Response to Comments from Editors and Referees

Information of previously submitted manuscript

- Manuscript number: egusphere-2023-1344
- Title: Evaluation of Calibration Performance of a Low-cost Particulate Matter Sensor Using Collocated and Distant NO_2
- Authors: Kabseok Ko, Ramesh R. Rao, and Seokheon Cho
- Status: Under Review

The authors would like to thank the referees for their careful reviews and valuable insight. We prepared our response to each of the editors' and referees' comments and revised our manuscript by reflecting all feedback.

Referee #2

The authors appreciate Referee #2's kind and valuable comments.

[Major Comments]

1. Very limited in scope and performance improvement: Given that the authors focused the comparison on calibration of a single sensor, the weight of "substantial contribution" of this manuscript falls on performance improvements of calibration models associated with that sensor. Unfortunately, the improvements on inclusion of NO2 are quite minimal. For example, in Tables 3 and 4, the best performances of models with and without NO2 are ~5% of each other. Does that qualify this work as "represent(ing) a substantial contribution to scientific progress" as is required by AMT? I disagree. I suggest that the authors conduct the analysis for the excluded sensor (sensor #8) that otherwise passes all checks, but was not included in the analysis for an unknown reason, as also pointed by reviewer 1.

(Response)

We appreciate the reviewer's perspective on the performance improvement for our proposed calibration models with the addition of NO_2 concentration as well as observation for one single PA-II unit. It may be that the improvements of around 5% observed in Tables 5 and 7 may not be substantial in absolute terms. However, it is important to consider the context and significance of these improvements.

First, even if calibration enhancement is modest in percentage, it can have practical implications in real-world application of low-cost $PM_{2.5}$ sensors, such as the PA-II units. A 5% improvement in low-cost $PM_{2.5}$ sensors can translate to more accurate and reliable measurements. The following three different feature vectors in Table 5 need to be addressed: feature vector #1, containing only $PM_{2.5;}$ feature vector #5 containing $PM_{2.5}$, temperature (T), and relative humidity (RH); and feature vector #16, consisting of a combination of $PM_{2.5}$, T, RH and NO_2.

We observed that the MLR-based calibration model considering feature vectors #1, #5, and #16 provides R^2 values of 0.731, 0.763, and 0.790, respectively. These values demonstrate that the MLR-based calibration model using $PM_{2.5}$, T, and RH results in an improvement of 4.4% in terms of R^2 compared to a calibration model only considering $PM_{2.5}$. Furthermore, the addition of NO_2 leads to an additional enhancement of 3.5% in comparison with the feature vector consisting of $PM_{2.5}$, T, and RH. Hence, feature vector #16 can achieve a calibration performance improvement of up to 8.1% over feature vector #1, which uses only $PM_{2.5}$ concentrations. We must also consider that the PA-II units measuring $PM_{2.5}$ are low-cost sensors and may therefore face constraints in their performance. In other words, an R^2 of 0.790 is not easily the calibration model for a low-cost sensor. Hua *et al.* showed that a generalized additive model (GAM) using four variables, such as $PM_{2.5}$, T, RH, and CO, brings about a 7.3% improvement of R^2 compared to a GAM using one variable of $PM_{2.5}$ under dry conditions (Hua et al. 2021). Therefore, the authors consider an R^2 of 0.790 and the calibration improvement of 8.1% achieved by considering T, RH, and NO_2 to be significant results for calibration performance, especially taken across all four seasons.

Second, the significance of our study's contribution does not lie solely with the magnitude of performance improvement. The study's impact can also be evaluated in terms of its methodology, its novelty, and its potential to inspire further research. Including NO_2 measured by an expensive FEM-based device for calibration models, and not a collocated low-cost sensor, might be a novel approach that opens up new possibilities for research in this area.

Regarding the suggestion to conduct the analysis using PA-II 8 rather than the PA-II 7 unit used in original manuscript, it is a valid point raised by the reviewer 1 and thus we added analysis results for PA-II 8. We studied three cases of the PA-II 8 unit and showed that reliable and consistent PA-II units, which contain two PMS 5003 sensors with high correlation to each other, demonstrate similar calibration performance. This implies that a proposed calibration method can be applied to reliable and consistent PA-II units generally. The three case studies are included as follows:

Case 1: Calibration model is learned with the measurements collected from PA-II 8 in 2018 and calibration performance for the trained model is evaluated using data measured from PA-II 8 in 2019.

Case 2: This is similar to Case 1, except that the calibration model is trained with data measured from PA-II 7 in 2018.

Case 3: The measurement data from PA-II 8 with collocated NO2 concentration in 2018 is used as a training dataset, while the data collected from PA-II 8 with either collocated NO_2 or distant NO_2 concentration in 2019 is used as a test dataset.

