Response to comments on "Machine Learning Calibration of Low-Cost Air Quality Gas Sensors" by Reviewer 2.

REVIEWER 2

Overall quality of the preprint - general comments

The overall impression of the paper is that it is very well-written and well-organised. The study is also interesting and, in my view, well worth publishing. However, before publication, the authors need to address the scientific questions and issues outlined below.

We thank the reviewer for the assessment of our manuscript and for recognizing the quality of the presentation, organization, and scientific relevance of the study. We appreciate the constructive comments provided and have carefully addressed all concerns raised. We believe that the revisions adequately address the reviewer's comments and improve the clarity of the manuscript.

Individual scientific questions/issues - specific comments

Issue #1: The calibrations have been carried out only over a 6-month winter period. It would have been a strength if they had covered a full year, including a summer period. It would be helpful if the authors could include a brief description of the potential challenges of including a summer period and how they believe it could have affected the results of the paper.

We thank the reviewer for this valuable suggestion and agree that extending the analysis to a full annual cycle would provide a more comprehensive assessment of calibration robustness under a wider range of environmental conditions. To acknowledge this limitation and discuss its potential implications, we have expanded both the Introduction and Results sections to describe how summer conditions may influence sensor response and calibration performance.

To address this point, we have added a discussion in the revised manuscript in the introduction section in lines 78-81: "The selected study period covers a wide range of pollutant concentrations and meteorological conditions representative of autumn and winter in the eastern Mediterranean region. However, it does not include summer conditions, which may further influence sensor response due to higher temperatures, higher RH, and enhanced photochemical activity (Spinelle et al., 2015)."

In addition, we have expanded the discussion section to acknowledge the potential impact of summer conditions on calibration performance. In lines 370-373 we added some discussion in section 3.2 Feature Importance after Figure 7 (previously Figure 6): "The strong contribution of temperature and RH to the calibration models suggests that sensor performance may vary under environmental conditions not represented in the present dataset. In particular, the high temperature and RH, and enhanced photochemical activity characteristic of Mediterranean summer periods may modify the response and the stability of the sensors, highlighting the need for future evaluations based on full annual datasets."

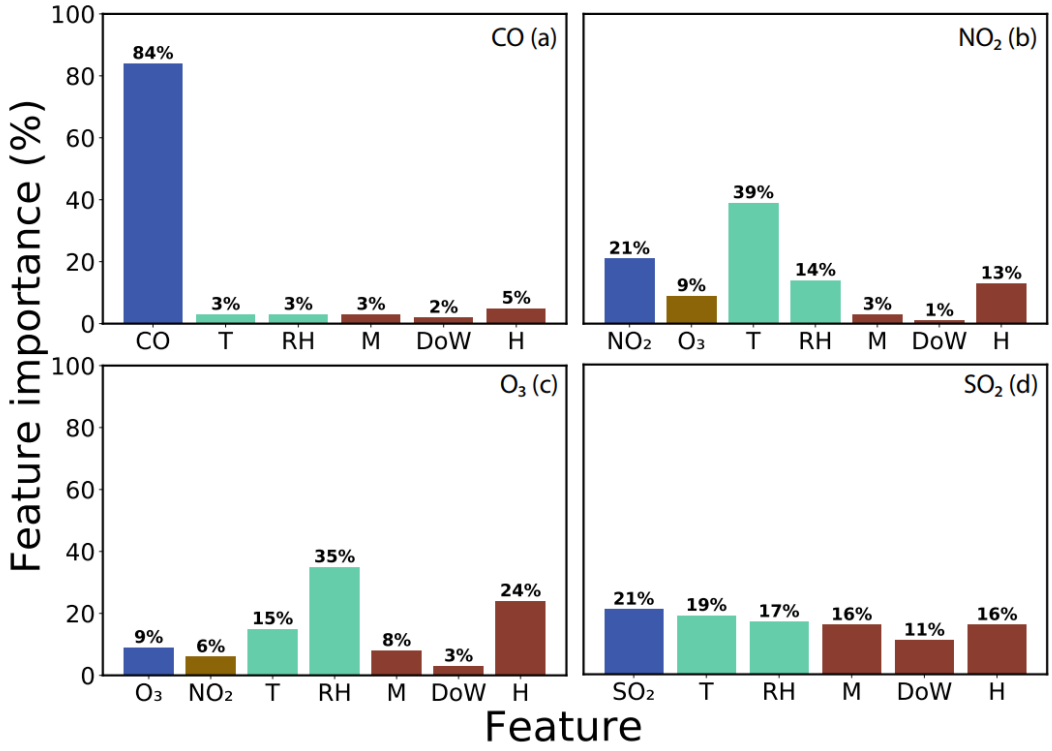

Figure 7: Estimated importance of each input feature in determining the variability of the signals reported by CO, NO_2, O_3 and SO_2 LCSs, determined for each sensor using the RF model. For CO and SO_2 the model is trained using respective reference concentrations of the target gases, temperature, relative 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 in order to assess cross-sensitivity.

