Rebuttal to the concerns raised by the reviewers for the paper titled "A Framework for Dynamic Hyper-local Source Apportionment using Low-cost Sensors for Real-time Policy Action"

(MS No.: egusphere-2025-5677)

Concerns of Reviewer 1:

This manuscript proposes a machine learning framework for real-time, hyper-local air pollution source apportionment (SA) using low-cost air quality (LCAQ) sensors co-located with Reference Grade Instruments (RGI). Multi-output linear regression models are trained on calibrated LCAQ data, with SA outputs from Positive Matrix Factorization (PMF) applied to HR-ToF-AMS and Xact-625i data serving as ground truth. The framework is evaluated at two sites in Lucknow, India, during a single month (October 2023). While the research topic is timely and relevant to the scope of Atmospheric Measurement Techniques, the manuscript has fundamental deficiencies in experimental scope, methodological rigor, and scientific integrity of its claims that preclude publication in its current form.

Concern #1: Overstated Claims of Robustness and Generalization:

The introduction asserts that the objective of this study is for "enhancing model robustness, calibration fidelity, and generalizability across diverse sensing environments" (p. 4, l. 115) and that "extensive experimentation has been carried out to validate the robustness of the proposed framework " (p. 5, l. 133). However, the entire study consists of two deployments within a single city during a single month. Each site yields approximately 350 observations at 30-minute resolution (roughly one week of data) (p. 15, l. 332-333). With an 80/20 train-test split, ***the test set contains only ~70 observations per site — less than two days of data****. These sample sizes are wholly insufficient to support claims of robustness or generalizations. To substantiate such claims, the authors would need to collect data across multiple seasons at the same sites, across multiple cities, or across multiple years. The current field monitoring scope supports only a proof-of-concept demonstration to conduct a full investigation, and* ***not sufficient for a manuscript proposing a new framework****.*

Response: We sincerely thank the reviewer for this important observation. We agree that the claims regarding **robustness** and **generalizability** in the original manuscript were stronger than what could be fully supported by the experimental evidence initially presented. The reviewer is correct that the original wording—particularly phrases such as *"enhancing model robustness and generalizability across diverse sensing environments"* and *"extensive experimentation validating robustness"*—may overstate the scope of validation performed in the submitted version.

In response to this concern, we **performed substantial additional experiments** and **will substantially revise the presentation and scientific framing of the manuscript**.

First, we will carefully tone down all overstated claims throughout the manuscript. The revised manuscript will no longer present the proposed framework as a fully generalized operational

solution deployable across arbitrary environments. Instead, we will explicitly position this work as a **proof-of-concept framework for near-real-time source apportionment using low-cost air quality sensors**, demonstrating promising but still preliminary evidence of transferability across spatial and temporal domains.

Second, to address the reviewer's concern regarding limited spatial and temporal diversity, we performed **five transfer-validation experiments**. These experiments were designed specifically to test model transferability across:

- **Different sites within Lucknow** (cross-site transfer)
- **Different seasonal periods** (cross-season transfer)
- **Different cities** (cross-city transfer)

The five transfer experiments considered in the revised study are:

1. **Lucknow (Train) → Kanpur (Test)** (cross-city validation)
2. **CSIR (Train) → BBAU (Test)** (cross-site validation)
3. **BBAU (Train) → CSIR (Test)** (cross-site validation)
4. **October (Train) → February–March (Test)** (cross-season validation)
5. **February (Train)–March → October (Test)** (cross-season validation)

The deployment site **BBAU is 'Site-B'** in original manuscript, while **CSIR is 'Site-C'** in original manuscript. These experiments provide a significantly more rigorous assessment of transferability compared to the originally submitted version.

For **elemental source apportionment**, the best-performing model (Extra Trees regressor) achieved mean Spearman rank correlations (SROCC) between **0.73 and 0.83**, with mean NDCG values between **0.94 and 0.97** across all five transfer experiments, indicating consistently strong preservation of source ranking even under cross-domain shifts.

For **organic source apportionment**, we further revisited the PMF representation and observed that the original 5-factor decomposition led to weak predictive performance for several highly overlapping organic factors. To address this, we merged the original five organic PMF factors into three broader and physically more meaningful source classes: **BBOA (BBOA-1 + BBOA-2), HOA, and OOA (LVOOA + SVOOA)**. Using this revised representation, the Extra Trees model achieved SROCC values between **0.74 and 0.90**, with mean NDCG values between **0.98 and 0.99** across all five transfer experiments, indicating strong capability for identifying dominant organic source contributions under varying environmental conditions.

