https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

Improving Imputation of Missing $PM_{2.5}$ Speciation Data Using PMF-Informed Source–Receptor Relationships

Wubin Zhu[1], Mingjie Xie[2], Qili Dai[1,3], Xiaohui Bi[1], Yufen Zhang[1], and Yinchang Feng[1]

[1]State 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
[2]Collaborative 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
[3]Tianjin Key Laboratory of Software Experience and Human Computer Interaction, Tianjin 300457, China

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

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 K-nearest neighbor (KNN), Bayesian principal component analysis (BPCA), and deep learning architectures, rely primarily on statistical correlations, requiring 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 against commonly used imputation techniques KNN, BPCA, and deep learning predictive model demonstrates that PMFr achieves superior accuracy and robustness under all 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, highlighting its potential as a robust tool for recovering reliable data in air quality studies.

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

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

scientific conclusions.

A wide range of methods have been developed to address missing values in $PM_{2.5}$ chemical component datasets, generally 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(Hopke, 2000), 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 K-nearest neighbors (KNN), Bayesian principal component analysis (BPCA), and deep learning architectures 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 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 multipling the estimated source-specific $PM_{2.5}$ mass, 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 mean substitution, linear interpolation, K-nearest neighbors (KNN), deep belief networks (DBN), and Bayesian principal component analysis (BPCA). Results highlight the potential of a source-informed strategy for robust and interpretable imputation.

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 NH_4^+, SO_4^{2-}, NO_3^-, Cl^-, Ca^{2+}, Mg^{2+}, K^+, and Na^+ were determined by MARGA (ADI 2080; Applikon Analytical B.V., 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

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

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 set up and maintenance, and chemical analysis were provided by previous study(Yu et al., 2019, 2020; Xie et al., 2022).

The dataset used in this study comprises inorganic ions (NH_4^+, SO_4^{2-}, and NO_3^-), 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 October 1, 2017, 1:00 AM (local time, GMT +8) to November 30, 2017, 11:00 PM, a period without any major pollution incidents(Xie et al., 2022). The summary for the missing of raw data can be seen in Table S1.

2.2 Missing Data Generation

Four factors are considered to impact the efficiency of imputation method: the generating 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).

The mechanisms of missing value are typically classified into three categories(Little and Rubin, 2019): i) missing completely at random (MCAR), where missing values are generated independently, namely not relying on any values. ii) missing at random (MAR), where missing values have relationship with observed data, and iii) missing not at random (MNAR), where missing values are related with unobserved data such as values below detection limitation (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 NEPB 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 Bondì, 2006; Betancourt et al., 2023; Jing et al., 2022; Richardson and Hollinger, 2007), categorizing the gap length (l) in different columns 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 (one to five days), which represents the longest gap observed in the raw dataset (Table S2).

MCMS refers to the simultaneous absence of multiple species at a single timestamp, while MCMI denotes missing values occurring independently across time.

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

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

malfunction of two instruments (Cases 6–8), and malfunction of all monitoring instruments (Case 9). Potassium (K) was treat as missing in multi-instrument malfunction scenarios(Cases 6, 7 and 9) due to its strong correlations with both ionic and carbonaceous species (Figure S1). Performance of imputation methods are compared by indicators including:the coefficient of determination (R^2), and mean absolute percentage error, and index of agreement (IoA).

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

Missing Scenario	Case	Missing Compositions	Proportion (%)	Missing Pattern	Gap Length
Scenario#1: Random Single Species Missing	1	NH_4, SO_4, NO_3, Ca, Fe, Si, OC, EC	15	Missing Separately	Short
Scenario#2: Instrument Failure Induced Missing	2	NH_4, SO_4, NO_3	10, 20	MCMS	Medium
	3	OC, EC	10, 20, 30	MCMS	Medium
Scenario#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

