Response to reviewer comments

Reviewer #3

This manuscript proposes a novel imputation framework (PMFr) that leverages source–receptor relationships derived from PMF to reconstruct missing $PM_{2.5}$ speciation data. Addressing missing data due to instrument failures and monitoring gaps is an important and practical issue, and the attempt to incorporate physically interpretable source profiles rather than relying solely on statistical covariance is a clear strength of this work.

However, several key aspects of the methodology require further elaboration. In particular, the role of the pre-imputation step, the justification for the assignment of uncertainties, and the comparison with standard PMF practices need to be addressed to fully demonstrate the robustness of the proposed approach. I recommend major revisions to clarify the workflow and to more rigorously validate the method before the manuscript can be considered for publication.

Major Comments:

Comment #1:

The methodological description in Section 2.3 would benefit from further clarification to address potential concerns about model independence. According to the text, tracer species are first imputed using another method, such as KNN, while non-tracers are filled using the geometric mean prior to the initial PMF run. Given that the PMFr framework relies on these pre-imputed values to derive source profiles and subsequently reconstruct missing data, it is currently difficult to isolate the performance of the PMFr method itself from the accuracy of the initial KNN imputation. To address this, the authors should clearly delineate the full workflow, including all intermediate steps. Providing a comprehensive flowchart in Figure 1 that details the full pipeline from raw data to pre-imputation, initial PMF, reconstruction, and the final PMF would greatly improve clarity. Additionally, conducting a sensitivity analysis to evaluate how different pre-imputation methods in pre-imputed tracers propagate into the final PMFr results is necessary to demonstrate the methodological robustness of the framework.

Response:

We sincerely thank the reviewer for this valuable suggestion. To address this concern, we revised Section 2.3 (Line 98-122) and updated Figure 1 to show the complete PMFr workflow, including raw data preprocessing, source identification (initial PMF run),

pre-imputation, PMF-based reconstruction using the $G \times F$ structure, and validation.
In addition, we added a sensitivity analysis of the first pre-imputation step in Text S5 and Table S14. Different algorithms, including LI, KNN, DBN, and BPCA, were used for the initial pre-imputation step, and the final PMFr reconstruction results were compared. The results show that PMFr remains relatively stable across different pre-imputation methods. For NH_4^+, the final PMFr R^2 values range from 0.92 to 0.96 and the MAPE values range from 20.67% to 24.58%. For NO_3^-, the final PMFr R^2 values range from 0.85 to 0.90, while the MAPE values range from 22.91% to 33.11%.
These results indicate that the final PMFr reconstruction is not simply determined by the initial KNN imputation. Although an initial estimate is required, PMFr refines the reconstructed values through PMF-resolved source profiles and source-receptor constraints.
Our revisions have been added in Section 2.3, Figure 1, Text S5, and Table S14.

1. Revised Section 2.3 text (Line 98-122):
"A tracer for imputation, hereafter referred to as a tracer, is defined as a key species that distinguishes a specific factor from others and reflects how that factor influences the receptor over time. Co-tracers refer to co-varying tracers within the same factor, collectively characterizing the temporal behavior of the corresponding source. As illustrated in Figure 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. Details of the usage of PMF for SA can be found in the literature, and the uncertainty settings are provided in Text S3. Based on the SA results with selected source profiles, species requiring imputation are classified as tracers or non-tracers through a knowledge-driven step.
When imputing tracers, the availability of co-tracers should be checked at each timestamp before reconstruction, because the source contribution vector *G* needs to be constrained by source-specific tracer information. If all tracers associated with a specific factor are simultaneously missing, the corresponding *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 Text S4.
The pre-imputed dataset and its associated uncertainty matrix are then input into the PMF model for reconstruction. The PMF run decomposes the dataset into factor profiles (*F*) and source contributions (*G*), and data reconstruction is achieved by multiplying

the G and 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, 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.