Third, we also revised the machine learning evaluation protocol to follow a stricter **three-way temporal data separation strategy**. Instead of directly using a train-test split for model selection, the training-domain data was split into:

- **80% training set**
- **20% validation set** (temporally held out, with no shuffling)

while the target domain was used as a **fully independent external test set**. This ensures that model selection and hyperparameter tuning are performed without any test-set leakage.

At the same time, we acknowledge the reviewer's broader point that even with these additional experiments, the dataset still spans a limited number of field campaigns and cannot fully capture long-term atmospheric variability (e.g., year-to-year changes or broader inter-city diversity). Therefore, in the revised manuscript we will explicitly state that the proposed framework should currently be interpreted as a **promising proof-of-concept with preliminary evidence of transferability**, rather than a universally generalized framework ready for unrestricted deployment.

We thank the reviewer for this valuable comment, which significantly improved both the rigor of our experimental evaluation and the scientific positioning of the manuscript. The following table shows the revised list of experiments that have been performed and will be included in the revised manuscript.

Experiment	Training Domain	Validation Domain	Independent Test Domain	Purpose
E1	Lucknow (Oct 2023 + Feb–Mar 2024)	20% of Lucknow	Kanpur (Dec 2024)	Cross-city + long-term temporal transfer
E2	CSIR	20% of CSIR	BBAU	Cross-site transfer
E3	BBAU	20% of BBAU	CSIR	Cross-site transfer
E4	October	20% of October	February–March	Cross-season transfer
E5	February–March	20% of February–March	October	Cross-season transfer

Changes to be made in the revised manuscript:

- Revise **Abstract** to remove overstated claims of robustness/generalization.
- Revise **Introduction (Section 1)** to frame the study as a **proof-of-concept.**
- Add results from **five new transfer experiments** (cross-city, cross-site and cross-season).
- Add discussion on **limitations of dataset size and transferability scope** in Discussion/Conclusion.

Concern #2: Micro-Aethalometer Calibration Not Described:

The Micro-Aethalometer (AethLabs AE-51) is deployed alongside the LCAQ sensor unit and its BC measurements are used as a key predictor. The paper states only that the device "comes lab calibrated" (p. 2, l. 176) and cites a field intercomparison study, but provides no site-specific calibration or cross-validation against the EBAM or other RGI at either deployment site. Amazingly, the abstract of that field intercomparison study reads, "Real-world quality assurance of these instruments should be performed through field IC against reference instruments with longer durations in areas of slowly changing eBC concentration" (Alas et al., 2020). ***For a manuscript focused on calibration methodology, the absence of any field calibration assessment of this instrument is a significant omission that should be rectified.***

Response: We agree with the reviewer that the manufacturer's laboratory calibration alone does not provide a complete assessment of the Micro-Aethalometer performance under site-specific ambient conditions. We have therefore revised Appendix to include additional information on the operation, quality-control procedures, and measurement uncertainty of the AE-51.

A research-grade black carbon instrument (e.g., AE-33) was not available during the deployments. Because the AE-51 was operated as part of a mobile monitoring platform, the limited space inside the mobile laboratory restricted the deployment of an additional research-grade Aethalometer. Therefore, a direct site-specific calibration or intercomparison of the AE-51 against an AE-33 could not be performed. We have now explicitly acknowledged this limitation in the revised manuscript. However, Vernooij et al. (2022) reported a high correlation (R2 = 0.95) for BC measured by MicroAeth AE-51 and AE-33 at the same wavelength during biomass burning period. Other studies have also compared the AE-51 with ground-based instruments such as the Multi-Angle Absorption Photometer (MAAP), Alas et al. (2019) reported a correlation of $R^2 = 0.95$, while Pikridas et al. (2019) observed an R2 = 0.76 in city pollution.

In response to the reviewer's suggestion, we have also compared the averaged AE-51 BC concentrations with the collocated EBAM PM2.5 measurements. The measurements showed Pearson correlation coefficients of R = 0.91 in CSIR-CIMAP and R = 0.84 in BBAU at Cycle-1, and R = 0.85 in CSIR-CIMAP and R = 0.87 in BBAU at Cycle-2 (Figure R1.1). However, in Kanpur, EBAM was not operational due to a hardware issue. Since, in Lucknow, it showed good correlation, we assume it was working well.

