Response to comments on “Machine Learning Calibration of Low-Cost Air Quality Gas Sensors” by Reviewer 1.

We sincerely thank Reviewer 1 for the time and effort devoted to review our manuscript. The comments and suggestions were thoughtful, constructive, and highly valuable. We have carefully addressed each point raised and revised the manuscript accordingly. We believe that these revisions have considerably improved the clarity, quality, and overall contribution of the work.

REVIEWER 1

egusphere-2026-897

Machine Learning Calibration of Low-Cost Air Quality Gas Sensors

Ioannidis et al.

General comments

Ioannidis et al. explore calibration approaches for low-cost gas sensors for environmental and air quality monitoring. A six-month dataset including different types of Alphasense gas sensors in Nicosia, Cyprus was used to explore various machine learning algorithms to calibrate the sensors' observations with co-located reference monitors which were considered the ground truth. The results show that the calibration models improved the measurement performance of the sensors, the random forest algorithm was the most performant algorithm, and recommendations are given for what aggregation periods, the fraction of training/testing sets, and the frequency of calibration are best for these types of sensors.

The manuscript is well written, is clear, and the figures are presented well. However, I have concerns about the novelty and generality of the work presented. The calibration of low-cost sensors for air quality monitoring is a well-researched topic, and Ioannidis et al. do not provide new or solid insights. The authors touch on the poor measurement performance of low-cost sensors being driven by manufacturers focusing their calibration procedures on laboratory testing, rather than the variability that is experienced in environmental monitoring applications. Although this is undoubtedly true, there are two other fundamental issues with low-cost sensors and their measurement performance that need to be solved: (i) many of these gas sensors respond to other environmental variables much more than the target gas, i.e., they are very unspecific. Figure 6 demonstrates this point very well. (ii) The calibration of low-cost sensors is unstable over time and this needs very careful management. The latter point is only discussed indirectly. Therefore, we have a situation where a gas sensor is unspecific to the measurand and the calibration is unstable.

These two features of low-cost sensors put into question the portability of the calibration strategy applied in Ioannidis et al. I would argue that the splitting of a dataset into training and testing sets is a mechanism to train and test a model, but a third validation set is required to truly determine if any given trained model is useful for a monitoring application. The training and testing set comes from the same calibration space, but a validation period should be outside the space, generally in the future from the model's perspective, and be tested with truly parallel observations of the ground truth. This situation gives rise to "true" field performance as could be expected if a low-cost sensor were deployed to a field location without a reference monitor, exactly the point of these devices.

I would recommend that the authors apply their testing to a validation set for more robust performance metrics and further evaluate the fundamental deficiencies of low-cost sensors in their interpretation of their results.

Response to general comments: We agree that the calibration of low-cost air quality sensors has been extensively studied and that sensor cross-sensitivities and temporal instability are among the most important limitations affecting their practical use. To address the concerns raised by the reviewer, we have structured our response around three main points: (a) the innovative aspects and contributions of the present study, (b) the influence of environmental variables on sensor response, and (c) the transferability of the proposed calibration framework beyond the conditions represented in the current dataset, as recommended by the reviewer.

a) Innovative aspects: The primary objective of this study is to evaluate practical aspects of low-cost sensor calibration that are directly relevant to operational air quality monitoring. Beyond comparing calibration models, the study systematically investigates the effects of temporal resolution, training dataset size, sampling strategy, and recalibration frequency across four atmospheric pollutants, while assessing performance against the EU Data Quality Objectives.

These practical deployment considerations, which are the focus of the manuscript, are described in lines 81-86 of the introduction of the revised manuscript: "After identifying the best-performing calibration model, we further investigate a number of practical aspects that are directly relevant to the deployment of LCS networks. These include the influence of the temporal resolution of the training data, the effect of different sampling strategies, the frequency at which recalibration should be performed, and the minimum amount of collocation data required for model development. Emphasis is placed on evaluating how these factors affect compliance with the EU Data Quality Objectives and, consequently, the suitability of calibrated LCS for air quality monitoring applications."