2. Revised Figure 1:

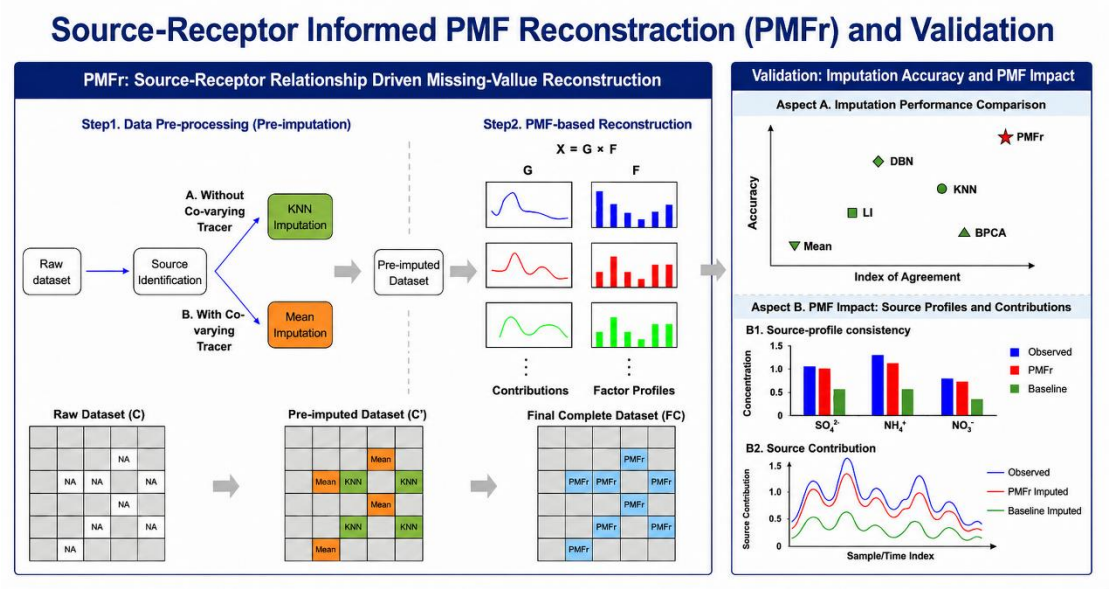

Figure 1. Flow chart of Source-Receptor Informed Positive Matrix Factorization Reconstruction (PMFr) and validation.

3. Added Text S5: Sensitivity Analysis of the First Pre-imputation Step

"Sensitivity analysis was conducted to quantify the impact of the pre-imputation step. As shown in Table S14, the final PMFr reconstruction metrics vary depending on the initial pre-imputation algorithm. For NH_4^+, utilizing KNN as the pre-imputation method yields an R^2 of 0.92, an IoA of 0.95, and a MAPE of 20.67%. DBN results in an R^2 of 0.95, an IoA of 0.92, and a MAPE of 23.61%. BPCA produces an R^2 of 0.96, an IoA of 0.81, and a MAPE of 24.58%. Across the tested algorithms for NH_4^+, the R^2 values range from 0.92 to 0.96, and the MAPE ranges from 20.67% to 24.58%. For NO_3^-, KNN achieves an R^2 of 0.85, an IoA of 0.95, and a MAPE of 23.92%. DBN yields an R^2 of 0.89, an IoA of 0.91, and a MAPE of 33.11%. BPCA results in an R^2 of 0.90, an IoA of 0.75, and a MAPE of 22.91%. For NO_3^-, the resulting R^2 values range from 0.85 to 0.90, while the MAPE spans from 22.91% to 33.11%. These results suggest that the PMFr maintains robust imputation performance regardless of the specific pre-imputation algorithm applied. Although the initial step is required, the PMFr method yields better performances than the baseline KNN approach. For NH_4^+ under MCMS, the PMFr MAPE is 20.67% versus the KNN MAPE of 24.00% at a 10% missing rate, and 25.87% versus 45.33% at a 20% missing rate. This improvement is achieved because subsequent PMF iterations impose source-profile constraints on the reconstructed values, such as maintaining a NO_3^-/NH_4^+ mass ratio of approximately 3 for the SN factor."