Issue #2: The interpretation of p-values in relation to the Shapiro-Wilk test for normality is incorrect. A low p-value indicates rejection of the null hypothesis that the data are normally distributed, thereby indicating non-normality, which is to be expected. Thus, further tests of differences in means and variances should use statistical methods that do not assume normality.

We thank the reviewer for identifying this error. The low p-values from the Shapiro–Wilk test indicate rejection of the null hypothesis of normality and therefore suggest that the datasets are not normally distributed. In response to that, we have corrected the interpretation of the Shapiro–Wilk test throughout the manuscript. Furthermore, because the concentration datasets were found to be non-normally distributed, we replaced the parametric t-test originally used to compare means with the non-parametric Mann–Whitney U test, which does not assume normality. The Fligner–Killeen test was retained for variance comparisons because it is robust to departures from normality.

This is clearly explained in the revised manuscript in lines 227-243: “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 LCSs 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 differences between the distributions of the LCSs 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 LCS, 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."

Issue #3: Referring to Fig. 4, XGBoost is actually slightly better than RF for all compounds except SO2, yet you describe RF as the best model. RF is indeed very close, so I guess it is quite OK to use it in the rest of the paper. However, you need to explain why you have chosen it over XGBoost. Are there other arguments, such as ease-of-use or robustness, for example?

We thank the reviewer for this observation. We agree that XGBoost showed slightly higher performance than RF for CO, NO_2, and O_3 in Figure 5 (previously Figure 4). However, the differences were small, whereas RF performed better for SO_2. Since our objective was to select a single calibration model for all pollutants, RF was considered the best overall compromise. In addition, RF is computationally more efficient than XGBoost.

We have clarified this rationale in the revised manuscript in lines 327-340: "Although XGBoost showed marginally higher performance for CO, NO_2, and O_3, RF was selected because it provided the best overall compromise across all pollutants, including SO_2. The performance differences between RF and XGBoost were small for CO, NO_2, and O_3, while RF performed better for SO_2. 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 ca. 5 and 14 s across the four pollutants. In comparison, the SVR model required between ca. 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 restrictive for an individual calibration, they 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."

Issue #4: SO2 is often below the detection limit. Perhaps you should consider removing it from the paper?

We thank the reviewer for this suggestion. While we agree that the SO_2 sensor exhibited poor performance due to the very low ambient concentrations encountered during the measurement campaign, we chose to keep the analysis of the data from this sensor because it provides a valuable test case demonstrating both the capabilities and limitations of calibration approaches, showing that although ML methods can improve sensor performance, they cannot fully overcome fundamental limitations when measurements are close to or below the sensor detection limit.

To further stress that we have updated the manuscript and now lines 265-271 state: "The measurements reported by the SO_2 LCS highly overestimate the concentrations of this pollutant as compared to those measured 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 SO2 LCS provide a useful test case for evaluating the extent to which statistical and calibration approaches presented in this paper can improve the performance of LCS operating close to their detection limits. Overall, we observe an improvement in the agreement between measurements reported by the LCS and the respective reference instruments after calibration."

List of technical corrections - typing errors

1. 145: Cal and Ref should have bars above them to indicate averages.

We thank the reviewer for the comment. The bars denoting the mean values of Cal and Ref are already included in Eq. (1). However, to improve readability and avoid any ambiguity, the equation formatting has been revised in the updated manuscript as follows:

$$r = \frac{\sum_{t=1}^{n}(Cal_t - \overline{Cal})(Ref_t - \overline{Ref})}{\sqrt{\sum_{t=1}^{n}(Cal_t - \overline{Cal})^2 \sum_{t=1}^{n}(Ref_t - \overline{Ref})^2}} \quad (1)$$

2. 239: Perhaps replace the term validation here with testing?

We thank the reviewer for this observation. We agree that the term "testing" is more appropriate in this context, as the 20% subset was used for final model evaluation rather than model tuning or hyperparameter optimization. The terminology has been corrected throughout the manuscript.

In lines 258-259 of the revised manuscript: "We should note that for all the tests we used 80% of the entire dataset for training and 20% for testing, whereas all the measurements had a 2-min time resolution."

3. 252: Referring to Fig. 3, perhaps use LAB-calibrated on the y-axis in the left column of plots and ML-calibrated on the y-axis of the right matrix of plots.