Also, several operational measures were adopted to reduce measurement uncertainty. The AE-51 was regularly checked for flow-related warnings, abnormal optical signals, and excessive filter attenuation. The filter ticket was replaced every 4–5 h to avoid excessive aerosol loading and was maintained at an attenuation below 100, as increased filter loading can lead to underestimation of black carbon concentrations. The instrument was placed on a stable surface inside the mobile laboratory to reduce vibration-related noise, and sampling was conducted through a silica-gel diffusion dryer to minimize the influence of relative humidity. Although the AE-51 provided measurements at 1-min resolution, the data were averaged to 30-min intervals before analysis to reduce short-term instrumental noise and occasional negative values.

In addition, the previous study by our group has estimated the uncertainty in the AE-51 BC mass concentration by propagating uncertainties of 34.8%, associated with the filter spot area, sample flow rate, optical reference signal, and sensing signal (Lakra et al., 2024). We will also state in the revised manuscript that future mobile deployments should include a longer-duration field intercomparison with an AE-33 or another research-grade absorption instrument, wherever operational and space constraints permit, consistent with the recommendation of Alas et al. (2020).

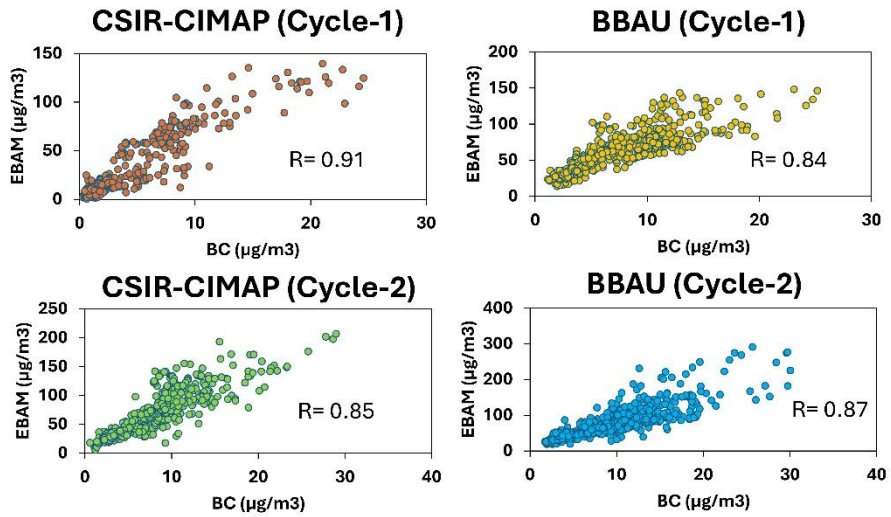


Figure R1.1: Scatter plot between EBAM and BC at both sites during Cycle-1 and Cycle-2. Where R represents the Pearson coefficient.

Changes to be made in the revised manuscript and Appendix:

- Revised in section-2 and in the Appendix of the manuscript.

References:

Vernooij, R., Winiger, P., Wooster, M., Strydom, T., Poulain, L., Dusek, U., Grosvenor, M., Roberts, G.J., Schutgens, N., Van Der Werf, G.R., 2022. A quadcopter unmanned aerial system (UAS)-based methodology for measuring biomass burning emission factors. Atmos. Meas. Tech. 15, 4271–4294. https://doi.org/10.5194/amt-15-4271-2022

Alas, H.D.C., Weinhold, K., Costabile, F., Di Ianni, A., Müller, T., Pfeifer, S., Di Liberto, L., Turner, J.R., Wiedensohler, A., 2019. Methodology for high-quality mobile measurement with focus on black carbon and particle mass concentrations. Atmos. Meas. Tech. 12, 4697–4712. https://doi.org/10.5194/amt-12-4697-2019

Pikridas, M., Bezantakos, S., Močnik, G., Keleshis, C., Brechtel, F., Stavroulas, I., Demetriades, G., Antoniou, P., Vouterakos, P., Argyrides, M., Liakakou, E., Drinovec, L., Marinou, E., Amiridis, V., Vrekoussis, M., Mihalopoulos, N., Sciare, J., 2019. On-flight intercomparison of three miniature aerosol absorption sensors using unmanned aerial systems (UASs). Atmos. Meas. Tech. 12, 6425–6447. https://doi.org/10.5194/amt-12-6425-2019

Lakra, A., Shukla, A.K., Bhowmik, H.S., Yadav, A.K., Jain, V., Murari, V., Gaddamidi, S., Lalchandani, V., Tripathi, S.N., 2024. Comparative analysis of winter composite-PM2.5 in Central Indo Gangetic Plain cities: Combined organic and inorganic source apportionment and characterization, with a focus on the photochemical age effect on secondary organic aerosol formation. Atmos. Environ. 338, 120827. https://doi.org/10.1016/j.atmosenv.2024.120827