4. Added Table S14:

Table S14. Sensitivity analysis of different pre-imputation methods under the MCMS mechanism (Case 4, 10% missing rate)

Species	Pre-Imputation Method	R^2	IoA	MAPE(%)
NH_4^+	KNN	0.92	0.95	20.67
	LI	0.82	0.88	40.32
	DBN	0.95	0.92	23.61
	BPCA	0.96	0.81	24.58
NO_3^-	KNN	0.85	0.95	23.92
	LI	0.76	0.90	42.04
	DBN	0.89	0.91	33.11
	BPCA	0.90	0.75	22.91

Comment #2:

While the manuscript comprehensively compares PMFr against LI, KNN, BPCA, and DBN, it omits the most widely used baseline in receptor modeling practice. The U.S. EPA PMF 5.0 User Guide recommends handling missing values by replacing them with

the species median and assigning an uncertainty of four times the median (400%). As this approach is routinely used in real-world PMF applications, including it as a baseline would help the receptor modeling community better assess the meaningful improvement provided by PMFr. The authors are encouraged to include this EPA-recommended method as a baseline and compare the PMFr performance against it under the various missing-data scenarios.

Response:

We thank the reviewer for this valuable suggestion from the perspective of source apportionment practice. The U.S. EPA PMF 5.0 User Guide recommends replacing missing values with the species median, while foundational PMF studies have also used geometric mean substitution; both approaches follow the same principle of replacing missing values with a robust central value and assigning a high uncertainty to reduce the influence of outliers and limit the impact of substituted values on model fitting (Norris et al., 2014; Polissar et al., 1994). In this study, geometric mean imputation was used as a representative conventional PMF missing-value treatment because the geometric mean is also a robust central value for skewed data and was adopted in our previous PMF analysis using the same hourly $PM_{2.5}$ speciation dataset. To make this rationale clearer, we have added an explicit explanation in **Section 2.3 (Line 121-126)**. Our quantitative evaluation shows that geometric mean imputation yields an overall MAPE of 66.75%. Because this approach replaces missing data with a constant value, it exhibits no temporal variation. This inherent structural limitation prevents it from capturing the dynamic fluctuations of pollutant concentrations, resulting in high absolute errors across the individual missing-data scenarios, as shown in Figures S11--S29. We have added **Section 3.2.1 (169-178)** to include the quantitative comparison for the geometric mean imputation method. Furthermore, we added a rationale stating that, given its lack of temporal variation, as shown across all individual scenarios in Figures S11--S29, and its high overall MAPE, its performance is quantified solely in this overall assessment and is excluded from further detailed trend comparisons in subsequent sections. In addition, the conventional geometric mean imputation method was included in the subsequent PMF analysis to evaluate its influence on PMF-resolved source profiles and source contributions (Lines 254-284).

The revised text in Section 2.3 and Section 3.2.1 is as follows:

1. Revised text in Section 2.3 (Line 121-126):

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, it was used here as a representative conventional PMF missing-value treatment for comparison with PMFr.

2. Revised text in Section 3.2.1 (Line 169-178)

"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 (Figures S11-S29), its performance is solely quantified by MAPE here and is excluded from further detailed comparisons in subsequent sections."

Reference:

Norris, G., Duvall, R., Brown, S., & Bai, S. (2014). *EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide* (EPA/600/R-14/108). U.S. Environmental Protection Agency, Washington, DC.

Polissar, A. V., Hopke, P. K., Paatero, P., Malm, W. C., & Sisler, J. F. (1998). Atmospheric aerosol over Alaska: 2. Elemental composition and sources. *Journal of Geophysical Research: Atmospheres*, 103(D15), 19045-19057. https://doi.org/10.1029/98JD01212

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, Environmental Pollution, 314, 120273, https://doi.org/10.1016/j.envpol.2022.120273, 2022.

