ.*, 518–519, 626–635, <https://doi.org/10.1016/j.scitotenv.2015.01.022>, 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.
- Chang, Y., Huang, K., Xie, M., Deng, C., Zou, Z., Liu, S., and Zhang, Y.: First long-term and near real-time measurement of trace elements in China's urban atmosphere: temporal variability, source apportionment and precipitation effect, *Atmos. Chem. Phys.*, 18, 11793–11812, <https://doi.org/10.5194/acp-18-11793-2018>, 2018.
- Cheng, K., Wang, Y., Tian, H., Gao, X., Zhang, Y., Wu, X., Zhu, C., and Gao, J.: Atmospheric Emission Characteristics and Control Policies of Five Precedent-Controlled Toxic Heavy Metals from Anthropogenic Sources in China, *Environ. Sci. Technol.*, 49, 1206–1214, <https://doi.org/10.1021/es5037332>, 2015.
- Choi, E., Yi, S.-M., Lee, Y. S., Jo, H., Baek, S.-O., and Heo, J.-B.: Sources of airborne particulate matter-bound metals and spatial-seasonal variability of health risk potentials in four large cities, South Korea, *Environ. Sci. Pollut. R.*, 29, 28359–28374, <https://doi.org/10.1007/s11356-021-18445-8>, 2022.
- Dai, Q., Liu, B., Bi, X., Wu, J., Liang, D., Zhang, Y., Feng, Y., and Hopke, P. K.: Dispersion Normalized PMF Provides Insights into the Significant Changes in Source Contributions to PM_{2.5} after the COVID-19 Outbreak, *Environ. Sci. Technol.*, 54, 9917–9927, <https://doi.org/10.1021/acs.est.0c02776>, 2020.
- Dai, Q., Ding, J., Song, C., Liu, B., Bi, X., Wu, J., Zhang, Y., Feng, Y., and Hopke, P. K.: Changes in source contributions to particle number concentrations after the COVID-19 outbreak: Insights from a dispersion normalized PMF, *Sci. Total Environ.*, 759, 143548, <https://doi.org/10.1016/j.scitotenv.2020.143548>, 2021.
- Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O.: A survey on missing data in machine learning, *Journal of Big Data*, 8, 1–37, <https://doi.org/10.1186/s40537-021-00516-9>, 2021.
- García-Laencina, P. J., Sancho-Gómez, J.-L., and Figueiras-Vidal, A. R.: Pattern classification with missing data: a review, *Neural Comput. Appl.*, 19, 263–282, <https://doi.org/10.1007/s00521-009-0295-6>, 2010.
- Hao, H., Wang, Y., Zhu, Q., Zhang, H., Rosenberg, A., Schwartz, J., Amini, H., van Donkelaar, A., Martin, R., Liu, P., Weber, R., Russell, A., Yitshak-sade, M., Chang, H., and Shi, L.: National Cohort Study of Long-Term Exposure to PM_{2.5} Components and Mortality in Medicare American Older Adults, *Environ. Sci. Technol.*, 57, 6835–6843, <https://doi.org/10.1021/acs.est.2c07064>, 2023.
- Hopke, P. K.: Review of receptor modeling methods for source apportionment, *J. Air Waste Manage.*, 66, 237–259, <https://doi.org/10.1080/10962247.2016.1140693>, 2016.
- Hu, J., Zhang, H., Chen, S., Ying, Q., Wiedinmyer, C., Vandenberghe, F., and Kleeman, M. J.: Identifying PM_{2.5} and PM_{0.1} Sources for Epidemiological Studies in California, *Environ. Sci. Technol.*, 48, 4980–4990, <https://doi.org/10.1021/es404810z>, 2014.
- Jing, X., Luo, J., Wang, J., Zuo, G., and Wei, N.: A Multi-imputation method to deal with hydro-meteorological missing values by integrating chain equations and random forest, *Water Resour. Manag.*, 36, 1159–1173, <https://doi.org/10.1007/s11269-021-03037-5>, 2022.
- Junger, W. and Ponce de Leon, A.: Imputation of missing data in time series for air pollutants, *Atmos. Environ.*, 102, 96–104, <https://doi.org/10.1016/j.atmosenv.2014.11.049>, 2015.
- Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M.: Methods for imputation of missing values in air quality data sets, *Atmos. Environ.*, 38, 2895–2907, <https://doi.org/10.1016/j.atmosenv.2004.02.026>, 2004.
- Khan, S. I. and Hoque, A. S. M. L.: SICE: an improved missing data imputation technique, *Journal of Big Data*, 7, 37, <https://doi.org/10.1186/s40537-020-00313-w>, 2020.
- Kim, Y., Yi, S.-M., Heo, J., Kim, H., Lee, W., Kim, H., Hopke, P. K., Lee, Y. S., Shin, H.-J., Park, J., Yoo, M., Jeon, K., and Park, J.: Is replacing missing values of PM_{2.5} constituents with estimates using machine learning better for source apportionment than exclusion or median replacement?, *Environ. Pollut.*, 354, 124165, <https://doi.org/10.1016/j.envpol.2024.124165>, 2024.

- Kim, Y., Hopke, P. K., Yi, S.-M., Lee, W., Kim, H., Heo, J., Kim, H., Lee, Y. S., Jeon, K., and Park, J.: Positive matrix factorization outperforms machine learning in imputing missing PM_{2.5} and further identifying spatial patterns in multi-sites without external data, *Urban Climate*, 62, 102552, <https://doi.org/10.1016/j.uclim.2025.102552>, 2025a.
- Kim, Y., Kang, C., Yi, S.-M., Heo, J., Kim, H., Lee, W., Kim, H., Hopke, P. K., Lee, Y. S., Shin, H.-J., Park, J., Yoo, M., Jeon, K., and Park, J.: Imputing missing data with statistical-learning estimates: impacts on mortality risks attributable to area- and source-specific PM_{2.5}, *Atmos. Pollut. Res.*, 102785, <https://doi.org/10.1016/j.apr.2025.102785>, 2025b.
- Kowarik, A. and Templ, M.: Imputation with the R Package VIM, *J. Stat. Softw.*, 74, 1–16, <https://doi.org/10.18637/jss.v074.i07>, 2016.
- Lai, W. Y. and Kuok, K.: A study on bayesian principal component analysis for addressing missing rainfall data, *Water Resour. Manag.*, 33, 2615–2628, <https://doi.org/10.1007/s11269-019-02209-8>, 2019.
- Lee, S.-J., Ju, J.-T., Lee, J.-J., Song, C.-K., Shin, S.-A., Jung, H.-J., Shin, H. J., and Choi, S.-D.: Mapping nationwide concentrations of sulfate and nitrate in ambient PM_{2.5} in South Korea using machine learning with ground observation data, *Sci. Total Environ.*, 926, 171884, <https://doi.org/10.1016/j.scitotenv.2024.171884>, 2024.
- Lee, Y. S., Choi, E., Park, M., Jo, H., Park, M., Nam, E., Kim, D. G., Yi, S.-M., and Kim, J. Y.: Feature extraction and prediction of fine particulate matter (PM_{2.5}) chemical constituents using four machine learning models, *Expert Syst. Appl.*, 221, 119696, <https://doi.org/10.1016/j.eswa.2023.119696>, 2023.
- Li, R., Wang, Q., He, X., Zhu, S., Zhang, K., Duan, Y., Fu, Q., Qiao, L., Wang, Y., Huang, L., Li, L., and Yu, J. Z.: Source apportionment of PM_{2.5} in Shanghai based on hourly organic molecular markers and other source tracers, *Atmos. Chem. Phys.*, 20, 12047–12061, <https://doi.org/10.5194/acp-20-12047-2020>, 2020.
- Li, R., Gao, Y., Chen, Y., Peng, M., Zhao, W., Wang, G., and Hao, J.: Measurement report: Rapid changes of chemical characteristics and health risks for highly time resolved trace elements in PM_{2.5} in a typical industrial city in response to stringent clean air actions, *Atmos. Chem. Phys.*, 23, 4709–4726, <https://doi.org/10.5194/acp-23-4709-2023>, 2023.