This is also highlighted in the conclusions section in lines 441-445: "We have evaluated the performance of various algorithms for calibrating LCSs data from CO, NO_2, O_3 and SO_2, and identified that the Random Forest model provides the best results. This model is then used (1) to investigate how different input variables affect the calibration performance, and (2) to investigate the extent to which practical parameters such as the temporal resolution of the measurements, how the training data is sampled, and the frequency of calibration reduce the fraction of data needed for training, while keeping the quality of calibrated data within the accepted levels."

b) Gas sensors response to other environmental parameters: We agree that many low-cost sensors are not exclusively responsive to their target gas and may exhibit substantial sensitivity to environmental conditions and interfering species. In fact, the feature-importance analysis presented in Figure 7 (Fig. 6 in the earlier version of the manuscript) provides direct evidence of this problem and allows us to quantify the relative contribution of target-gas concentrations, environmental variables, and temporal parameters to the performance of the calibration models. To better address this point, we have expanded the discussion of Fig. 7 of the revised manuscript in lines 362-367: "These results highlight a fundamental limitation of low-cost electrochemical gas sensors, namely their sensitivity to environmental conditions in addition to the target pollutant. For the NO_2, O_3 and SO_2 sensors, temperature and RH account for a substantial fraction of the model performance, indicating that sensor response is strongly influenced by ambient conditions. This explains why multivariate calibration approaches are required and suggests that calibration performance may vary when applied under environmental conditions not represented in the training dataset."

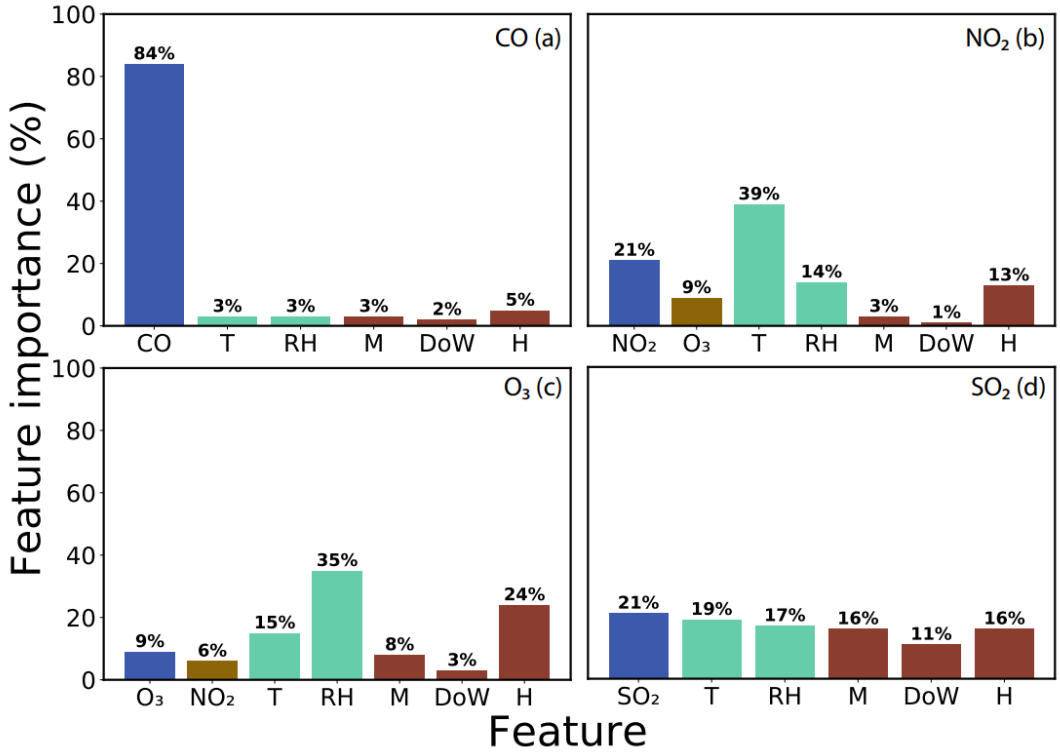

Figure 7: Estimated importance of each input feature in determining the variability of the signals reported by CO, NO_2, O_3 and SO_2 LCSs, determined for each sensor using the RF model. For CO and SO_2 the model is trained using respective reference concentrations of the target gases, temperature, relative humidity, month, day of week and hour as input variables, and the NSS as dependent variable. For NO_2 and O_3, the training of the RF algorithm also includes respectively, the O_3 and NO_2 reference concentrations as input variables to assess their cross-sensitivity.

c) Calibration method transferability and stability: We thank the reviewer for raising this important point regarding the temporal portability of ML calibration models. In the present study, each calibration interval (1, 3, or 6 months) was divided into training and previously unseen testing datasets, allowing the models to be validated on measurements that were not used during training. Although this does not represent a true long-term transferability assessment, it provides an evaluation of model performance on independent observations within each calibration period while enabling us to investigate the influence of recalibration frequency, sampling strategy, and training data fraction on calibration performance.