2.3 Source–Receptor Informed Positive Matrix Factorization Reconstruction (PMFr)

A tracer for imputation (referred to as tracer) is defined as key species that distinguishes a specific factor from others and reflects how that factor influences the receptor over time. Co-tracers refer to additional tracers within the same factor that collectively characterize the temporal behavior of the corresponding source. As illustrated in Figure 1a, 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. Details of the usage of PMF for SA can be found in the literature and the sets of uncertainty can be seen in Text S3(Hopke, 2016; Paatero, 1999; Paatero and Hopke, 2003). Based on the prior SA results with selected source profiles, a knowledge-driven step(Bi et al., 2019), the species to be imputed can be classified as either tracers or non-tracers. For tracers, missing values requires to be firstly imputed by another imputation method (KNN recommended for its simplicity and efficiency) and correspondent uncertainty is set as 0.1 as their imputed value. For tracers with the existence of co-tracer and non-tracers, missing values are imputed with geometric mean of certain species and correspondent uncertainty are set as 8 times the geometric mean. After data processing, the dataset is divided into Fs and Gs through a PMF run. The reconstruction is by multiplying factor profiles (Fs) with contributions (Gs). After these steps, the dataset and its associated uncertainty are input into the PMF run. In this study, US EPA PMF v5.0 was utilized for PMFr run for imputation.

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

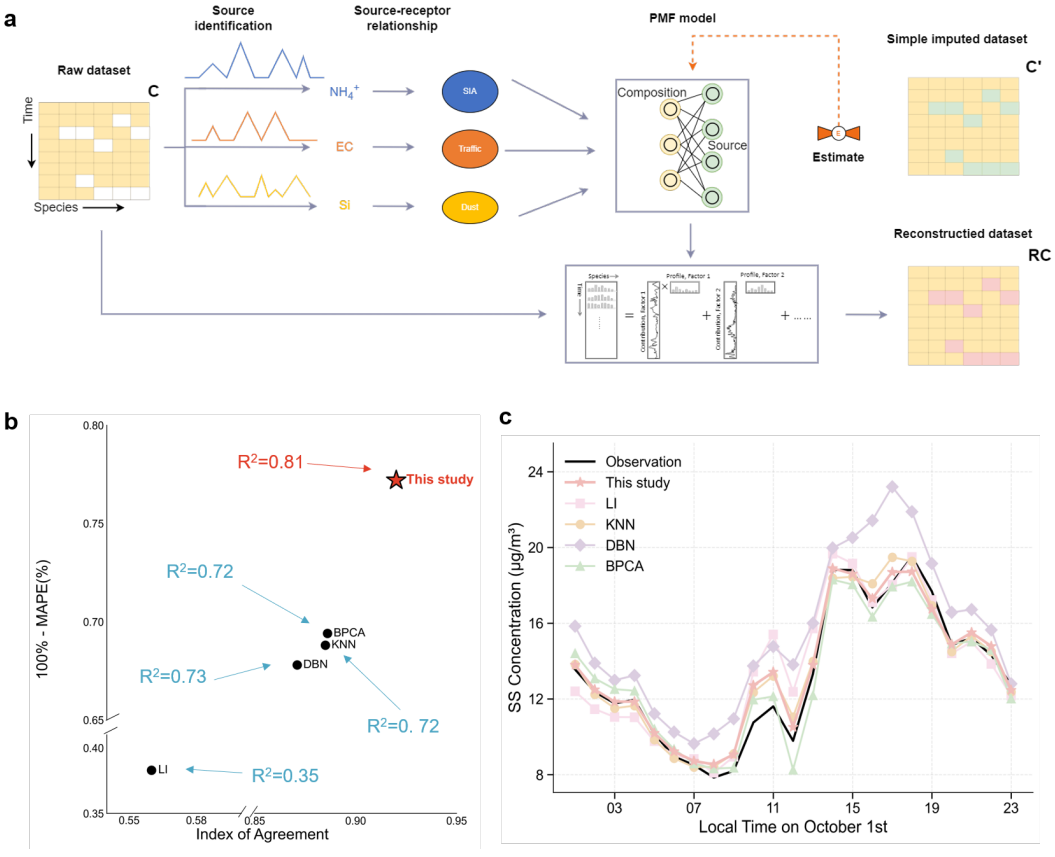

Figure 1. Source–Receptor Informed Positive Matrix Factorization Reconstruction. **a**, Workflow of the proposed PMFr imputation framework. **b**, Mean performance of the evaluated imputation methods under all cases. **c**, Comparison of PMF solved source contributions obtained from datasets imputed using different methods with those derived from the non-missing dataset under Case 2.