Concern #3: Potential NO_2 Sensor Degradation Mid-Deployment:

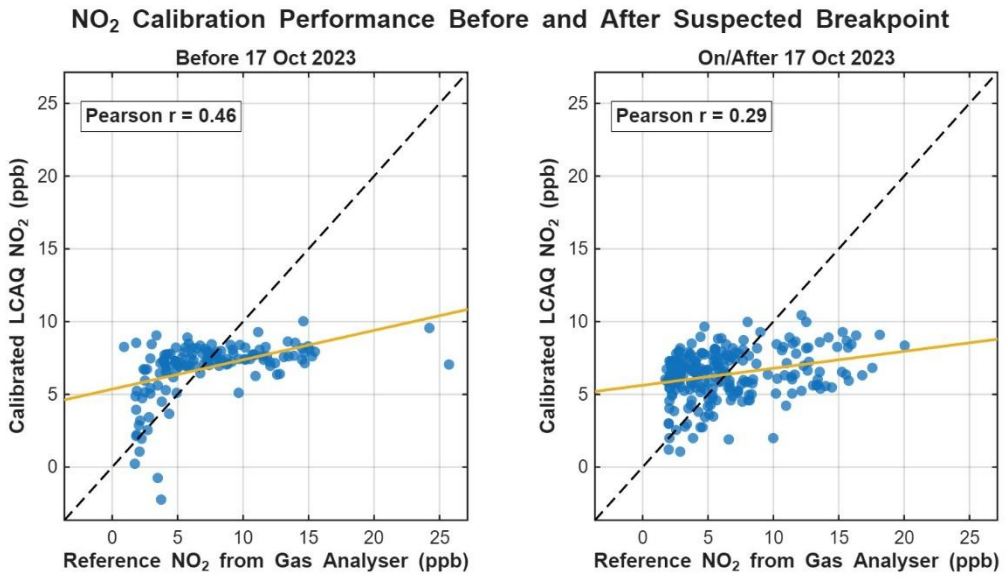
Figure 3 displays the NO_2 calibration time series at Site-B and visually suggests a ***step-change in sensor behavior*** *around 16–17 October 2023, consistent with* ***sensor performance degradation****. The authors must formally test this by computing the calibration correlation coefficient separately for data before and after this break-point. If a statistically significant change is identified, NO_2 data collected after that date should be excluded from training and inference. Given the already limited dataset size (~350 observations), such an exclusion could substantially constrain the usable training data and reduce the available inputs to three reliably calibrated pollutants (CO, O_3, and $PM_{2.5}$), given the acknowledged below-detection-limit issues with SO_2.* ***The implications for model validity need to be fully addressed, making the underlying data in its current form unusable.***

Response: We thank the reviewer for this careful observation regarding the possible degradation of the NO_2 sensor at **Site-B (BBAU)** around 16–17 October 2023.

In response to this concern, we performed the suggested split analysis by separately evaluating the calibration performance before and after the suspected breakpoint (taken as **17 October 2023, 00:00 hrs**). We quantified calibration quality by computing the Pearson correlation coefficient between the **reference NO_2 concentration measured by the gas analyser** and the **calibrated NO_2 concentration predicted by the LCAQ sensor unit**.

The correlation analysis revealed:

- **Before 17 Oct 2023:** Pearson correlation = **0.46**
- **On/After 17 Oct 2023:** Pearson correlation = **0.29**



Thus, the calibration correlation decreases by approximately **37%** after the suspected breakpoint, indicating a clear deterioration in NO_2 calibration performance during the latter part of the deployment.

While this degradation may arise from sensor drift, environmental interference, or reduced calibration validity under changing atmospheric conditions, the result confirms the reviewer's concern that the NO_2 predictor becomes increasingly unreliable after mid-deployment.

We agree that retaining such a weak predictor may adversely affect the machine learning framework for Source Apportionment.

To further quantify this impact, we performed an additional **feature ablation study** using the best-performing regression model (Extra Trees) under the most stringent transfer-learning setting (Lucknow (Train) → Kanpur (Test)). The mean Spearman rank correlation (SROCC) for elemental source apportionment changed as follows:

- Full predictor set: **0.8288**
- Without SO_2: **0.8291**
- Without SO_2 and NO_2: **0.8303**
- Without SO_2, NO_2, and VOC: **0.8349**

These results demonstrate that removing NO_2 does **not degrade model performance**; instead, it leads to a small but consistent improvement in generalization performance.