2. Choice of performance parameters and lack of uncertainty analysis: While the authors include three performance measures, despite considering models with multiple and changing number of variables, the authors fail to include the most important one: adjusted R2. The authors have clearly used the multiple R2 squared value to compare model fits; however, multiple R2 will increase on addition of even poorly correlated variables. I suggest that the authors report adjusted R2 results. Additionally, presentation of such calibration results would also benefit from an uncertainty analysis, and a key manuscript cited by the authors uses bootstrapping to do just that (Hua et al., 2021). I strongly recommend uncertainty (in terms of standard deviation) be considered when presenting performance metrics associated with such comparisons. The authors can then answer the question: are the distributions of performance parameters statistically significantly different with or without NO2? I would consider answering that question as a significant contribution.

(Response)

We appreciate the reviewer's perspective about the performance metric R^2. The adjusted R^2 is formulated as follows:

$$adj\ R^2 = \frac{(N-1)R^2-(M-1)}{N-M},$$

where N is the number of observations and M is the number of independent variables. To more accurately gauge the relationship between a PA-II unit and regulatory measurements over seasonality, we used whole-year data for training and test datasets, which are measured in 2018 and 2019, respectively. Our training and test datasets contained 7,198 and 7,621 samples, respectively. These numbers are much larger than the number of independent variables. Thus, from the equation of adjusted R^2 above, M has little effect on the value of adjusted R^2. In other words, adjusted R^2 is not significantly different from R^2 in our study. Table R1 shows the values of both R^2 and adjusted R^2 for an MLR-based calibration model on test datasets. The maximum difference between two values for every feature vector is 0.01.

Table R1. Comparison of R^2 and adjusted R^2 for MLR-based calibration model.

Feature Vector	R^2	Adjusted R^2	Feature Vector	R^2	Adjusted R^2
1	0.731	0.731	10	0.741	0.741
2	0.755	0.755	11	0.741	0.741
3	0.760	0.760	12	0.789	0.789
4	0.763	0.763	13	0.789	0.789

5	0.763	0.763	14	0.792	0.792
6	0.765	0.765	15	0.793	0.793
7	0.772	0.772	16	0.790	0.790
8	0.772	0.772	17	0.789	0.789
9	0.771	0.771	18	0.798	0.797
			19	0.796	0.796
			20	0.797	0.797
			21	0.797	0.797

We performed an uncertainty analysis of the MLR-based calibration model by using a bootstrapping technique on a test dataset. Table R2 shows statistics of uncertainty analysis for each feature vector and t-values between two feature vectors whose difference is the existence of NO_2. We selected 8 feature vectors with various independent variables to verify whether the addition of NO_2 affects the performance of our calibration model. The 4 feature vectors we considered are {$PM_{2.5}$}, {$PM_{2.5}$, T}, {$PM_{2.5}$, RH}, and {$PM_{2.5}$, T, RH}. We also added NO_2 to create four other feature vectors, {$PM_{2.5}$, NO_2}, {$PM_{2.5}$, T, NO_2}, {$PM_{2.5}$, RH, NO_2}, and {$PM_{2.5}$, T, RH, NO_2}. We generated 1,000 test sets using a bootstrapping technique with replacement. We evaluated mean and standard deviation values of RSME calculated over 1,000 test sets for each feature vector. In addition, we applied a t-test to verify the effectiveness of adding NO_2 to each feature vector. Consideration of NO_2 additionally reduces mean values of RMSE for all 4 feature vectors. Contrary to mean value, standard deviation of RMSE for every feature vector increases slightly with the addition of NO_2.

We evaluated t-value for the mean values of RMSE for two feature vectors, with and without NO_2; for example, the t-value between {$PM_{2.5}$} and {$PM_{2.5}$, NO_2}. Hence, we can evaluate 4 t-values. Degree of Freedom (DoF) is 1,998, so the relevant p-values are much less than 0.00001. Therefore, the difference in the mean RMSE values of the NO2–included and NO2-excluded groups is significant.

From these results, we can conclude that the performance of the MLR-based calibration model can be enhanced with consideration of NO_2 concentrations.

Table R2. Statistics of uncertainty analysis to selected feature vectors and t-values.