- Liao, K., Wang, Q., Wang, S., and Yu, J. Z.: Bayesian Inference Approach to Quantify Primary and Secondary Organic Carbon in Fine Particulate Matter Using Major Species Measurements, *Environ. Sci. Technol.*, 57, 5169–5179, <https://doi.org/10.1021/acs.est.2c09412>, 2023.
- Little, R. J. and Rubin, D. B.: *Statistical analysis with missing data*, John Wiley & Sons, <https://doi.org/10.1002/9781119482260>, 2019.
- Liu, B., Wu, J., Zhang, J., Wang, L., Yang, J., Liang, D., Dai, Q., Bi, X., Feng, Y., Zhang, Y., and Zhang, Q.: Characterization and source apportionment of PM_{2.5} based on error estimation from EPA PMF 5.0 model at a medium city in China, *Environ. Pollut.*, 222, 10–22, <https://doi.org/10.1016/j.envpol.2017.01.005>, 2017.
- Liu, M. and Matsui, H.: Aerosol radiative forcings induced by substantial changes in anthropogenic emissions in China from 2008 to 2016, *Atmos. Chem. Phys.*, 21, 5965–5982, <https://doi.org/10.5194/acp-21-5965-2021>, 2021.
- Liu, X., Fu, Y., Wang, Q., Bi, Y., Zhang, L., Zhao, G., Xian, F., Cheng, P., Zhang, L., Zhou, J., and Zhou, W.: Unraveling the process of aerosols secondary formation and removal based on cosmogenic beryllium-7 and beryllium-10, *Sci. Total Environ.*, 821, 153293, <https://doi.org/10.1016/j.scitotenv.2022.153293>, 2022.
- McKenzie, E. R., Money, J. E., Green, P. G., and Young, T. M.: Metals associated with stormwater-relevant brake and tire samples, *Sci. Total Environ.*, 407, 5855–5860, <https://doi.org/10.1016/j.scitotenv.2009.07.018>, 2009.
- Moritz, S. and Bartz-Beielstein, T.: imputeTS: Time Series Missing Value Imputation in R, *R J.*, 9, 207–218, <https://doi.org/10.32614/RJ-2017-009>, 2017.
- Norris, G., Duvall, R., Brown, S., and Bai, S.: EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide, U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC, USA, EPA/600/R-14/108, https://www.epa.gov/sites/default/files/2015-02/documents/pmf_5.0_user_guide.pdf (last access: 19 June 2026), 2014.
- 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, <https://doi.org/10.1080/10618600.1999.10474853>, 1999.
- 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](https://doi.org/10.1016/S0003-2670(02)01643-4), 2003.
- 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.
- Pekney, N. J., Davidson, C. I., Robinson, A., Zhou, L., Hopke, P., Eatough, D., and Rogge, W. F.: Major source categories for PM_{2.5} in Pittsburgh using PMF and UNMIX, *Aerosol Sci. Tech.*, 40, 910–924, <https://doi.org/10.1080/02786820500380271>, 2006.
- Peng, X., Xie, T.-T., Tang, M.-X., Cheng, Y., Peng, Y., Wei, F.-H., Cao, L.-M., Yu, K., Du, K., He, L.-Y., and Huang, X.-F.: Critical Role of Secondary Organic Aerosol in Urban Atmospheric Visibility Improvement Identified by Machine Learning, *Environ. Sci. Technol. Letters*, 10, 976–982, <https://doi.org/10.1021/acs.estlett.3c00084>, 2023.
- Plaia, A. and Bondì, A.: Single imputation method of missing values in environmental pollution data sets, *Atmos. Environ.*, 40, 7316–7330, <https://doi.org/10.1016/j.atmosenv.2006.06.040>, 2006.
- Polissar, A. V., Hopke, P. K., Paatero, P., Malm, W. C., and Sisler, J. F.: Atmospheric aerosol over Alaska: 2. Elemental composition and sources, *J. Geophys. Res.-Atmos.*, 103, 19045–19057, <https://doi.org/10.1029/98JD01212>, 1998.
- Reff, A., Eberly, S. I., and Bhawe, P. V.: Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods, *J. Air Waste Manage.*, 57, 146–154, <https://doi.org/10.1080/10473289.2007.10465319>, 2007.