Based on both the calibration split analysis and the feature ablation study, we agree with the reviewer that retaining NO_2 in the final model is not justified. Therefore, in the revised manuscript, we will exclude **NO_2**, along with **SO_2** (which remained below the sensor detection limit) and **VOC** (which showed negligible feature importance in the organic SA model), from the final elemental SA framework.

The revised predictor set will therefore include only the more reliable predictors:

- CO
- O_3
- $PM_{2.5}$
- BC
- Temperature
- Relative Humidity
- Hour-of-day cyclical features

We thank the reviewer for raising this important issue, as it led to a more rigorous assessment of predictor reliability and a more robust final model design.

Changes to be made in revised manuscript:

- Add predictor sensitivity / feature ablation analysis.
- Exclude **NO_2**, **SO_2**, and **VOC** from final elemental SA model.
- Revise Section 3.2 discussion on predictor reliability.

Concern #4: Use of Below-Detection-Limit SO_2 Data in Predictive Models:

The authors acknowledge that all ambient SO_2 concentrations at both sites fall below the minimum detection limit (MDL) of the Alphasense B4 sensor (~5 ppb), resulting in R^2 values near zero compared to reference grade instrumentation (p. 10, l. 261-263; 0.03 at Site-B, −0.17 at Site-C). Yet SO_2 appears to be retained as a predictor in the SA regression models, and the authors do not explicitly state that all SO_2 measurements are excluded. ***Using as model predictor a pollutant where all measurements are sub-MDL introduces noise rather than signal and violates fundamental principles of analytical measurement.***

Response: We sincerely thank the reviewer for highlighting this important issue regarding the inclusion of **SO_2** as a predictor in the machine learning framework.

We agree with the reviewer that retaining a predictor with poor calibration reliability can introduce noise and adversely affect model performance. In the original manuscript, SO_2 was retained among the input predictors primarily because it was one of the measured variables in the LCAQ sensor unit and was initially included during exploratory model development. However, upon re-evaluation, we agree that its inclusion in the final model is not sufficiently justified.

First, the calibration performance of the SO_2 sensor was indeed extremely poor at both deployment sites. The coefficient of determination between the **reference-grade gas analyser measurements** and the **calibrated LCAQ SO_2 measurements** in the original manuscript was:

- **Site-B (BBAU): R^2 = 0.03**
- **Site-C (CSIR): R^2 = –0.17**

These values indicate essentially no meaningful predictive relationship between the calibrated SO_2 signal and the reference measurements.

Second, we revisited the technical specifications of the **Alphasense SO2-B4 electrochemical sensor** used in our LCAQ unit. Although the manufacturer does **not explicitly specify a formal minimum detection limit (MDL)**, the datasheet reports a sensor **noise floor of 5 ppb (±2 standard deviations, ppb equivalent)**. This implies a practical field detection threshold of approximately **5–10 ppb**, below which sensor measurements become increasingly dominated by noise rather than true signal.

Furthermore, the datasheet indicates substantial **cross-sensitivity** of the SO_2 sensor to other gases, particularly:

- **NO_2 sensitivity: < –120%**
- **O_3 sensitivity: < –120%**

These strong cross-sensitivities suggest that under ambient urban conditions, fluctuations in NO_2 and O_3 may significantly distort the SO_2 sensor response, further reducing its reliability as an independent predictor.

To quantitatively assess the impact of SO_2 on the machine learning framework, we performed an additional **feature ablation study** using the best-performing regression model (**Extra Trees**) under the most stringent transfer-learning setting (**Lucknow (Train) → Kanpur**

(Test)). The mean Spearman rank correlation (SROCC) for elemental source apportionment changed as follows:

- Full predictor set: **0.8288**
- Without SO_2: **0.8291**
- Without SO_2 and NO_2: **0.8303**
- Without SO_2, NO_2, and VOC: **0.8349**

These results show that removing SO_2 does **not degrade performance**; instead, it leads to a small but consistent improvement in model generalization.

Based on the poor calibration performance, the practical detection threshold inferred from the sensor noise characteristics, the strong cross-sensitivity to NO_2 and O_3, and the feature ablation analysis, we agree with the reviewer that retaining SO_2 in the final model is unnecessary and potentially detrimental.

Therefore, in the revised manuscript, we will **exclude SO_2 from the final machine learning framework**. In conjunction with additional predictor sensitivity analysis, we will also remove **NO_2** and **VOC**, retaining only the more reliable predictors:

- CO
- O_3
- $PM_{2.5}$
- BC
- Temperature
- Relative Humidity
- Hour-of-day cyclical features

We thank the reviewer for this important comment, which helped improve both the physical validity and robustness of the final model.