We acknowledge that this does not constitute a true long-term transferability assessment, which would require training a model once and validating it on future observations without recalibration. We have clarified in the revised manuscript that one of the objectives is to identify practical calibration strategies that minimise the amount of collocated reference measurements while maintaining the data quality required for operational air quality monitoring (Section 2.4) in lines 213-225:

"2.4 Practical calibration strategy

To determine the minimum amount of collocated data required to develop an accurate calibration model and the influence of recalibration frequency on calibration performance, different calibration intervals (1, 3 and 6 months), training data fractions and sampling strategies were evaluated. For each calibration scenario, the calibration models were trained using only a selected fraction of the collocated measurements, while the remaining measurements within the same calibration period were excluded from model development and used exclusively for performance evaluation against the corresponding

reference measurements. This approach enables the calibration strategy to be assessed using previously unseen observations while investigating practical approaches for minimizing the duration of collocation with reference-grade instruments.

For each calibration scenario, the calibrated measurements obtained from the held-out testing dataset were evaluated using the statistical metrics described in Section 2.2 together with the EU Data Quality Objectives presented in Section 2.3. The influence of calibration interval, training data fraction, and sampling strategy on calibration performance is presented and discussed in Section 3.3. This analysis provides practical guidance for optimizing calibration effort while maintaining the data quality required for operational air quality monitoring applications."

To further address the transferability of the proposed calibration strategy, we expanded the discussion at the end of Section 3.3 to clarify its practical implementation in long-term monitoring applications. We now emphasize that only a relatively short co-location period is required before deployment.

In lines 378-384: "Using RF model, here we investigate how the amount of training data can be minimized, while not significantly sacrificing their quality, by varying practical parameters such as (1) the temporal resolution of the training data, (2) how the training is sampled, and (3) the frequency of calibration. For each calibration scenario, the RF model was trained using only the selected fraction of the collocated dataset, while its performance was evaluated on the remaining previously unseen measurements by comparison with the corresponding reference observations. This approach enables the calibration strategy to be assessed under conditions representative of practical deployment while ensuring that model performance is evaluated on data not used during training."

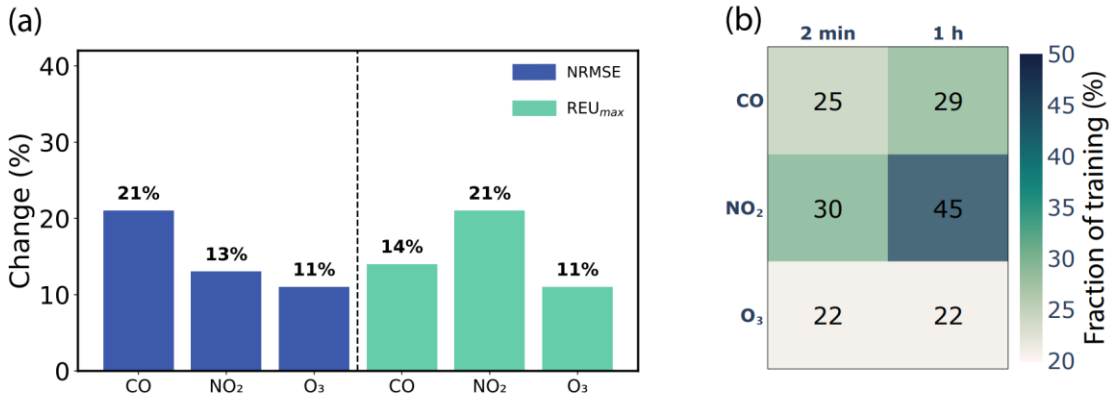

Figure 8: (a) Percentage change in the normalized root mean squared error (NRMSE) and the average relative expanded uncertainty calculated at maximum concentration observed (REUmax) when the resolution of the data used to train the RF model is increase from 1 h to 2 min, and (b) the minimum fraction for 2 min and 1 h measurements required to train the RF model that will guarantee its calibrations to meet the data quality objective for indicative measurements defined by the EU directive.

