Response to Editor's and Reviewers' Comments

Information of previously submitted manuscript

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

The authors would like to thank the referees for their careful reviews and insightful comments. We prepared responses for each comment from the editors and referees and revised our manuscript to reflect feedback.

Reviewer: 1

The authors appreciate Reviewer #1's kind and valuable comments.

[Major Comments]

1. Your article idea is interesting and relevant. However, the data set is very small (two PA-II units, when only one is used in most of the analysis). Would it be possible to add other locations/states to make your method validation more robust?

(Response)

We sought to evaluate the applicability of our proposed methods across diverse locations. We used to download the PM2.5 concentrations from the open platform operated by PurpleAir Inc. that manufactures the PA-II units. However, we encountered their recent policy change, which now imposes charges for data downloads and then makes us challenging to obtain a sufficient volume of data.

Hence, we searched for the data we had downloaded from PurpleAir open platform. Fortunately, we identified one single suitable PA-II unit that meets our criteria, which a PA-II unit is collocated to an EPA monitor measuring both $PM_{2.5}$ and NO_2 (used as collocated NO_2), as well as there exists at least other EPA monitor measuring NO_2 (used as nearby NO_2) not being far away from the collocated EAP monitoring station. This particular PA-II unit, named SACRAMENTO, is located in Sacramento, CA, USA. It is installed near an EPA monitoring site, designated as 06-067-0010. Another EPA monitoring site to measure NO_2, 06-067-0015, is located in the vicinity of the collocated EAP station.

We found around 13 months of data from December 2020 to December 2021 that we had collected from the PA-II named SACRAMENTO earlier. In our manuscript, to utilize and scrutinize the seasonality patterns of $PM_{2.5}$ concentrations, we divided the two-year dataset into a training dataset and test dataset consisting of data collected in 2018 and 2019, respectively. We cannot apply our previous approach onto the SACRAMENTO PA-II unit due to the limited data we had stored. However, we could keep considering seasonality for reliable performance by employing stratified random sampling to partition the dataset. The stratified strategy was applied on a monthly basis with the split ratio of 80:20 for training and testing. We assessed

the calibration performance of the SACRAMENTO PA-II unit using the MLR algorithm and the corresponding results are presented in Table R1.

When NO_2 is not included in feature vectors, the best performance is achieved for feature vector #9. Also, in case of using NO2, all feature combination sets from #12 to #20 except #15 result in higher R^2 as well as lower RMSE and MAE than feature vector #9. This is the same result as we showed with two PA-II units in our manuscript.

Table R1. Calibration results of hourly $PM_{2.5}$ concentrations measured from the SACRAMENTO PA-II using MLR-based calibration model.

NO_2 not included				NO_2 included			
Feature Vector	R^2	RMSE	MAE	Feature Vector	R^2	RMSE	MAE
1	0.743	4.844	3.070	10	0.745	4.824	3.018
2	0.764	4.646	3.032	11	0.773	4.552	2.909
3	0.769	4.597	2.982	12	0.790	4.385	2.790
4	0.783	4.458	2.874	13	0.791	4.372	2.767
5	0.769	4.590	2.986	14	0.791	4.371	2.768
6	0.776	4.520	2.943	15	0.776	4.527	2.903
7	0.783	4.452	2.875	16	0.791	4.371	2.787
8	0.784	4.441	2.859	17	0.792	4.359	2.767
9	0.784	4.441	2.857	18	0.792	4.357	2.767
				19	0.793	4.346	2.757
				20	0.793	4.354	2.765

We tried to evaluate the performance of MLR-based calibration model to get the comparison results using both collocated NO_2 (06-067-0010) and distant NO_2 (06-067-0015) measurements. However, the site we need to refer to as distant NO_2 has discontinued measuring NO_2 since Aug. 2021. Hence, we can gather and use only around 7 months of NO_2 data, which precludes a fair and appropriate performance comparison between collocated NO_2 and distant NO_2 as we provided in our manuscript.

2. Line 77: check spelling of "gase".

(Response)

We had already used “gases” on Line 77 as you suggested, so we don’t change the word.

3. Sections 2.1.2, 2.1.3, and 2.1.4:
- What is the relevance of explaining POC and other EPA minutia?.
- Why give so much attention to sections 2.1.2 and 2.1.3 but not detail the test and training data separation strategy on 2.1.4? - - I recommend using the level of detail used in section 2.2

(Response)

We modified on line 103 (removed sentences from line 104 to line 112) in subsection 2.1.2 as follows:

Outdoor air quality data collected from across the U.S. is publicly available through the U.S. Environmental Protection Agency (EPA) website (https://epa.gov/outdoor-air-quality-data). ~~The EPA has a description file for monitors, which includes state code, county code, site number, location (latitude and longitude), parameter code, parameter occurrence code (POC), and last method. A combination of state code, county code, and site number can uniquely identify a monitoring site. For example, a monitoring station located at Bakersfield, CA has a state code of 06, a county code of 029, and a site number of 0014. The parameter code is an air quality system (AQS) code corresponding to the parameter measured by a monitor. For example, parameters regarding PM$_{2.5}$ and NO$_2$ are 88101 and 42602, respectively. A POC is used to identify an instrument among multiple ones with the same parameter code at a site. For example, two FRM instruments with a parameter of 88101 at the Bakersfield site are used to measure daily PM$_{2.5}$ concentrations and are identified with POC 1 and 2. The last method descriptor describes the measurement scheme used by the monitor for its most recent sample.~~

We added a new paragraph about test and training data separation strategy after line 170 in subsection 2.1.4 of the original manuscript as follows:

The period of valid measurement data collected from the PA-II units we selected is 24 months, such as from Jan. 2018 to Dec. 2019. The measurement data in the years 2018 and 2019 from the two-year dataset were used for training and testing for our calibration models, respectively. The reason why we split the two-year dataset at a 1:1 ratio is that PM2.5 as well as the other environmental parameters, such as temperature and relative humidity, which we considered for calibration models, have a seasonal pattern. Also, we used whole-year dataset for training to learn the relationship between PA-II and regulatory measurement over seasonality and thus enhance the performance of the calibration models over all 4 seasons.

4. Line 152: "...measure 'for' obtaining..."? Are you sure it is 'for' and not 'by'?

(Response)

We changed “for” to “by”.

Reviewer: 2

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

[Minor Comments] Comments to the Corresponding Author

1. Line 62: Please change to "two months have shown good correlation"

(Response)

We corrected "have shown" as you suggested.

2. Line 322, 367: Pleas unitalicize ug/m^3

(Response)

We changed italic ug/m^3 into unitalic ug/m^3.