**Reviewer #1**

This manuscript provides a seven-step methodology for the calibration and quality assurance of low-cost air quality sensors. Thanks to the generalised nature of this method, it can be applied to a wide range of sensors and potentially be used as a standard calibration procedure. The data processing script was made publicly available which maximises the applicability of this method and the impact of this research.

The authors have pointed out current challenges in the use of low-cost sensors including the lack (or incomparability) of calibration procedures in many low-cost sensor application studies. They stress the need of a reliable and reproducible data calibration and post-processing method. This manuscript is an important step towards this aim and, therefore, a valuable contribution to the literature in this field as it has the potential to improve the data quality in future applications of low-cost sensors. The manuscript is well structured and clearly written.

My main suggestions to further improve the scientific quality of the manuscript are:

- Discuss the limitations of this method in more detail (Point 1)
- Add physical explanations of the found observations (Point 5)

*We thank the reviewer for their positive comments regarding this manuscript and their recognition of the need for such a standardized methodology in the calibration of low-cost sensors. Their comments are very helpful and have improved the quality of this manuscript. What follows are point by point responses to their main comments. All technical comments were accepted and changed in the manuscript, unless otherwise specified.*

**Specific comments**

**1. Please discuss the limitations of your calibration method in more detail (Point 2.1)**

- **Application range of calibrated sensors (indoor vs outdoor vs mobile)**

  You stressed the importance of calibrating the sensors under conditions that are similar to those under which they will be (or have been) operated during the experimental application. This needs to be considered when defining the application range of the sensors.

  Thanks to their silent operating conditions and small size, low-cost sensors are suited for indoor as well as mobile applications (e.g. wearable sensors for personal exposure assessment). However, if the calibration is conducted outdoors, the sensors might not be suited for such applications as the environmental conditions may differ significantly in these environments. Furthermore, mobile deployments would require further data cleaning and validation steps as rapidly changing environments may have an impact on the sensor performance (e.g. Alphasense Ltd., 2013).

*We thank the reviewer for this comment. Some text has been added to this section to clarify the study's focus on stationary field deployment and to highlight that different calibration considerations should be used depending on the environment in which the sensors will be deployed.*

*The added text reads, "In this study, we focus primarily on stationary field deployment of low-cost sensor systems. There are, however, other forms of deployment, including indoor and mobile, for which these criteria also apply. It is important to mention that there may be other considerations required in such alternative forms of deployment, e.g. more scrutinous data cleaning in mobile deployments due to impacts of rapidly changing environments on sensor performance."*

- **Sensor systems**

  As you have pointed out, low-cost sensors are often temperature and RH dependent as well as cross-sensitive to other pollutants. Therefore, it should be recommended to apply the presented calibration method to sensor systems (with additional sensors for T, RH and cross-sensitive gases) rather than individual sensors.

*We thank the reviewer for this comment. In the manuscript we have tried to be coherent in describing that this methodology is intended for use on small sensor systems, which often include individual sensors for temperature, relative humidity, and other cross-sensitivities. We have gone through the manuscript to ensure that this is consistently presented. Please see the track changes version of the manuscript for these changes.*

- **Data cleaning (Point 2.2.2)**

  In this step, point outliers are removed based on the assumption of a slowly changing airfield where peak exposures over a few seconds do not occur. However, such short-term (< 10 sec) emissions may occur in certain settings (e.g. traffic emissions of nearby passing vehicles, cigarette emissions of passengers etc.). One advantage of the high spatial and temporal resolution of low-cost sensors is that such peak exposures may be captured. The proposed method, however, excludes such events. Please include this argument when defining the application range of the sensors (Point 2.1).

*We thank the reviewer for this comment. Indeed, this is a significant challenge in data cleaning and unfortunately requires, in some cases, subjective assessment for an accurate determination to be reached. While it is possible that this data cleaning method removes some non-outlier measurements during peak emissions, it is equally possible that such events are indeed outliers due to technical sensor error. For this reason, we recommend that identified outliers be graphically compared with neighboring points to determine if their removal is justified.*

*With the optimized moving window and threshold identified in this study, for sensor system s71, a total of 58 outliers were detected from >500,000 data points. In this case individual assessment of each point's 'outlierness' was practical, but there may be cases where this is impractical. In such cases, we recommend a random subset of outliers be graphically assessed to determine the extent to which the data cleaning function is removing actual outliers. This is imperfect, but it is unlikely that a data cleaning method exists which can perfectly separate outliers from peak events. It is with such peak events that other tested methods such as the AutoRegressive Integrated Moving Average performed particularly poorly, identifying most peak events as outliers. If such events are expected due to the deployment environment, particular care in the evaluation of potential outliers should address this.*