Changes to be made in revised manuscript:

- Add predictor sensitivity / feature ablation analysis.
- Exclude **NO_2**, **SO_2**, and **VOC** from final elemental SA model.
- Revise Section 3.2 discussion on predictor reliability.

Concern #5: Cross-Sensitivity of Alphasense B4 Sensors Not Validated:

The authors note (p. 34, l. 607–608) that the Alphasense B4 auxiliary electrode compensates for cross-sensitivities from interfering gases; however, the calibration model (Equation 1) does not include any terms to explicitly test for or remove residual cross-sensitivities among pollutant channels. The authors must either demonstrate empirically that cross-sensitivities are negligible in their deployment context (e.g., by showing that adding cross-sensitivity terms explains negligible additional variance), or account for these effects within the calibration framework. This is particularly important given the poor NO_2 and SO_2 calibration results. The currently ***developed LCS calibration model is unfit for use.***

Response: We sincerely thank the reviewer for this insightful comment regarding the possible influence of residual cross-sensitivities among the electrochemical gas sensors.

We agree that electrochemical sensors are susceptible to cross-gas interference and that evaluating whether information from neighbouring gas channels can improve calibration is an important consideration. In response to the reviewer's suggestion, we performed an additional calibration study by extending the original calibration framework to explicitly include voltage outputs from the remaining gas sensors.

Specifically, two calibration models were investigated for each target gas. The **original calibration model** employed only the working and auxiliary electrode outputs (OP1 and OP2) of the target gas sensor together with temperature and relative humidity. The **extended calibration model** additionally incorporated the OP1 and OP2 outputs from the remaining gas sensors to account for potential cross-gas interactions.

To ensure a rigorous evaluation, the calibration models were developed using the combined **Lucknow datasets (CSIR and BBAU; October 2023 and February–March 2024)** and subsequently evaluated on an **independent dataset collected at Kanpur (December 2024)**.

The results indicate that the influence of cross-gas information is **gas dependent** rather than universal. For **CO**, the extended calibration model improved the test MAE from **0.44 to 0.37**, increased the Spearman correlation coefficient from **0.81 to 0.88**, and improved the coefficient of determination from **0.77 to 0.83**. A more substantial improvement was observed for **NO_2**, where the test MAE decreased from **12.12 to 8.94**, the RMSE decreased from **14.15 to 10.23**, and the Spearman correlation coefficient increased from **0.68 to 0.85**. These observations indicate that incorporating cross-gas information can partially compensate for residual cross-sensitivity effects in certain electrochemical sensors.

In contrast, only marginal changes were observed for **O_3**, while the calibration performance for **SO_2** deteriorated after including cross-gas terms. This behaviour is consistent with the fact that the ambient SO_2 concentrations during the measurement campaigns were predominantly close to or below the effective operating range of the sensor, where the measured signal is largely noise dominated. Under such conditions, incorporating additional gas-channel information does not provide meaningful calibration improvements.

These results demonstrate that residual cross-gas interactions do exist for certain sensors, particularly **NO_2**, but their influence is not sufficiently consistent across all gases to justify replacing the original calibration framework within the scope of the present work. Since the primary objective of this manuscript is the development and validation of a **real-time source apportionment framework**, rather than the development of a new calibration methodology, we have retained the original calibration approach while expanding the discussion to acknowledge the observed gas-dependent behaviour. A comprehensive investigation of advanced cross-sensitivity-aware calibration models constitutes an important direction for future work and will be pursued separately.

We thank the reviewer for this valuable suggestion. The additional analyses have improved our understanding of residual cross-gas interactions and have enabled us to present a more balanced discussion of the calibration methodology in the revised manuscript.

N.B: We would also like to clarify that the calibration performance that will be reported in the revised manuscript should not be directly compared with that presented in the originally submitted version. The revised study employs a substantially different experimental protocol based on independent cross-city evaluation (Lucknow (Train) → Kanpur (Test)), whereas the original manuscript reported results obtained using within-site train–test partitions. The revised calibration results therefore represent a considerably more challenging and realistic assessment of model generalization.

Changes to be made in the revised manuscript:

- Include a new subsection discussing the influence of **cross-gas information** on calibration performance.
- Present a summary table comparing the original and cross-gas-aware calibration models for **CO, NO_2, O_3, and SO_2**.
- Expand the discussion to clarify that cross-gas effects are **gas dependent**, with substantial improvements observed for **NO_2**, moderate improvements for **CO**, negligible changes for **O_3**, and no meaningful benefit for **SO_2**.
- Clarify that the objective of the present work is **real-time source apportionment**, while a comprehensive investigation of advanced calibration methodologies is beyond the scope of this manuscript and will be considered in future work.

