Response to reviewer comments

Reviewer #1

Missing data in $PM_{2.5}$ speciation monitoring, due to instrumental drift, calibration, and maintenance, poses challenges for source apportionment and health risk assessments. Conventional imputation methods, including statistical techniques and deep learning, depend on mathematical correlations and often lack physical interpretability. This study presents a novel Positive Matrix Factorization-based reconstruction method (PMFr) that integrates source profile characteristics into the imputation process. Unlike traditional models that rely solely on data covariance, this approach uses "low-entropy structures" to reconstruct latent information, ensuring chemical consistency and physical interpretability. Given its potential to improve data quality in atmospheric research, the reviewer recommends this work for publication with some revisions and clarifications.

Major Comments:

Comment #1:

The manuscript introduces a novel framework for data imputation based on low-entropy structures, but lacks practical guidelines on its limits of applicability for specific timestamps. It does not define conditions under which the method may fail due to insufficient observational constraints. The methodology assumes the source contribution vector (G) can be uniquely resolved from observed species, which requires at least one key tracer species for each source factor. However, the manuscript does not address scenarios where all characteristic species for a specific source are missing, leading to an under-constrained system that undermines the imputation's reliability.

The authors should include a section on practical principles for validity checks. It must state that before imputation, users should ensure each identified source factor has at least one non-missing key tracer. If any time point lacks all diagnostic tracers for a source, that data point should be flagged as un-imputable or handled with caution.

To operationalize this principle, the authors should add a table listing the "Non-Missable Key Tracers" for each "pollution source". This table should clearly map each source factor to its essential diagnostic species. This will serve as a vital reference for practitioners to assess data quality and imputation feasibility before applying the model.

Addressing these points is essential to prevent the misapplication of the method and to clarify the boundary conditions under which the proposed imputation remains scientifically valid.

Response:

We sincerely thank the reviewer for this important comment. We agree that the availability of key tracers is critical for PMFr, because the source contribution vector (G) needs to be sufficiently constrained by source-specific chemical information. When all diagnostic tracers for a specific source factor are simultaneously missing at a given timestamp, the corresponding G vector becomes less constrained by observations, and the reliability of the reconstruction should be carefully evaluated.

We would like to clarify that this situation is one of the motivations for including the pre-imputation step in PMFr. When all tracers associated with a specific factor are missing, PMFr first uses a pre-imputation method to provide an initial estimate for the missing tracer values, allowing the subsequent PMF run to proceed. In this study, KNN is recommended for this initial step because of its simplicity, efficiency, and ability to provide a reasonable initial estimate of temporal variation. The subsequent PMFr reconstruction then further constrains the imputed values using PMF-resolved source profiles and source-receptor relationships. Therefore, such cases are not automatically treated as unusable; rather, they are considered cases with weakened source constraints and should be flagged and handled with caution. Our sensitivity analysis (Text S5 and Table S14) further shows 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.

To address the reviewer's concern and to prevent misapplication of the method, we have revised the manuscript as detailed below:

1. **Listing Non-Missable Key Tracers (Table S13):** We have added **Table S13** in the Supplementary Information, which maps each PMF-resolved pollution source in this study to its essential diagnostic species. This table provides a practical reference for users to evaluate whether each source factor retains sufficient observational constraints before applying PMFr. When PMFr is applied to other datasets or to other species requiring imputation, the non-missable key tracers should be checked based on the resolved source profiles and prior knowledge of local source characteristics.

Table S13. Pollution sources and corresponding non-missable key tracers

Factor Identity	Non-Missable Key Tracers
Secondary Nitrate	NH_4^+, NO_3^-
Secondary Sulfate	NH_4^+, SO_4^{2-}
On-road Traffic	OC, EC, Ba, Cu
Coal Combustion	OC, EC, As, Se, K, Pb
Metal Smelting	Mn, Pb, Cr, Ni, Zn
Heavy Oil Combustion	V, Ni
Crustal Dust	K, Fe, Ca, Si, Ti

2. **Adding a validity-check procedure in Section 2.3 (Line 106-111):** We have revised the Methods section to explicitly state that, before PMFr reconstruction, the availability of key tracers or co-tracers should be checked at each timestamp to identify cases with weakened source constraints. This check helps users determine whether the source contribution vector (G) is sufficiently constrained by observed species or whether initial estimation (e.g., KNN) is required for the PMF reconstruction.

"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."

3. **Expanding the discussion of applicability and limitations in Section 3.4 (Line 292-306):** We have added a dedicated discussion in the revised Section 3.4, "Applicability and Limitations of PMFr", to clarify the significance of validity checks. In this section, we state that PMFr is most reliable when at least one key tracer remains available for each source factor. If all key tracers for a specific source are simultaneously missing, the corresponding G vector is less directly constrained by observed species and the resulting reconstruction should be interpreted with caution. We also clarify that the pre-imputation step is designed to provide the initial estimate needed for PMFr reconstruction under such weakened source-constraint conditions.