*Clarification text has been added to section 2.2.2 regarding this and now reads, "The points flagged as outliers with this method were then graphically assessed against neighboring datapoints to prevent inadvertent removal of peak emission events. In other cases where assessing all outliers is impractical, it is recommended to do so with a random subset of outliers. Furthermore, if substantial short-term events are expected due to the deployment environment, such as during mobile measurements, a more thorough check of potential outliers should be done."*

**2. Line 115:** You state that, while demonstrated here with MOS, the proposed calibration method can equally be applied to electrochemical sensors. To strengthen this argument, please add a brief physical explanation, a reference, or experimental proof.

*We thank the reviewer for their comment. A brief physical explanation has been added in the text. While different in their design, both MOS and EC sensors produce a measure of voltage/resistance which varies in response to changing concentrations of gas-phase species, and hence can be calibrated using the same methodology. In more recent published work, we have successfully applied this methodology for the calibration of EC sensors, see Schmitz et al., (2021).*

*The added text reads "Furthermore, while it was applied here to sensor systems containing metal oxide LCS, this methodology is also equally as applicable to electrochemical LCS or photoionization detectors (PID), as these produce a similar measure of voltage that varies in response to changing concentrations of gas-phase species and have similar cross-sensitivities to temperature and relative humidity. It is not directly applicable for optical particle counters (OPC) for the measurement of particulate matter, as the transformation of the raw data into concentrations during calibration functions differently, though some of the principles discussed here are still relevant. For an application of this methodology to EC sensors, please see Schmitz et al., (2021)."*

**3. Line 221, line 240:** Please explain how you have determined the splitting ratio between training and validation period. How much differ you results when using other ratios?

*We thank the reviewer for their comment. In this case a standard splitting ratio of 75:25 for training and validation test sets was used, as this is common practice in model building. However, we have conducted a robustness cross-check with various splitting ratios and found that our results did not differ substantially when using other ratios. The training to validation splitting ratios tested were 90:10, 80:20, 75:25, 70:30, 60:40, and 50:50. For MLR, the median $R^2$ across all blocks in Step 5 for $NO_2$ models ranged between 0.78 – 0.83 for ambient T/RH and between 0.59-0.74 for internal T/RH. For $O_3$ models, the median $R^2$ ranged between 0.90 – 0.93 and 0.60 – 0.85 for ambient and internal T/RH, respectively. For RF, the median $R^2$ across all blocks for $NO_2$ ranged between 0.71 – 0.77 and 0.61 – 0.74 for ambient and internal T/RH, respectively. For $O_3$ the median $R^2$ ranged between 0.89 – 0.93 and 0.70 – 0.89 for ambient and internal T/RH, respectively. Given these results, we feel that using a splitting ratio of 75:25 is justified, as the results do not differ significantly based on this choice.*

*Text has been added in section 2.2.5 that reads, "A robustness cross-check with various splitting ratios was conducted and found that changing the splitting ratio did not significantly impact the results."*

**4. Table 6:** Please explain why you are using the medians and not the means of your statistical parameters. (whereas in Line 221 you were speaking about the average RMSE)

*We thank the reviewer for pointing out this discrepancy. The median was chosen instead of the mean as it is less susceptible to the influence of extreme values. The discrepancy is due to the*

*nature of the R function 'train()' from the 'caret' package used in Step 4, which provides only the mean RMSE during operation. In step 5 the calculations were done manually and thus the median was preferred. To correct this, the mean RMSE, MAE, and $R^2$ from Step 4 were replaced with manually calculated medians. This change is reflected in the text and in tables 2-5.*

**5.** While the manuscript nicely discusses the implications of a finding, it sometimes does not offer **physical explanations** for them:

- **Line 245:** "If the graphs showed instability across the various folds, Step 4 was repeated and a new model was selected for validation"

  What causes this instability and how can you ensure that the model stays stable under field conditions?