3 Results and Discussion

3.1 Source-receptor relationship resolved by PMF

PMF solutions were explored with four to nine factors using datasets containing 10% missing values. The best-fitting solution were selected by the model performance, including the interpretability of the factor profiles, which is the key basis for determining the optimal factor number and imputation, and the distributions of scaled residuals(Reff et al., 2007; Brown et al., 2015) (Figures S2 and S3). Bootstrapping (BS), displacement (DISP), and combined BS-DISP analyses were also performed for these solutions(Paatero et al., 2014). The factor profiles for 7-factor solution were chosen as the optimal fits for the data. The model predicted concentrations of tracers such as Ca, V, NH_4^+, and NO_3^- 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 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

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

0 and a variance of 1. For bulk species like NH_4^+, NO_3^-, and SO_4^{2-}, the scaled residuals exhibit a light-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. The change in the Q/Q_{exp} ratio with the factor number, where Q is the sum of the squared and scaled residues of PMF results and Q_{exp} equals to the number of elements in input data matrix, is also an indicator for the selection of an appropriate factor number(Liu et al., 2017; Wang et al., 2018). After each base case analysis, 100 BS runs with the r value set to 0.8 were conducted to obtain factor mapping rates between bootstrapped and base-case factor contributions, and the characteristic species in all factor profiles (Ca, Ti, As, V, Ba, Cr, Se, Ni, NH_4^+, SO_4^{2-}, NO_3^-, OC, and EC) were displaced for BS-DISP analysis. As shown in Table S3 and Figure S7, the 7-factor solution has the second highest mapping rates of BS runs >90 with no swap in DISP. The Q/Q_{exp} ratio drops less dramatically from the 7- to 8-factor solutions (8.5%) than it does when the factor number increases from six to seven (11.2%). The BS-DISP run shows the largest decrease in Q for 7-factor solution is 0.040%. When considering factor profiles, the 6-factor solution lumped on-road traffic emissions with the SS factor (Figure S6), and the metal smelting and coal combustion factor were divided into three unexplainable factors when increasing the factor number to eight (Figure S8). These two factors are further divided into four unexplainable factors when the factor number increased from 8 to 9 (Figure S9). All these results support the use of a 7-factor solution to explain the input data(Kim et al., 2005; Kim and Hopke, 2007; Tian et al., 2016).

The first factor was interpreted as Coal Combustion (CC), with the high Pb, As and Se explained variances(Cheng et al., 2015) and exhibiting higher daytime concentrations(Li et al., 2020) (Figure S10a). 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 (Figure S10d). Crustal Dust (CD) factor are 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 (Figure S10f and Figure S10g). 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 Scenario#1: Random Single Species Missing