In the end of section 3.3 we added to the revised text in lines 432-439: "The results presented in this section provide practical guidance for the deployment of LCS networks, demonstrating that a relatively short co-location period, combined with interceptive sampling, minimizes the amount of reference data required while maintaining compliance with the EU data quality objectives and improving the cost-effectiveness of long-term monitoring. Furthermore, by training the calibration models using only a selected fraction of the collocated measurements and evaluating their performance on the remaining previously unseen measurements, the proposed methodology provides an objective assessment of model performance on data not used during model development. The results also indicate that calibration performance is highest when the training and application periods are characterized by similar

environmental conditions, highlighting the need for periodic recalibration during long-term deployments."

Itemised comments

- Line 72. Linear regression is probably not considered a machine learning model

We thank the reviewer for addressing this. We agree that Linear Regression is not considered a machine learning model but we want to keep it in the analysis. Considering that, we edited lines 74-78 of the manuscript stating: "In this work, we use concentration measurements of atmospheric pollutants recorded by LCSs and reference instruments over a period of 6 months in a traffic station in the city of Nicosia, Cyprus, in order to train and evaluate the performance of one statistical method, namely Linear Regression (LR; Kumar and Sahu 2021), and four ML algorithms: i.e., Support Vector Regression (SVR; Kumar and Sahu 2021); Random Forest Regressor (RF; Zimmerman et al. 2018); Artificial Neural Network (ANN; Spinelle et al. 2015); and Extreme Gradient Boosting (XGBoost; Chen and Guestrin 2016) for low-cost sensor calibration."

In the remaining manuscript we replaced the "ML calibration approaches" with just "calibration approaches" in all occurrences as it is more general and includes non-ML models such as LR.

- Line 104. Is this a good idea? If there is uncertainty in the positive direction, there should be an uncertainty in the negative direction to avoid a bias.

We thank the reviewer for this insightful comment. Negative net sensor signals were removed because they are physically non-meaningful, as they represent the difference between the working and auxiliary electrode responses and are expected to be positive under normal operating conditions. To make this more clear lines 112-116 in the updated version of the manuscript state: "Subsequently, data cleaning was done by dropping all rows with the missing values and those with negative values of the net sensor signal (NSS; i.e., obtained by subtracting the signals reported by the auxiliary electrode from those reported by working electrode). Negative NSS values were assumed to arise from low noise-to-signal ratios. While their exclusion may introduce a small bias under very low-concentration conditions, their occurrence represented only a minor fraction of the dataset."

- Line 148. Only one of r or R2 probably needs to be used; they mostly overlap and show the same thing

We thank the reviewer for this comment. We agree that Pearson's r and R^2 are related metrics and may provide partially overlapping information. For that reason, we have removed the values of R^2. This is revised in the text in lines 152-155: "The Pearson correlation coefficient, r, provides information on the strength of associations between the calibrated concentrations and their corresponding reference concentrations whereas NRMSE expresses the error between the calibrated and reference concentrations normalized by the mean of the reference concentrations."

Figure 5 now only includes NRMSE analysis as stated in lines 289-291 of the revised manuscript: "Figure 5 provides heat maps that show the performance of all the calibrations (including LAB) performed in this study based NRMSE. The good performance of the ML calibrations was reflected by the relatively low NRMSE values (see Fig. 5), which exhibit a significant improvement compared to the values of their corresponding LAB calibration."