*We thank the reviewer for this comment. A sentence for clarification has been added to the text. We use instability to refer to major changes in $R^2$ and RMSE between folds in the model validation process. This is likely caused by differences in field conditions between the training and test folds. The best way to ensure the model remains stable under field conditions is with repeated co-location over longer time periods, in coordination with meteorological changes due to seasonality. The more training data available for calibration, the better the chances that the final model will be stable under field conditions.*

*The added text now reads, "In this case, instability refers to major differences in $R^2$ and RMSE between folds likely caused by differing field conditions among the training and test folds. If this is seen, it indicates that the model may be too sensitive to changes in field conditions."*

- **Section 3.4.4 (model selection):** Different relationships between the input variables were found for different models, e.g. an inverse temperature dependence for $NO_2$ was found for the best fitting MLR but no temperature dependence was found in the case of the best fitting RF. How can you explain this and what type of physical relationship (e.g. temperature dependence) would you expect?

*We thank the reviewer for this comment. A dependence on temperature was expected for the $NO_2$ models and was therefore included during the initial model selection process for both the MLR and RF models. However, the nature of this physical relationship was not clear, as the sensor specifications indicated that expected temperatures during field deployment would not impact the functioning of the MOS sensors. Rather, the dependence on temperature was expected due to the impact that temperature, as a proxy for insolation, has on daytime chemistry. An inverse relationship in this sense makes sense, as $NO_2$ is photolyzed in VOC-sensitive environments to produce $O_3$, which is normally the case in urban environments such as Berlin.*

*Following the update of the tables in Step 4 to reflect the median RMSE/MAE/$R^2$ instead of the mean for each model in response to the reviewer's previous comment, the best RF model for $NO_2$ was found to include T, which was previously not the case. All subsequent tables and graphs throughout the example were updated to reflect this change in the $NO_2$ RF model. Therefore, the reviewer's comment is partially answered, as there is in fact a temperature dependence in the RF model.*

*In the case of MLR, the final temperature dependence was determined to require an inverse transformation, whereas for RF, the relationships are equal in nature, as inverse, logarithmic, etc.*

*transformations do not affect the outcome of the RF model. This is principally due to different calculations that occur within the mechanisms of each model. For an RF model, this involves randomly choosing a variable by which to split the decision tree. This occurs at each node until no more splits are possible or the data are collected into final bins containing 5 data points. Therefore, any physical transformation of the data will not lead to a change in the calculations that occur in an RF model.*

*Text has been added to section 3.4.4, which now reads, "This is in line with what would be expected in urban environments, as T can be seen as a proxy for insolation, which causes the photolysis of $NO_2$."*

- The model performance was found to be higher when using the ambient environmental conditions (T and RH) as parameters (e.g**. Tables 6 and 7**). However, you pointed out in the discussion (**Line 619**) that the internal conditions are more representative for the operating conditions of the sensor. What are possible explanations for this observation?

*We thank the reviewer for this comment. This is indeed an interesting finding that is challenging to explain. The ambient T and RH would be expected to be more accurate models, as they are better representative of the conditions under which chemical processes occur that produce the concentrations of $NO_2$ and $O_3$ measured by the reference instruments. In this regard it makes sense that the model performance during validation was better with models trained using ambient T and RH than with internal. However, since the actual chemical reactions being measured are those that are taking place on the surface of the MOS, it seems that the internal T and RH better represent the conditions of the chemistry inside of the sensor system. The signal produced from the internal T and RH sensors is then used alongside the MOS signal in the models as markers for the chemistry that is occurring inside the sensor system at the time the same parcel of air reaches the sensor system and the reference instrument. However, that there are equally valid explanations for both outcomes warrants a closer investigation into these results. We feel that this would require much more detailed inspection of model predictions and would be outside the scope of this paper, which intends primarily to present and explain a methodology for the calibration of LCS. Future work will take a closer look at these results to determine why this occurs.*

*A sentence has been added to the discussion in line with this comment and another from Reviewer #2 that reads, "However, given that models using ambient data were more accurate during the validation step and significant differences between predictions of models trained with internal vs ambient T and RH were identified, these results require closer inspection, which should be the subject of future research."*

**6. Line 292:** Please specify "decent" and "good" agreement (e.g. with mean R2 & RMSE)

*Done. $R^2$ of these intercomparisons have been added to the text in addition to the reference to the supplemental information.*

**7. Line 327:** You deployed (at least) two low-cost sensors. Have you quantified the agreement between the two sensors? If so, add a small sentence here as it may be a strong argument why it is sufficient to only look at the data of one representative sensor. Perhaps summarise the performance of the second sensor briefly in the main text. How can you explain the non-linear response of sensor s72 (Figure S8)?