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

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

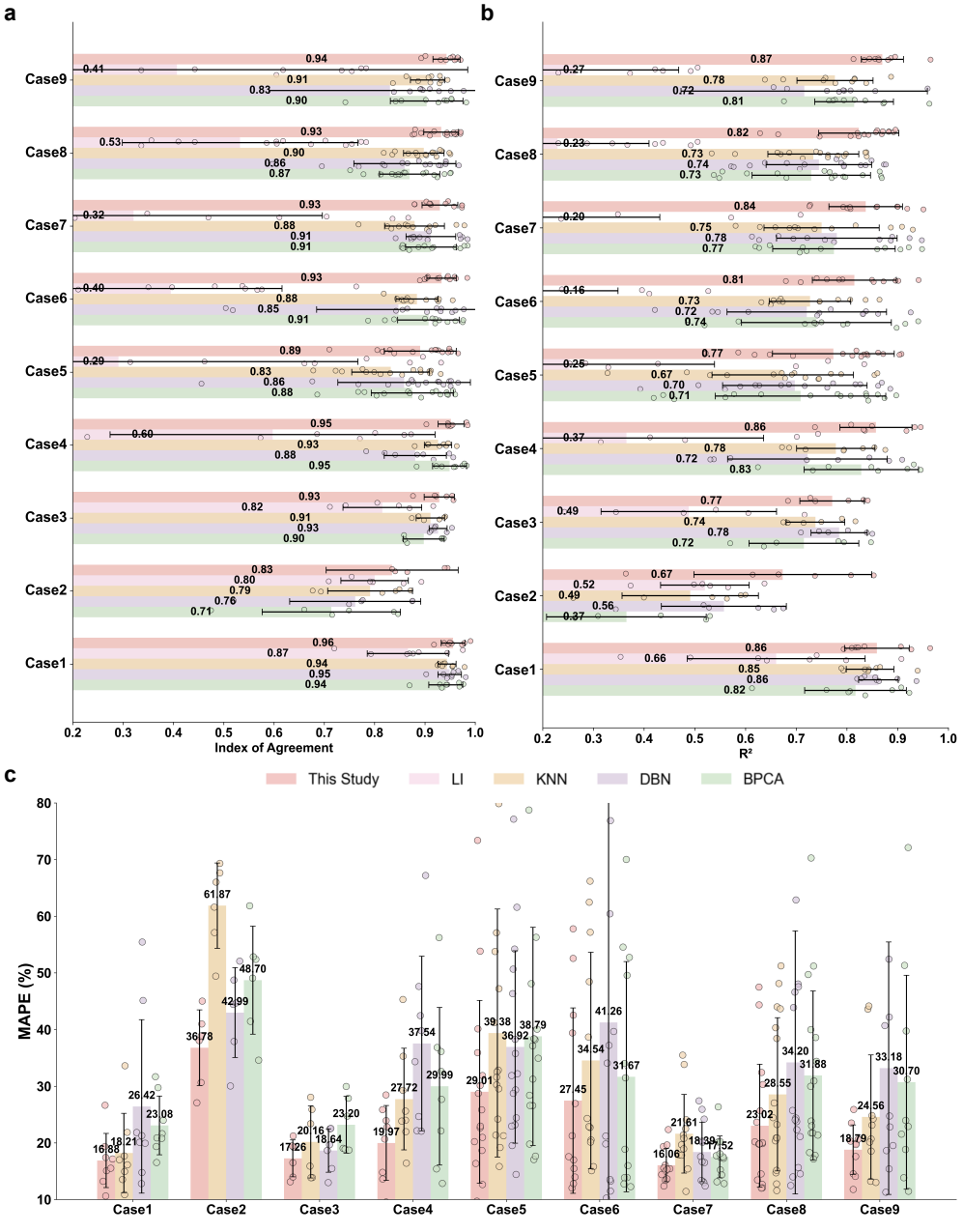

Figure 2. Performance of five imputation methods across nine Cases. Asymmetric error bars indicate the standard deviation. Points show the performance for individual species. **a** R^2, **b** IoA, and **c** MAPE, where LI method is excluded due to poor performance under system failure conditions.

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

For inorganic ions, PMFr performs best when imputing NH_4^+ and NO_3^- according to the trend indicator R^2, 0.96 and 0.91, respectively. PMFr exhibits highest accordance when imputing NH_4^+ and NO_3^- compared to other methods, especially for low and high missing values (Figures S11 and S12). The performance of PMFr decline when imputing SO_4^{2-}, with R^2=0.79, IoA=0.92 and MAPE=15.09%. Nevertheless PMFr still outperforms LI, KNN, and BPCA. DBN 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 Figure S12, values imputed by PMFr show better agreement with true observations when the missing data correspond to low SO_4^{2-} concentrations. All methods except LI struggle to impute high SO_4^{2-} concentrations accurately. The absence of other cations like Na^+ and Mg^{2+} impact the imputation efficiency when missing SO_4^{2-} concentration values are high. The formation of NH_4NO_3 dominates nitrate, while $(NH_4)_2SO_4$ account for only part of sulfate. SN can capture source contributions when one tracer is missing, while SS, lacking enough information on pollutant sources, cannot capture source contributions as accurately in the absence of its only tracer.

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 deviation. Compared with PMFr, DBN performs better when imputing Fe, yielding higher R^2 and IoA, but also higher MAPE. Conversely, PMFr shows better performance when imputing Ca, particularly for high concentration values (Figure 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 (Figure S16). LI performs better when imputing elements than ions, indicating that element concentration fluctuates more steadily.