Comment #3:

The treatment of uncertainty in Section 2.3 requires further justification. The manuscript states that tracers are assigned an uncertainty equal to 10% of their imputed value, while non-tracers are assigned an uncertainty equal to eight times the geometric mean. The physical or statistical rationale for these specific multipliers is currently missing. Since the uncertainty matrix directly controls the PMF objective function (Q-value) and strongly influences the model solution, these parameters are critical. The authors should provide a justification for these choices, whether through literature references, empirical evidence, or theoretical reasoning, and briefly discuss or conduct a sensitivity analysis demonstrating how different uncertainty assignments might affect the PMFr performance and subsequent PMF outputs.

Response:

To address this concern, we added Text S4, "Discussion on Data Treatment of PMFr

Uncertainty", to explain the rationale for assigning different uncertainties to pre-imputed tracers, geometrically filled tracers, and non-tracers.

For missing tracers without available co-tracers, the corresponding G vector is less directly constrained by observed species. Therefore, these missing tracer values are first estimated using another imputation method, with KNN recommended for its simplicity, efficiency, and ability to provide a reasonable estimate of temporal variation. Their uncertainty is set to 10% of the imputed concentration so that these pre-imputed tracer values retain sufficient statistical weight in the PMF calculation and can provide source-specific temporal information for constraining G, rather than being effectively ignored during factorization. This setting is supported by previous PMF analysis using the same observation site and hourly $PM_{2.5}$ speciation dataset (Xie et al., 2022).

For missing tracers with available co-tracers and for non-tracers, the imputed values are not intended to provide the primary temporal constraint because available co-tracers or other observed species already provide stronger information for resolving G. Therefore, these values are assigned a much larger uncertainty, defined as eight times the geometric mean. Standard receptor-modeling practice commonly assigns missing data an uncertainty of four times the geometric mean concentration (Polissar et al., 1994). In this study, we further tested larger multipliers and selected eight times the geometric mean because it provided an appropriate balance: it sufficiently downweighted these geometrically filled values while maintaining stable PMF reconstruction. This treatment minimizes the influence of potentially biased geometric-mean substitutions on the Q-value objective function and ensures that the PMF solution is primarily driven by reliable observed species and available source-related tracers.

Text S4. Discussion on Data Treatment of PMFr Uncertainty

"The uncertainty matrix directly determines the statistical weight of individual data points in the Positive Matrix Factorization (PMF) objective function (Q). In PMFr, uncertainty assignment is designed according to the role of each filled value in constraining the source contribution matrix (G). Specifically, when imputing tracers, the availability of co-tracers should first be checked at each timestamp because G needs to be constrained by source-specific tracer information.

If all tracers associated with a specific factor are simultaneously missing, the corresponding G vector is less directly constrained by observed species. In such cases, the missing tracer values are first estimated 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. This uncertainty setting allows the pre-imputed tracer values to retain sufficient statistical weight in the PMF calculation, so that they can provide source-specific temporal information for constraining G, rather than being effectively

ignored during factorization. This setting is supported by previous PMF analysis using the same observation site hourly $PM_{2.5}$ speciation dataset. For missing tracers with available co-tracers, as well as for non-tracers, missing values are replaced by the species-specific geometric mean. In these cases, the filled values are not intended to provide the primary temporal constraint for the corresponding source factor, because available co-tracers or other observed species already provide stronger information for resolving G. Therefore, these filled values are assigned a much larger uncertainty, defined as eight times the geometric mean. Standard receptor-modeling practice commonly assigns missing data an uncertainty of four times the median or geometric mean concentration. Here, a larger multiplier was adopted to more strongly downweight these filled values. This treatment minimizes the influence of potentially biased geometric mean substitutions on the Q -value objective function and ensures that the PMF solution is primarily driven by reliable observed species and available source-related tracers."

Reference:

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, Environmental Pollution, 314, 120273, https://doi.org/10.1016/j.envpol.2022.120273, 2022.

Polissar, A. V., Hopke, P. K., Paatero, P., Malm, W. C., & Sisler, J. F. (1998). Atmospheric aerosol over Alaska: 2. Elemental composition and sources. *Journal of Geophysical Research: Atmospheres*, 103(D15), 19045-19057. https://doi.org/10.1029/98JD01212