*We thank the reviewer for their comment. For this study, we use the two low-cost sensors primarily as examples of how to use the seven-step methodology and did not consider their intercomparison*

*as we felt it might distract from the main focus of the work. However, we have added graphs depicting the agreement between standardized raw LCS data of s71 and s72 during the 2 co-locations of the winter measurement campaign into the supplemental information (Figures S4 and S5). The Oxa and O3a MOS sensors of each sensor system are linearly related, but due to differences in sensor sensitivity, have different baselines. In the summer campaign, the relationship between the O3a sensors of s71 and s72 during co-location 2 is non-linear but returns to linear agreement in co-location 3 and in the winter campaign.*

*A reference to the added graphs in SI and a brief discussion of this point in the text was added [section 3.3] and reads, "To compare sensor performance between s71 and s72, an intercomparison of available co-location raw data was conducted for the oxidizing MOS (Oxa) and ozone MOS (O3a). During all co-locations, the sensors had a linear relationship and an $R^2 > 0.95$ (Figures S4 and S5). In only one instance was this not the case (co-location 2, O3a), where the $R^2$ was 0.59 and a deviation from linearity was detected. This relationship was linear in all other co-locations.".*

**8. Figure 8 (optional):** Adding histograms showing the overlap between colocation and experiment would make the Figure easier to comprehend and help to understand the flagging procedure.

*We thank the reviewer for this comment. We have decided not to include extra figures to the manuscript, as there are already very many. Instead, since the violin plots in Figure 4 would help understand the overlap between co-location and experimental data, we have added text that compares Figure 8 to Figure 4.*

*The added text in section 3.4.3 reads, "This shows the utility of comparing the results of Step 1 with the flags generated in Step 3."*

**9. Line 596:** Replace "for those who enjoy" with "to achieve"

*Done.*

**Technical comments**

**10.** Please use **subscripts** for $NO_2$ and $O_3$ and superscripts for $R^2$ throughout the document.

*Done.*

**11. Lines 93 and 96:** What means SVM? Do you mean SVR (support vector regression)?

*Yes, this was a mix-up between Support Vector Machines and Support Vector Regression. SVM has been changed to SVR to match earlier text in this section.*

**12. Line 149:** Delete "for use in statistical calibration" (the general quality of the final data is likely to be higher)

*Done.*

**13. Line 154 (Style, optional):** Replace "What follows in this section is a" with "This section provides a"

*Done.*

**14. Line 196:** How do you define the range of the colocation data? As the range between the minimum and maximum observations? (Or percentiles?)

*Yes, the range between minimum and maximum observations is meant here. This has been added to the text.*

**15. Line 219:** Please provide references for AIC and VI

*Done.*

**16. Line 263 (optional):** Perhaps add a sentence or reference explaining the term "smearing" as the audience might not be familiar with this practice.

*Done.*

**17. Line 295:** "more information 295 in section 3.2" – this is section 3.2

*This has been changed to "section 3.3", as is correct.*

**18. Table 1:** Is it correct that the sensor models for the reducing and the oxidising gases are identical? (SGX Sensortech MICS- 4514)

*Yes, this is correct. This sensor detects both reducing and oxidizing species.*

**19. Figure 2 (optional):** Adding a timeline with (rough) dates would help to comprehend the paragraph above quicker.

*Figure 2 has been updated and dates have been added to the timeline.*

**20. Figures 4 c, d; 6; 7 a, b; 10 etc:** Make sure that all axes have units (even if only arbitrary units).

*Done.*

**21. Figures 14 and 15 (optional):** Although you have already mentioned them in Tables 8 and 9, add the $R^2$ and RMSE values to the graphs to provide a comprehensive overview.

*Done.*

**22. Line 503:** "the reference instruments did not impact the predictive accuracy of the models and can therefore [in this case] be ignored as a potential interference" – can this be generalised for all sensors? If not, add "in this case"

*Done.*

**23. Line 508:** "The uncertainty between RF models and MLR models was fairly similar" - replace "between" with "of"

*Done.*

**Reference**

Alphasense Ltd. (2013). Alphasense Application Note 110: Environmental Changes: Temperature, Pressure, Humidity. Retrieved from www.alphasense.com, pages 1–6.

***References***

*Schmitz, S., et al. (2021). "Do new bike lanes impact air pollution exposure for cyclists?—a case study from Berlin." Environmental Research Letters 16(8). https://doi.org/10.1088/1748-9326/ac1379.*