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 also performs when imputing OC and EC, achieving 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 (Figure 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 consists of both primary organic carbon (POC) and secondary organic carbon (SOC); POC is directly emitted, while SOC forms in the atmosphere through secondary processes. POC can partially originate from motor vehicles, whereas SOC is 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 missing.

3.2.2 Scenario#2: Instrument Failure Induced Missing

As shown in Figure 2, PMFr achieves the highest mean R^2 (0.67) and IoA (0.83) with broad narrow bar, and the lowest mean MAPE (36.78%), all with relatively low standard deviation in Case 2. 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 percentage are 10% and 20%,

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

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 (Figures S18, S19, S20, and S21), owing to constructed source–receptor relationships, which effectively address the difficulty machine-learning methods face in capturing extreme values. When all ions are missing, all methods except LI exhibit a decline in performance, suggesting that LI can also serve as a simple and viable imputation approach for PMFr. The proposed method remains robust for NH_4^+ and NO_3^- and maintains acceptable performance for SO_4^{2-}, highlighting the importance of incorporating chemical balance principles to constrain the plausible range of the missing species.

In Case 3, the proposed method achieves the highest mean IoA (0.93) and the lowest mean MAPE (17.26%), both with low standard deviation. Although performance declines relative to Case 1, PMFr remains comparable to DBN, which leverages inter-variable correlations for imputation. The decline is attributable to the absence of key tracers, consistent with the tracer-dependent variability observed at the NEPB site—where the strong OC–EC correlation reflects their common origin in motor-vehicle emissions(Yu et al., 2020). PMFr is affected because PMF overestimate the loading of OC and EC in the OT factor, thereby underscoring 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. Overall, PMFr is well-suited for missingness due to instrument failure, given its accuracy and low variability. As illustrated in Figure 1b, SO_4^{2-} imputed by PMFr shows best agreement with observed values. When the PMF model is applied to the dataset imputed by PMFr, its results exhibit the best agreement with those from the complete dataset.

3.2.3 Scenario#3: Station-Wide Instrument Malfunctions

As illustrated in Figure 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 capture 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 (Figures S26, S27, S28, and S29). This improvement arises because Ti is predominantly emitted from dust sources, enabling PMFr to estimate missing values using 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 Case 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, the degradation being most pronounced in Case 6. In contrast, PMFr remains stable with low standard deviation 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