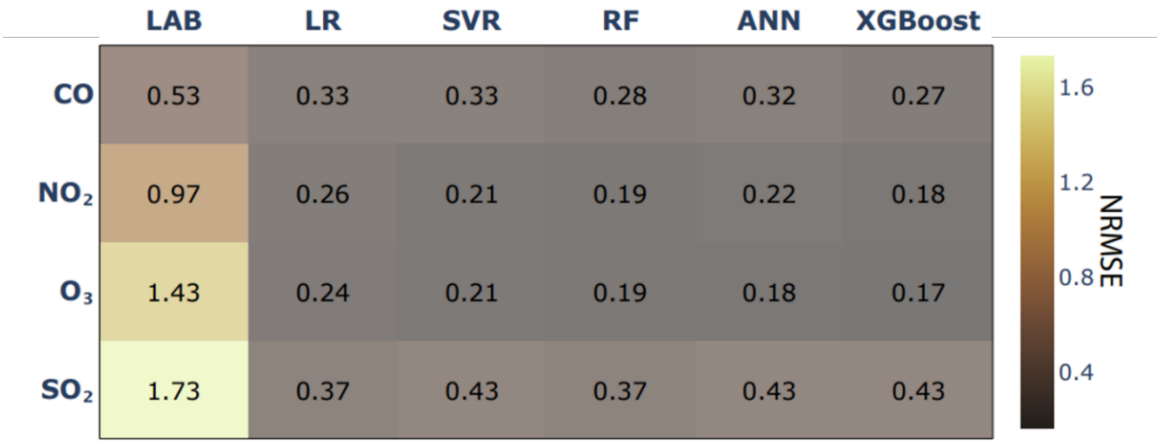

Figure 5: Heat map showing the performance of the LCSs in terms of normalized root mean squared error (NRMSE) after calibration using the laboratory tests carried out by the manufacturer (LAB) and the five ML algorithms employed in this work.

The estimated R^2 values were added to the Supplement (Figure S2):

"We evaluated the accuracy of the calibration models based on the coefficient of determination (R^2) shown in equation S1:

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(Cal_t - Ref_t)^2}{\sum_{t=1}^{n}(Ref_t - \overline{Ref})^2}, \quad \text{(S1)}$$

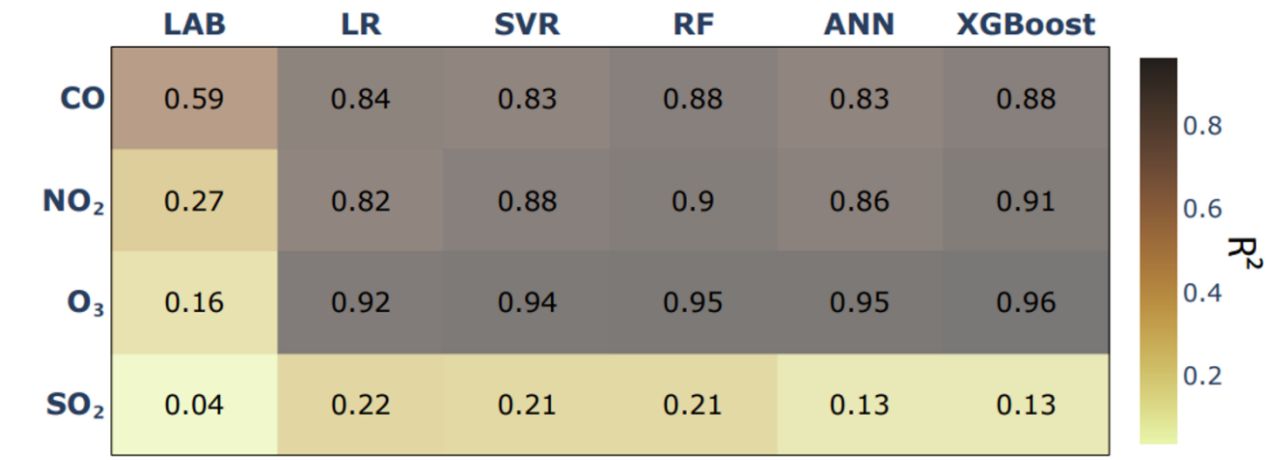

Figure S2: Heat map showing the performance of the LCSs in terms of coefficient of determination (R^2) after calibration using the laboratory tests carried out by the manufacturer (LAB) and the five ML algorithms employed in this work."

- Line 200 onward. Many sub- and super-scripts are missing in the text

We thank the reviewer for pointing out this inconsistency. Every case is now corrected in the manuscript (R2→R^2), (SO2→ SO_2), (O3→O_3) and so on.

- Line 221. When `n` is so high, statistical significance between two groups will almost always be detected, so this might not be a very useful test

We agree that, given the very large sample sizes considered in this study, statistical significance can be detected even for relatively small differences between datasets. We also acknowledge, as noted by

Reviewer 2, that the concentration datasets were not normally distributed. Therefore, the original t-tests were replaced with the non-parametric Mann–Whitney U test, which does not assume normality.

This is clearly explained in lines 227-243 of the revised manuscript: "Table 1 provides summary statistics of the concentration measurements reported by the reference instruments and the LAB calibrated LCSs. The differences in the mean and standard deviation between the measurements obtained by the LCSs and the reference instruments for the four pollutants range between ca. 10 and 1300% and ca. 30 and 2600%, respectively. To evaluate the statistical significance of these differences, we first checked normality of all LCS and reference measurements by running Shapiro-Wilk tests, which returned p-values less than 0.05, indicating that all the LCSs and reference measurements were normally distributed. Next, we performed Mann–Whitney U tests to assess the statistical significance of the differences between the distributions of LCS and reference concentration measurements for each pollutant. The Mann–Whitney U test was selected because it does not assume normality and is therefore appropriate for comparing non-normally distributed datasets (Mann and Whitney, 1947).