This comment is well taken. We have now changed Figure 3 to include the new names on the axes

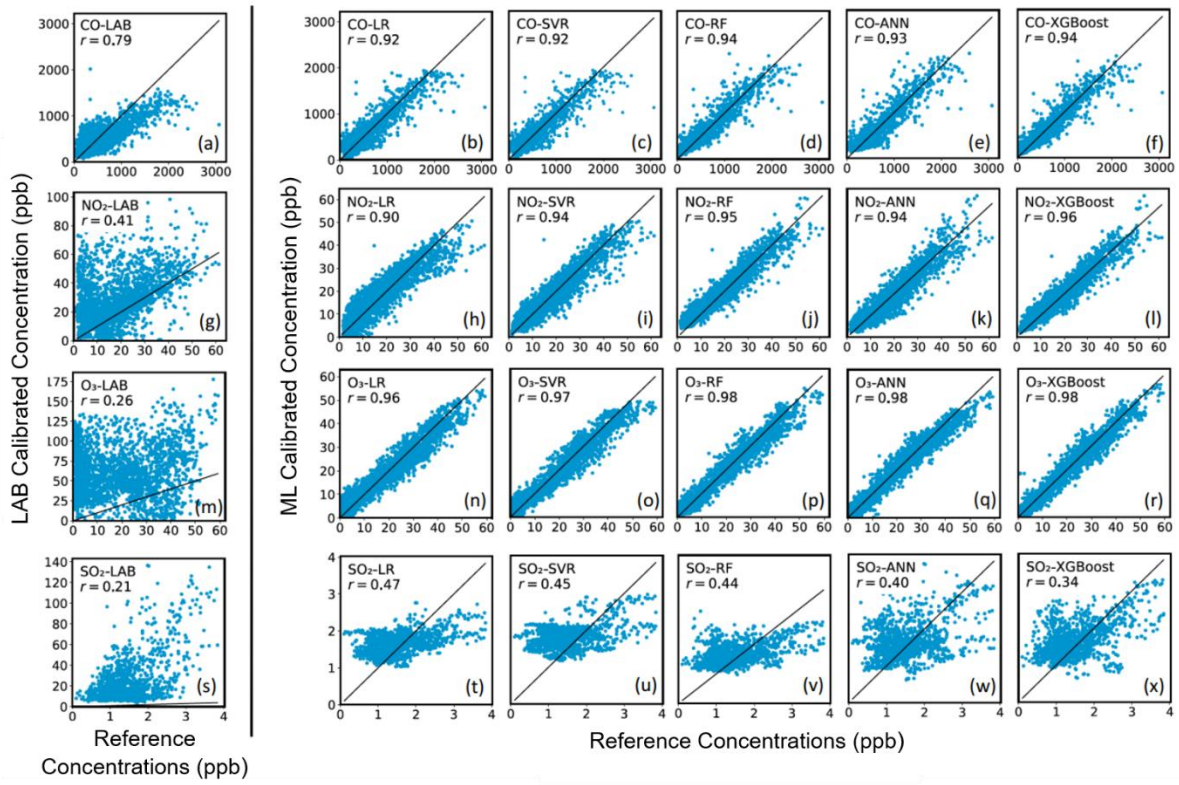

Figure 1: Correlation between calibrated LCS and reference concentration measurements for CO, NO2, O3 and SO2. The black dotted correspond to 1:1 relation.

4. 358: Referring to Fig. 8, the sentence should read: At least 80% for CO, 70% for NO2 and 50% for O3.

We have now corrected the text accordingly in lines 415-418 of the revised manuscript referring to Figure 9 (previously Figure 8) as follows: "It is also evident that under the continuous data sampling, a larger fraction of data (at least 80% for CO, 70% for NO_2 and 50% for O_3) is needed to train the models for the quality of calibrated data to meet the EU directive DQOs for indicative measurements (see Fig. 9d-f)."

5. 370: Referring to Fig. 9, the sentence should read: 22% for CO and O3 and 30% for NO2.

We thank the reviewer for pointing this out. We have edited the caption of both Figure 9 (prev. Fig. 8) and Figure 10 (prev. Fig. 9) accordingly: "The dotted lines in the bottom show the EU DQOs for indicative measurements, which are set to 22% for CO and O_3 and 30% for NO_2."

6. 392: 22% should be replaced by 22% for CO and O3 and 30% for NO2.

We have now edited the manuscript to correct this error in lines 455-458 that read as follows: "In those cases, if the sampling of the training data is collected over specific periods every month, but the entire training dataset is used to calibrate the measurements over 3 or 6 months, then the amount of data required for qualifying the LCSs for indicative measurements can reduce to 22 % for CO and O_3 and 30% for NO_2."

7. 497: Replace “measure- ment” with “measurement” and use the following for reference to the report: (NILU rapport 1/2020, in Norwegian), NILU, https://nilu.no/publikasjon/1809472/, 2020.

We have changed the citation to the correct version in lines 572-573:

“Walker, S.-E. and Schneider, P.: A study of the relative expanded uncertainty formula for comparing low-cost sensor and reference measurements, (NILU rapport 1/2020, in Norwegian) NILU, https://nilu.no/publikasjon/1809472/, 2020.”

References:

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.

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.

Citation: https://doi.org/10.5194/egusphere-2026-897-RC2