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

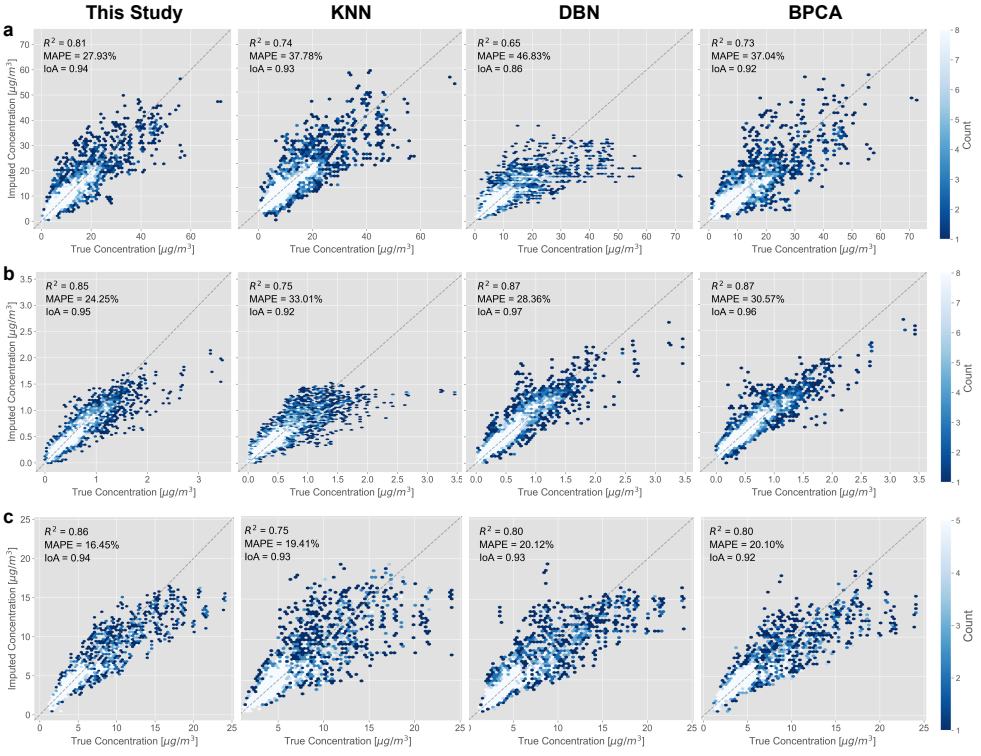

Figure 3. Comparison of observed and imputed values derived from different imputation methods under Scenarios#3 (Cases 4–9) stratified by chemical species: **a** Inorganic ions; **b** Trace elements; and **c** Carbonaceous materials.

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 Figure 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 due to insufficient valid training samples and variables caused by long missing gaps and increasing number of missing species(Liu et al., 2022)(Figures S22, S23, S24, and S25). NH_4^+ and NO_3^- are strongly correlated due to the predominance of NH_4NO_3 during fall in NEPB site(Yu et al., 2020). The absence of either species therefore degrades the performance of machine learning method, whereas PMFr can reconstruct the NH_4^+-NO_3^- relationship using the existing source profiles. When imputing OC and EC, PMFr performs best at low concentration ranges (0-10 $\mu g/m^3$), likely due to their relatively stable emission patterns. The limitations of machine-learning methods for imputing ionic and carbonaceous species have also been proved by Lee et al.(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

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

and co-varying elements(Li et al., 2023), and elemental species are generally emitted directly without undergoing chemical reactions(Choi et al., 2022), which contributes to the strong performance of DBN when imputing elemental species.

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. By assuming temporal stability in source chemical compositions, the approach yields chemically consistent and physically meaningful estimates. Benchmarking against commonly used imputation techniques LI, KNN, BPCA and deep learning predictive model demonstrates improved accuracy and robustness, preserving physical and chemical interpretability, especially for key marker species. The method is not limited to particulate matter and can be extended to other atmospheric datasets that contain source-related information, such as VOC measurements. Its performance holds potential for continued enhancement as more advanced source apportionment techniques evolve. Therefore, this work offers a simple, 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 mean imputation 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/web/packages/imputeTS/index.html). 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/web/packages/VIM/index.html). DBN is a deep learning method which is capable of solving non-linear problems (https://github.com/albertbup/deep-belief-network). 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://rdocumentation.org/packages/pcaMethods).

Author contributions. Wubin Zhu: Writing – original draft, Writing – review and editing, Visualization, Methodology, Formal analysis, Data curation. Mingjie Xie: Data curation, Resources. Qili Dai: Conceptualization, Supervision, Writing – review and editing. Xiaohui Bi: Writing – review and editing. Yufen Zhang: Writing – review and editing. Yinchang Feng: Supervision, Writing – review and editing.

Competing interests. The authors declare no conflicts of interest relevant to this study.

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

Acknowledgements. This work was financially 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) and the robotic AI-Scientist platform of Chinese Academy of Sciences. This work was also supported by Tianjin Natural Science Foundation Project (grant no. 24JCYBJC01870)

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

References

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, 2024.

Becagli, S., Sferlazzo, D. M., Pace, G., di Sarra, A., Bommarito, C., Calzolai, 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, Atmospheric Chemistry and Physics, 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, Environmental Science & Technology, 57, 18 246–18 258, https://doi.org/10.1021/acs.est.3c05104, pMID: 37661931, 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, Atmospheric Chemistry and Physics, 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, 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, Science of The Total Environment, 518-519, 626–635, https://doi.org/https://doi.org/10.1016/j.scitotenv.2015.01.022, 2015.

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, Atmospheric Chemistry and Physics, 18, 11 793–11 812, 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, Environmental Science & Technology, 49, 1206–1214, https://doi.org/10.1021/es5037332, pMID: 25526283, 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, Environmental Science and Pollution Research, 29, 28 359–28 374, 2022.