“PMFr is applicable when source-related chemical structures can be constrained by at least one tracer for each source factor, and the completed dataset is suitable for 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 NH_4^+ 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 NO_3^- 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 NH_4^+, and from 22.81% to 28.46% for NO_3^-. 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 (Text S5 and Table S14).”

Comments #2:

Mixed missing data patterns (MCMS vs. MCMI) in Cases 4~8. MCMS is inherently much more challenging than MCMI because it removes the identifiability of the source, whereas MCMI only removes temporal continuity. A model might perform well under MCMI but fail catastrophically under MCMS. Combining these two patterns into a single performance metric for each Case obscures the specific source of error. Therefore, the reviewer suggests that reporting the results for pure MCMS scenarios and pure MCMI scenarios separately is more scientifically valid.

Response:

We appreciate the reviewer’s rigorous comment. The reviewer is correct from a theoretical standpoint: MCMS and MCMI represent fundamentally different mechanisms of information loss, and isolating them is highly valuable for understanding specific algorithmic vulnerabilities.

Regarding the main text, our primary goal in proposing the PMFr framework is to provide a robust, practical imputation tool tailored for realistic, operational datasets. In actual continuous monitoring stations, complex instrument malfunctions frequently result in a simultaneous occurrence of both MCMS and MCMI. Therefore, Cases 4-8 in the main manuscript were designed to evaluate whether PMFr can maintain its reconstruction stability when both mechanisms occur simultaneously, which is precisely the challenge faced in practice.

However, we fully agree with the reviewer that reporting the isolated scenarios provides essential scientific insights. To address this concern without diluting the real-world focus of the main manuscript, we have now provided the evaluations in the **Section 3.4 Applicability and Limitations of PMFr—**the **most challenging** MCMS scenario, in which all key tracers are missing simultaneously, and the corresponding MCMI scenario. Specifically, the detailed performance metrics for all original mixed cases are already comprehensively listed in **Supplementary Tables S7-S12**.

The added discussion comparing MCMS and MCMI reads as follows (Line 293-306):

"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 NH_4^+ 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 NO_3^- 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 NH_4^+, and from 22.81% to 28.46% for NO_3^-. 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 (Text S5 and Table S14)."

Comments #3:

The PMFr method relies on the assumption that source chemical profiles remain stable over time. However, real-world atmospheric conditions lead to dynamic source

signatures that can vary significantly due to seasonal changes, fuel composition, and combustion conditions. This variability can introduce biases in reconstructed data if profiles differ from reality. Though this study uses a short two-month dataset, concerns about using this method over longer periods (e.g., multi-year datasets) highlight issues with profile stability. The manuscript currently lacks guidance on determining the appropriate temporal window for stable profiles. The reviewer advises the authors to provide clear, quantitative guidelines for assessing this assumption, including metrics or statistical tests (like rolling window analysis or change-point detection) to identify when profiles need recalibration or updating.

Response:

We thank the reviewer for this thoughtful comment. Indeed, PMFr does rely on the assumption that source chemical profiles remain sufficiently stable within the reconstruction period. If source profiles change substantially over time, the source-receptor relationships resolved by PMF may no longer represent the true receptor-based pollution sources, which could introduce biases into the reconstructed data. This is a key limitation of PMFr.

In the present study, we intentionally used a relatively short two-month dataset to reduce the potential influence of long-term source-profile changes. Within such a short period, source profiles are more likely to remain stable, making the PMF-resolved source-receptor relationships suitable for missing-value reconstruction. As mentioned by the reviewer, source-profile changes may occur under various real-world conditions, such as changes in fuel composition, implementation or removal of end-of-pipe control technologies, changes in industrial production processes, shutdown or relocation of major emission sources, changes in source suppliers, or strong seasonal shifts in atmospheric processing. Therefore, the appropriate temporal window for applying PMFr should not be fixed universally, but should be determined according to the stability of source profiles in the specific dataset and the local emission context.

To address this concern, we have revised the manuscript by adding a dedicated discussion in the Section 3.4 "**Applicability and Limitations of PMFr**" section (Line 319-328). In this section, we now explicitly state that source-profile stability should be evaluated before applying PMFr to extended datasets. We further discuss that rolling PMF approaches can be used to examine temporal changes in source profiles and to determine whether the PMF solution remains stable over time. Because real-world source-profile changes are often not known a priori, we believe that such diagnostic checks are essential for long-term applications. When substantial changes in source profiles are detected, the dataset should be divided into shorter time windows, or the PMF model should be recalibrated before applying PMFr. Future improvements of PMFr could also incorporate time-dependent source profiles to better support

reconstruction under changing atmospheric conditions.

The corresponding revision has been added to Section 3.4, "Applicability and Limitations of PMFr", as follows (Line 319-328):

"Another limitation of the PMFr framework 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 may change substantially. Therefore, source-profile stability can be evaluated before applying PMFr in extended applications. As for long-term data applications, moving-window evolving PMF approaches provide a promising way to track time-dependent factor profiles within short moving windows. Improvements of PMFr could incorporate time-dependent source profiles to address this limitation and better support reconstruction under changing source emissions."