Concern #6: Absence of a Validation Set — Risk of Overfitting and Contaminated Evaluation:

The paper describes only a training/test split (80/20). No separate validation set is used for hyperparameter tuning or model selection (Appendix C evaluates multiple regression models including gradient boosting and random forests with tuned hyperparameters). Without a held-out validation set, the reported test-set performance risks being optimistic: if any model selection decisions were informed — even informally — by test-set behavior, the test set no longer represents a true independent evaluation. Standard practice in supervised machine learning requires a three-way split (e.g., 60/20/20 or 70/15/15) when multiple models or hyperparameters are compared. In the future, the authors must clarify their model selection procedure and, if the test set was used in any capacity for model comparison, must provide corrected evaluation on a genuinely held-out partition.

Response: We sincerely thank the reviewer for highlighting this important methodological issue regarding the use of an independent validation set during model development.

In the originally submitted manuscript, the machine learning framework employed only a training and test partition. Although the test data were never used during model fitting, we agree that when multiple machine learning algorithms and hyperparameter combinations are compared, the use of an explicit validation set provides a more rigorous and unbiased model selection procedure.

In response to the reviewer's comment, we have restructured the complete machine learning workflow adopted in this study. For every transfer-learning experiment, the **training-domain dataset** is now partitioned chronologically into **80% training data** and **20% validation data**, while preserving the temporal ordering of the measurements (**shuffle = False**). The **target-**

domain dataset is used exclusively as an independent external test set and is not involved in either model training or hyperparameter selection.

Accordingly, hyperparameter tuning for all machine learning models (Random Forest, Gradient Boosting, Extra Trees, XGBoost, Weighted k-Nearest Neighbours and Support Vector Regression) was performed solely using the validation dataset. After selecting the optimal hyperparameter combination based on the minimum validation MAE, the corresponding model was retrained using the combined training and validation datasets before being evaluated on the completely unseen external test dataset.

This revised evaluation strategy eliminates any possibility of information leakage from the test set during model selection and provides a significantly more rigorous assessment of model generalization performance.

Furthermore, all additional transfer-learning experiments reported in the revised manuscript—including the cross-city (Lucknow → Kanpur), cross-site (CSIR → BBAU and BBAU → CSIR), and cross-season (October → February–March and February–March → October) evaluations—were conducted using this revised three-way data separation strategy.

We appreciate the reviewer for pointing out this important methodological issue, which has substantially strengthened the evaluation protocol adopted in the revised manuscript.