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, 2021.

García-Laencina, P. J., Sancho-Gómez, J.-L., and Figueiras-Vidal, A. R.: Pattern classification with missing data: a review, Neural Computing and Applications, 19, 263–282, 2010.

Hao, H., Wang, Y., Zhu, Q., Zhang, H., Rosenberg, A., Schwartz, J., Amini, H., van Donkelaar, A., Martin, R., Liu, P., Weber, R., Russel, A., Yitshak-sade, M., Chang, H., and Shi, L.: National Cohort Study of Long-Term Exposure to PM2.5 Components and Mortality in Medicare American Older Adults, Environmental Science & Technology, 57, 6835–6843, https://doi.org/10.1021/acs.est.2c07064, pMID: 37074132, 2023.

Hopke, P. K.: A guide to positive matrix factorization, in: Workshop on UNMIX and PMF as Applied to PM2, vol. 5, p. 600, 2000.

Hopke, P. K.: Review of receptor modeling methods for source apportionment, Journal of the Air & Waste Management Association, 66, 237–259, https://doi.org/10.1080/10962247.2016.1140693, pMID: 26756961, 2016.

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

Hu, J., Zhang, H., Chen, S., Ying, Q., Wiedinmyer, C., Vandenberghe, F., and Kleeman, M. J.: Identifying PM2.5 and PM0.1 Sources for Epidemiological Studies in California, Environmental Science & Technology, 48, 4980–4990, https://doi.org/10.1021/es404810z, pMID: 24552458, 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 Resources Management, 36, 1159–1173, 2022.

Junger, W. and Ponce de Leon, A.: Imputation of missing data in time series for air pollutants, Atmospheric Environment, 102, 96–104, https://doi.org/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, Atmospheric Environment, 38, 2895–2907, https://doi.org/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, 2020.

Kim, E. and Hopke, P. K.: Comparison between sample-species specific uncertainties and estimated uncertainties for the source apportionment of the speciation trends network data, Atmospheric Environment, 41, 567–575, 2007.

Kim, E., Hopke, P. K., and Qin, Y.: Estimation of organic carbon blank values and error structures of the speciation trends network data for source apportionment, Journal of the Air & Waste Management Association, 55, 1190–1199, 2005.

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 PM2.5 constituents with estimates using machine learning better for source apportionment than exclusion or median replacement?, Environmental Pollution, 354, 124 165, https://doi.org/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 PM2.5 and further identifying spatial patterns in multi-sites without external data, Urban Climate, 62, 102 552, https://doi.org/https://doi.org/10.1016/j.uclim.2025.102552, 2025a.

Kim, Y., Kang, C., Yi, S. M., Heo, J. B., Kim, H., Lee, W., Kim, H., Hopke, P. K., Lee, Y. S., Shin, H. J., et al.: Imputing missing data with statistical-learning estimates: impacts on mortality risks attributable to area-and source-specific PM2. 5., Atmospheric Pollution Research, p. 102785, 2025b.

Kowarik, A. and Templ, M.: Imputation with the R Package VIM, Journal of Statistical Software, 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 Resources Management, 33, 2615–2628, 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 PM2.5 in South Korea using machine learning with ground observation data, Science of The Total Environment, 926, 171 884, https://doi.org/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 (PM2.5) chemical constituents using four machine learning models, Expert Systems with Applications, 221, 119 696, https://doi.org/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, Atmospheric Chemistry and Physics, 20, 12 047–12 061, https://doi.org/10.5194/acp-20-12047-2020, 2020.

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

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, Atmospheric Chemistry and Physics, 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, Environmental Science & Technology, 57, 5169–5179, https://doi.org/10.1021/acs.est.2c09412, pMID: 36940370, 2023.

Little, R. J. and Rubin, D. B.: Statistical analysis with missing data, John Wiley & Sons, 2019.

Liu, B., Wu, J., Zhang, J., Wang, L., Yang, J., Liang, D., Dai, Q., Bi, X., Feng, Y., Zhang, Y., et al.: Characterization and source apportionment of PM2. 5 based on error estimation from EPA PMF 5.0 model at a medium city in China, Environmental Pollution, 222, 10–22, 2017.