Table 1: Summary statistics of concentration measurements of CO, NO_2, SO_2 and O_3 recorded by the LCSs, using the laboratory (LAB) calibrations, and the reference (REF) instruments.

	CO (REF)	CO (LAB)	NO_2 (REF)	NO_2 (LAB)	O_3 (REF)	O_3 (LAB)	SO_2 (REF)	SO_2 (LAB)
Number of data points	111,025	111,025	102,427	102,427	94,520	94,520	44,520	44,520
Average (ppb)	470	430	18	33	18	62	1.6	27
Standard deviation (ppb)	288	262	13	26	15	39	0.98	23
Minimum (ppb)	2.87	1.79	0.26	0.030	0.002	0.070	0.001	0.170
Maximum (ppb)	3,573	3,110	99	1,365	64	1,789	96	252

The Mann–Whitney U tests yielded p-values $< 10^{-10}$ for all four pollutants, indicating statistically significant differences between the distributions of the laboratory-calibrated LCS measurements and the corresponding reference measurements."

- Line 242. The SO2 sensor does not seem suitable for ambient SO2 monitoring

Indeed, the SO_2 sensor exhibited poor performance, which can be mainly attributed to the low ambient concentrations encountered during the measurement campaign. However, the objective of this study was not to assess the suitability of individual sensors for deployment, but rather to investigate the extent to which statistical and ML calibration approaches can improve the performance of commercially available low-cost sensors. For this reason, we chose to retain the analysis of the data from the SO_2 sensor despite its poor initial performance. We believe that this result is valuable because it demonstrates both the potential and the limitations of calibration approaches when applied to sensors operating close to their detection limits.

To better highlight this point, lines 262-266 in the updated version of the manuscript state: "The measurements reported by the SO_2 LCS highly overestimate the concentrations of this pollutant as compared to those reported by the reference instrument, indicating limited suitability of the sensor for ambient monitoring under the low-concentration conditions encountered during the campaign. Despite that, the measurements provided by the SO_2 LCS provide a useful test case for evaluating the extent to which statistical and calibration approaches employed in this paper can improve the performance of LCSs operating close to their detection limits."

- Line 245-ish. Displaying time series would be useful too. Maybe a clear single example could be added to the text?

We thank the reviewer for this helpful suggestion. Following the recommendation, we have added a new figure (Figure 4) showing a representative time series comparison between the reference measurements, the laboratory-calibrated LCS observations, and the outputs of all calibration models. Figure 4 provides a direct visual comparison of the ability of the different calibration approaches to reproduce both the temporal variability and concentration peaks observed by the reference instrument.

The corresponding discussion has been added in Section 3.1 (Lines 278-285): "Figure 4 shows representative time series of the calibrated measurements by the CO and O_3 LCSs in comparison with measurements provided by the reference instruments from 1 to 7 December 2019. While the laboratory-calibrated CO measurements already show good agreement with the reference measurements, all (ML) models further improve the representation of concentration peaks and temporal variability. In contrast, the laboratory-calibrated O_3 measurements exhibit poor agreement with the reference observations, whereas all ML models substantially improve its performance. Among the evaluated methods, RF provides the closest overall agreement with the reference measurements for both pollutants, followed closely by XGBoost and ANN, while linear regression shows larger deviations during rapidly changing concentration events. These observations are consistent with the statistical performance metrics presented above."