Feature Vector	Mean of RMSE	Std. Dev. of RMSE	Feature Vector	Mean of RMSE	Std. Dev. of RMSE	t-value	DoF
{$PM_{2.5}$}	4.5095	0.1026	{$PM_{2.5}$, NO_2}	4.4202	0.1037	19.3580	1,998

{$PM_{2.5}$, T}	4.3084	0.1000	{$PM_{2.5}$, T, NO_2}	3.9979	0.1173	63.7008	1,998
{$PM_{2.5}$, RH}	4.2598	0.0995	{$PM_{2.5}$, RH, NO_2}	4.1548	0.1074	22.6792	1,998
{$PM_{2.5}$, T, RH}	4.2387	0.1050	{$PM_{2.5}$, T, RH, NO_2}	3.9865	0.1156	51.0686	1,998

3. Poor presentation: Large sections of the manuscript are unnecessarily detailed, and could be moved into tabular form whether in the main manuscript or the supplement. These include large portion of the lines 198-222 and 233-246 which are two representative examples. Additional examples include lists of variables shown in text format, which is laborious to read or keep track of (e.g., Lines 355-364). Additionally, key details of the authors' methodology such as performance metrics and intercomparison exercises are dispersed throughout the Results section (Sect. 3.1 to 3.6). I suggest that authors separate the methods portions of these results and discuss them in a separate subsection under Methods called "Instrument intercomparisons".

(Response)

We appreciate the reviewer's suggestion to streamline our presentation by consolidating certain portions into tabular format and creating a dedicated subsection within the Methods section. Hence, we implemented these changes to improve the overall clarity and accessibility of our work. We carefully reorganized the manuscript to enhance readability and ensure that we present key methodological details more cohesively.

1) We restructured Sections 2 and 3 as follows:

2. Methods
2.1 Measurement data
2.1.1 PurpleAir PA-II units
2.1.2 Air quality measurement data from EPA
2.1.3 Selection of PA-II units and reference monitoring sites
  - Note: We merged Subsection 2.3 with Subsection 3.1
2.1.4 Data preprocessing of PA-II units

2.2 Instrument intercomparisons

- Note: We merged Subsections 3.2 and 3.3. We also eliminated redundancy by creating a table for summary statistics of daily and hourly $PM_{2.5}$ measurement data, and removing detailed explanations of maximum, minimum, mean, and standard deviations of various measurement data.

2.3 Feature vector selection for calibration models

- We merged Subsections 3.3 and 3.4. We then shifted the merged text into this subsection and simplified the contents for greater cohesion from the viewpoint of feature selections.

2.4 Calibration models
2.4.1 Multiple Linear Regression (MLR)
2.4.2 Random Forest (FR)

2.5 Performance evaluation metrics

3. Results and discussion
3.1 Calibration performance
3.1.1 MLR-based calibration model
3.1.2 RF-based calibration model

3.2 Effect of distant NO_2 on calibration performance

3.3 Applicability of other PA-II Units

3.4 Effect of training period

3.5 Uncertainty analysis

2) We added two Tables describing the selected feature vectors used in analyzing our MLR- and RF-based calibration models, which had previously been written as text in our original manuscript. In the original manuscript, this included Lines 311-318 and 354-362.

3) We added a subsection on performance evaluation metrics to improve readability.

[Minor Comments]

1. Lines 123-139 The authors start off with a large dataset but remove data points using some filters. I suggest that the authors add a supplementary table showing how many data points were removed at each step.

(Response)

We included the following information in the supplementary document regarding the number of data points processed for each step of pre-processing.

Table R1 Number of data points processed for each step of pre-processing.

Applied Method	Number of data points
Original (01/01/2018 – 12/31/2019)	703,369
Remove data with N/A	703,369
Valid data with 0<=Temperature<=200	703,368
Valid data with 0<=RH<=100	703,339
Valid data with $PM_{2.5}$ <= 2,000	703,339
Averaging data for hourly $PM_{2.5}$	17,507
Hourly Averaging with sufficient data points	17,198
Comparison of PMS 5003 A and B using SPE	16,966

2. Lines 412-413 and Lines 287-290 The language used by the authors is unclear. I suggest either expanding on these sentences or rephrasing them so that the point is made clearly.

(Response)

We rewrote Lines 410-413 as follows:

To evaluate the usefulness of distant NO_2 measurements on the calibration of a low-cost PM sensor, we used NO_2 data measured from monitoring sites near the PA-II 7 unit as a test dataset, rather than data from the collocated Rubidoux site. When we trained calibration models with the measurements from the PA-II 7 unit over 2018, we used highly accurate NO_2 concentrations measured by FEM instruments at the Rubidoux site. Subsequently, to verify the trained calibration models, we utilized a separate test dataset featuring distant NO_2 measurements taken by FEM instruments at sites 06-065-8005 and 06-071-0027. We considered this scenario to evaluate our proposed calibration models, previously trained with collocated NO_2 concentrations and distant NO_2 concentrations, when collocated NO_2 measurements cannot be collected.

We rewrote Lines 287-290 as follows:

These remarkable results suggest that NO_2 is generally a key factor that can improve the performance of PA-II units over a year, even though the enhancement by NO_2 does not meet the values of 0.7 of R^2 and 3.5 $\mu g/m^3$ of RMSE during certain months, such as July 2018, August 2019, and October 2019.