- Richardson, A. D. and Hollinger, D. Y.: A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO₂ flux record, *Agr. Forest Meteorol.*, 147, 199–208, <https://doi.org/10.1016/j.agrformet.2007.06.004>, 2007.
- Samal, K. K. R., Babu, K. S., and Das, S. K.: Multi-directional temporal convolutional artificial neural network for PM_{2.5} forecast-

- ing with missing values: A deep learning approach, *Urban Climate*, 36, 100800, <https://doi.org/10.1016/j.uclim.2021.100800>, 2021.
- Shen, H., Li, T., Yuan, Q., and Zhang, L.: Estimating Regional Ground-Level PM_{2.5} Directly From Satellite Top-Of-Atmosphere Reflectance Using Deep Belief Networks, *J. Geophys. Res.-Atmos.*, 123, 13875–13886, <https://doi.org/10.1029/2018JD028759>, 2018.
- Song, L., Dai, Q., Feng, Y., and Hopke, P. K.: Estimating uncertainties of source contributions to PM_{2.5} using moving window evolving dispersion normalized PMF, *Environ. Pollut.*, 286, 117576, <https://doi.org/10.1016/j.envpol.2021.117576>, 2021.
- Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J.: *pcaMethods* – a bioconductor package providing PCA methods for incomplete data, *Bioinformatics*, 23, 1164–1167, <https://doi.org/10.1093/bioinformatics/btm069>, 2007.
- van Donkelaar, A., Martin, R. V., Li, C., and Burnett, R. T.: Regional Estimates of Chemical Composition of Fine Particulate Matter Using a Combined Geoscience-Statistical Method with Information from Satellites, Models, and Monitors, *Environ. Sci. Technol.*, 53, 2595–2611, <https://doi.org/10.1021/acs.est.8b06392>, 2019.
- Wang, Q., Qiao, L., Zhou, M., Zhu, S., Griffith, S., Li, L., and Yu, J. Z.: Source Apportionment of PM_{2.5} Using Hourly Measurements of Elemental Tracers and Major Constituents in an Urban Environment: Investigation of Time-Resolution Influence, *J. Geophys. Res.-Atmos.*, 123, 5284–5300, <https://doi.org/10.1029/2017JD027877>, 2018.
- Xie, J.: Deep Neural Network for PM_{2.5} Pollution Forecasting Based on Manifold Learning, in: 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), 236–240, <https://doi.org/10.1109/SDPC.2017.52>, 2017.
- Xie, M., Lu, X., Ding, F., Cui, W., Zhang, Y., and Feng, W.: Evaluating the influence of constant source profile presumption on PMF analysis of PM_{2.5} by comparing long- and short-term hourly observation-based modeling, *Environ. Pollut.*, 314, 120273, <https://doi.org/10.1016/j.envpol.2022.120273>, 2022.
- Yu, Y., Yu, J. J., Li, V. O. K., and Lam, J. C. K.: Low-rank singular value thresholding for recovering missing air quality data, in: 2017 IEEE International Conference on Big Data (Big Data), 508–513, <https://doi.org/10.1109/BigData.2017.8257965>, 2017.
- Yu, Y., He, S., Wu, X., Zhang, C., Yao, Y., Liao, H., Wang, Q., and Xie, M.: PM_{2.5} elements at an urban site in Yangtze River Delta, China: High time-resolved measurement and the application in source apportionment, *Environ. Pollut.*, 253, 1089–1099, <https://doi.org/10.1016/j.envpol.2019.07.096>, 2019.
- Yu, Y., Ding, F., Mu, Y., Xie, M., and Wang, Q.: High time-resolved PM_{2.5} composition and sources at an urban site in Yangtze River Delta, China after the implementation of the APPCAP, *Chemosphere*, 261, 127746, <https://doi.org/10.1016/j.chemosphere.2020.127746>, 2020.
- Zaini, N., Ean, L. W., Ahmed, A. N., and Malek, M. A.: A systematic literature review of deep learning neural network for time series air quality forecasting, *Environ. Sci. Pollut. R.*, 1–33, <https://doi.org/10.1007/s11356-021-17442-1>, 2022.