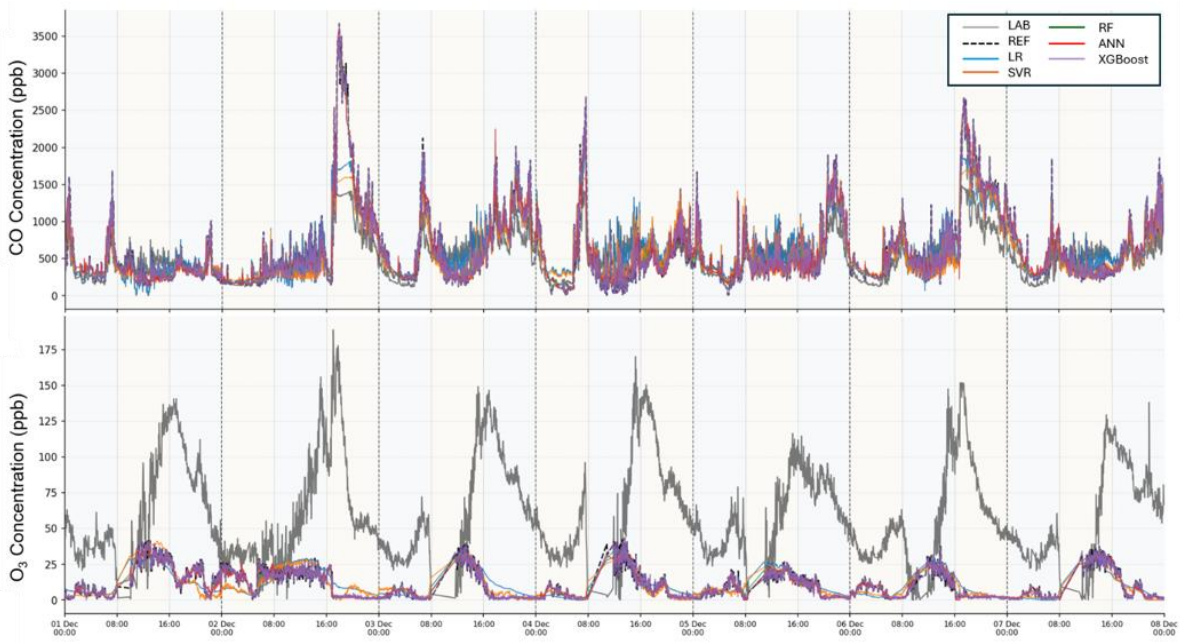

Figure 4: Time series of reference (REF), low-cost sensor (LAB), and calibrated CO and O_3 concentration measurements from 1 to 7 December 2019 using the five calibration approaches evaluated in this study.

."

- Line 260-ish. Do any of the models require much more training and/or prediction time than others? Is time or computation resource requirements a practical issue for this dataset?

We thank the reviewer for this insightful comment. To assess the practical computational requirements of the different calibration approaches, we recorded the training times of all models. While training times are not a limiting factor for the present dataset, the lower computational cost of RF becomes

increasingly important when processing larger datasets, performing frequent recalibrations, or calibrating large networks consisting of multiple sensors.

We added an extra explanation for the choice of RF in terms of computational efficiency in lines 330-340: “Another practical consideration when selecting a calibration model is its computational cost. The reported training times were obtained using a standard desktop computer and therefore represent practical runtimes under typical operating conditions. For the datasets used in this study, training times varied considerably among the evaluated algorithms. The LR model required less than 0.03 s for training, while the RF model, which consistently achieved the best overall calibration performance, required between approximately 5 and 14 s across the four pollutants. In comparison, the SVR model required between approximately 2 and 448 s, XGBoost between 62 and 371 s, and ANN between 23 and 194 s, depending on the pollutant. Although these computational requirements are not at all restrictive for an individual calibration, they can become increasingly relevant when calibration procedures are performed repeatedly throughout long-term monitoring campaigns or across multiple sensors. In this regard, the RF model provides an attractive balance between calibration accuracy and computational efficiency, achieving the best overall calibration performance while maintaining substantially lower training times than the ANN, SVR, and XGBoost models.”

- Line 356. This is a key point that could be tested with a validation dataset.

We thank the reviewer for this observation. A detailed response to this comment is provided in the general comment response under c) Calibration method transferability and stability section.

References:

Kumar, V. and Sahu, M.: Evaluation of nine machine learning regression algorithms for calibration of low-cost PM2. 5 sensor, Journal of Aerosol Science, 157, 105 809, 2021.

Zimmerman, N., Presto, A. A., Kumar, S. P., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., and Subramanian, R.: A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring, Atmospheric Measurement Techniques, 11, 291–313, 2018.

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M., and Bonavitacola, F.: Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide, Sensors and Actuators B: Chemical, 215, 249–257, 2015.

Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, 2016.

Henry B. Mann and Donald R. Whitney (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, 18(1), 50–60.

Citation: https://doi.org/10.5194/egusphere-2026-897-RC1