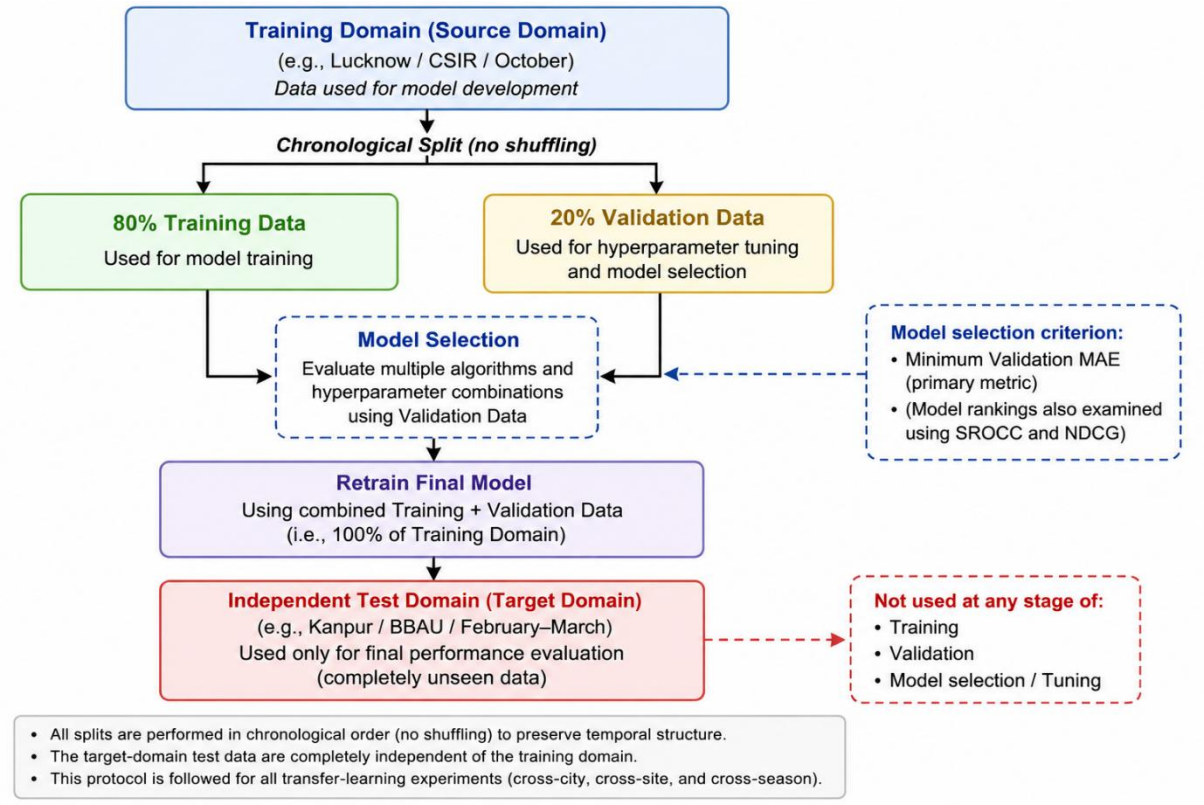


Figure: Modified train-test protocol used for all our experiments.

Changes to be made in the revised manuscript:

- Revise the machine learning methodology to explicitly describe the **training–validation–test** workflow.
- Clarify that the **training-domain data** are split chronologically into **80% training** and **20% validation** (**shuffle = False**).
- State explicitly that the **target-domain data** are used solely as an independent external test set.
- Clarify that **all hyperparameter tuning and model selection** are performed exclusively on the validation set.
- Update the descriptions of all machine learning experiments to reflect the revised evaluation protocol.

Concern #7: Use of R^2 for Calibration Assessment is Inappropriate; Spearman Correlation Required:

The paper uses Pearson R^2 to evaluate sensor calibration (Figures 3 and 4) and the feature correlation heat map (Figure 6). The R^2 values reported for NO_2 (0.22 at Site-B) and SO_2 (−0.17 at Site-C) are strikingly poor and should disqualify these sensors from use, yet the manuscript characterizes the calibration as "reasonably good" (p. 10, l. 259-260) and proceeds to include these pollutants as model inputs. Pearson R^2 is sensitive to outliers and is poorly suited to noisy electrochemical sensor data. In general, the authors must replace Figures 3, 4, and 6 with Spearman rank-order correlation coefficients, which are more appropriate for this type of data and provide an honest characterization of calibration quality.

Response: We sincerely thank the reviewer for this valuable suggestion regarding the statistical assessment of sensor calibration.

We agree that electrochemical sensor responses may exhibit nonlinear behaviour and may be influenced by occasional outliers arising from environmental variability. Under such circumstances, **Spearman's rank-order correlation coefficient (ρ)** provides useful complementary information by quantifying the strength of the monotonic relationship between the calibrated sensor measurements and the corresponding reference-grade observations.

At the same time, we respectfully note that the **coefficient of determination (R^2)** remains one of the most widely adopted performance metrics for regression-based sensor calibration, as it quantifies the proportion of variance explained by the calibration model and directly reflects goodness-of-fit. Consequently, Pearson-based R^2 and Spearman correlation characterize different aspects of calibration performance and should be regarded as complementary rather than mutually exclusive evaluation metrics.

In response to the reviewer's suggestion, we will revise the manuscript to report **both R^2 and Spearman's rank correlation coefficient** for the calibration results presented in Figures 3 and 4. This will allow readers to simultaneously assess both the regression quality and the monotonic agreement between the calibrated LCAQ measurements and the reference-grade instrument observations.

Furthermore, we agree that the feature relationship analysis presented in **Figure 6** is better represented using rank-based statistics. Accordingly, the correlation heatmaps in the revised manuscript will be recomputed using **Spearman rank correlations** instead of Pearson correlations.

We believe that these revisions provide a more comprehensive and robust statistical characterization of the calibration performance while preserving the conventional regression-based evaluation of the calibration models.

We thank the reviewer for this constructive suggestion, which has improved the statistical presentation of the calibration results.

Changes to be made in the revised manuscript:

- Report both **Pearson R^2** and **Spearman's rank correlation coefficient (ρ)** in the calibration plots (Figures 3 and 4).
- Replace the Pearson correlation matrix in **Figure 6** with a **Spearman rank correlation matrix**.
- Expand the discussion to explain the complementary roles of R^2 (goodness-of-fit) and Spearman correlation (monotonic agreement) in evaluating electrochemical sensor calibration.