Comments #4:

The PMFr framework relies on a linear mixing model ($C=G\times F$), assuming observed concentrations are linear combinations of primary emissions. However, secondary components like sulfates, nitrates, and Secondary Organic Carbon (SOC) arise from complex, non-linear photochemical reactions, which the linear assumption may fail to accurately capture, particularly during heavy pollution or specific weather conditions. The manuscript does not sufficiently address the uncertainty introduced by this assumption in reconstructing secondary species. It is recommended that the authors discuss the limitations of the linear model in secondary aerosol formation and consider conducting a sensitivity analysis to quantify the uncertainty.

Response:

We thank the reviewer for raising this important point. The linear mixing model used in PMFr cannot explicitly quantify or simulate the nonlinear photochemical reactions responsible for the formation of secondary components. However, PMFr is not intended to model the chemical reaction pathways of secondary aerosol formation. Instead, similar to conventional PMF receptor modeling, PMFr uses the observed chemical dataset to quantify the amount of secondary particles whose temporal variations differ from those of primary sources. These different temporal patterns allow secondary components to be identified as separate PMF factors. Therefore, PMFr uses the PMF-resolved secondary particles to quantify the abundance of already formed secondary components rather than simulating their nonlinear formation pathways. In this way, PMFr can be used for imputing missing secondary species without explicitly parameterizing the complex and nonlinear reactions that produced them.

Minor Comments:

Comments #1:

Line 82. MCMS and MCMI should be defined in the first paragraph of section 2.2.

Response:

Revised as suggested.

Comments #2:

Figure 1b illustrates model performance metrics through a scatter plot comparing MAPE (y-axis) and IoA (x-axis), with R^2 values annotated. However, it does not visualize the standard deviation (σ) of modeled data against observations. A model may show high IOA and low MAPE but still misrepresent variability, indicating "amplitude bias," which is crucial for accurate source contribution estimates. The authors should include a Taylor Diagram as a supplementary figure for a comprehensive statistical assessment of variance and correlation in the observed data.

Response:

We have generated a Taylor diagram (**Figure S31**) as a supplementary figure to visualize the variance and correlation. The diagram indicates that the PMFr reconstructed data yield a normalized standard deviation of 0.93, compared to the observational reference ($\sigma = 1.0$).

To systematically present this comprehensive statistical assessment, we have optimized the manuscript structure by introducing a new section, “Section 3.2.1 Overall Performance under All Missing Scenarios”. The quantitative evaluation of the Taylor diagram and the corresponding variance assessment have been fully integrated into this new section.

The revised text in Section 3.2.1 is as follows (Line 170-178):

“As shown in Figure 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 Mean imputation method 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. Furthermore, the Taylor diagram (Figure S31) illustrates that the

PMFr reconstructed data yield a normalized standard deviation (σ) of 0.93, closely matching the observational variance (σ = 1.0), suggesting its capability to capture the amplitude of data variations."

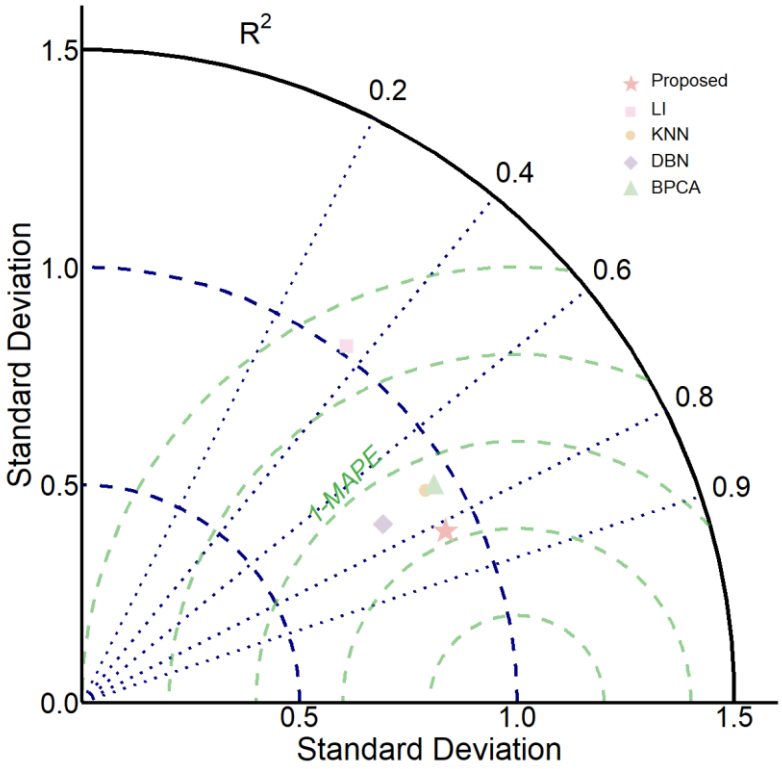

Figure S31. Taylor diagram summarizing the statistical performance of the proposed and baseline imputation models.