Liu, M. and Matsui, H.: Aerosol radiative forcings induced by substantial changes in anthropogenic emissions in China from 2008 to 2016, Atmospheric Chemistry and Physics, 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., et al.: Unraveling the process of aerosols secondary formation and removal based on cosmogenic beryllium-7 and beryllium-10, Science of The Total Environment, 821, 153 293, 2022.

McKenzie, E. R., Money, J. E., Green, P. G., and Young, T. M.: Metals associated with stormwater-relevant brake and tire samples, Science of The Total Environment, 407, 5855–5860, https://doi.org/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, The R Journal, 9, 207–218, https://doi.org/10.32614/RJ-2017-009, 2017.

Paatero, P.: The Multilinear Engine: A Table-Driven, Least Squares Program for Solving Multilinear Problems, including the n-Way Parallel Factor Analysis Model, Journal of Computational and Graphical Statistics, 8, 854–888, http://www.jstor.org/stable/1390831, 1999.

Paatero, P. and Hopke, P. K.: Discarding or downweighting high-noise variables in factor analytic models, Analytica Chimica Acta, 490, 277–289, https://doi.org/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, Atmospheric Measurement Techniques, 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 PM2. 5 in Pittsburgh using PMF and UNMIX, Aerosol science and technology, 40, 910–924, 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, Environmental Science & Technology 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, Atmospheric Environment, 40, 7316–7330, https://doi.org/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, Journal of Geophysical Research: Atmospheres, 103, 19 045–19 057, https://doi.org/https://doi.org/10.1029/98JD01212, 1998.

Reff, A., Eberly, S. I., and Bhave, P. V.: Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods, Journal of the Air & Waste Management Association, 57, 146–154, 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, Agricultural and Forest Meteorology, 147, 199–208, https://doi.org/https://doi.org/10.1016/j.agrformet.2007.06.004, 2007.

https://doi.org/10.5194/egusphere-2026-474
Preprint. Discussion started: 5 March 2026

Samal, K. K. R., Babu, K. S., and Das, S. K.: Multi-directional temporal convolutional artificial neural network for PM2.5 forecasting with missing values: A deep learning approach, Urban Climate, 36, 100 800, https://doi.org/https://doi.org/10.1016/j.uclim.2021.100800, 2021.

Shen, H., Li, T., Yuan, Q., and Zhang, L.: Estimating Regional Ground-Level PM2.5 Directly From Satellite Top-Of-Atmosphere Reflectance Using Deep Belief Networks, Journal of Geophysical Research: Atmospheres, 123, 13,875–13,886, https://doi.org/https://doi.org/10.1029/2018JD028759, 2018.

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.

Tian, S., Pan, Y., and Wang, Y.: Size-resolved source apportionment of particulate matter in urban Beijing during haze and non-haze episodes, Atmospheric Chemistry and Physics, 16, 1–19, 2016.

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, Environmental Science & Technology, 53, 2595–2611, https://doi.org/10.1021/acs.est.8b06392, pMID: 30698001, 2019.

Wang, Q., Qiao, L., Zhou, M., Zhu, S., Griffith, S., Li, L., and Yu, J. Z.: Source Apportionment of PM2.5 Using Hourly Measurements of Elemental Tracers and Major Constituents in an Urban Environment: Investigation of Time-Resolution Influence, Journal of Geophysical Research: Atmospheres, 123, 5284–5300, https://doi.org/https://doi.org/10.1029/2017JD027877, 2018.

Xie, J.: Deep Neural Network for PM2.5 Pollution Forecasting Based on Manifold Learning, in: 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), pp. 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 PM2.5 by comparing long- and short-term hourly observation-based modeling, Environmental Pollution, 314, 120 273, https://doi.org/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), pp. 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.: PM2.5 elements at an urban site in Yangtze River Delta, China: High time-resolved measurement and the application in source apportionment, Environmental Pollution, 253, 1089–1099, https://doi.org/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 PM2.5 composition and sources at an urban site in Yangtze River Delta, China after the implementation of the APPCAP, Chemosphere, 261, 127 746, https://doi.org/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, Environmental Science and Pollution Research, pp. 1–33, 2022.