Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC1
https://doi.org/10.5194/amt-2021-154-RC1, 2021

Comment on amt-2021-154

Anonymous Referee #1

Referee comment on "Evaluation methods for low-cost particulate matter sensors" by Jeffrey K. Bean, Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-154-RC1, 2021

This paper presents and discusses an improved air sensor evaluation method the prediction interval. Overall, this appears to be a valuable addition to statistics more commonly used to evaluate air sensor performance. However, the methods are lacking to understand how to implement this method in future work. Some of the discussion and conclusions are also lacking.

Major:

It seems like this paper would be more helpful to the air sensor community if the sensor type was provided.

Line 98: I think this method is super helpful to think about determining how to pick a threshold to remove outliers based on precision moving forward! Did you consider whether percentage and absolute concentration criteria would be more useful than just 50 ug/m3 by itself (as used in https://doi.org/10.5194/amt-14-4617-2021)? What was the range of concentrations experienced by the sensors during this period (as reported by the sensors since that is the criteria you are using to exclude)?

Figure 1: How much does removing disagreeing datapoints clean up the bias? Or does it only clean up the R2 as shown in this figure? It would be helpful to include a 3rd panel on Fig 1 that would be average concentration reported by the sensor versus the allowed difference and also the average of the FEM as well.

Line 120: How was no correlation defined? How did you decide which parameters to include in the figures and which not to? What are the correlations of the parameters in the

figures? It would be helpful to also show the correlation between PM10 or PM10-2.5 and the error since as you say in the paper particle size can influence the accuracy.

Figure 2 seems overly complicated and challenging to interpret. It seems like 2A would be easier to interpret if it just showed sensor 1 versus sensor 2 (not the bias) and then if figure 2d showed sensor 1 versus the BAM concentration. I think this would be valuable even if you wanted to add a figure showing the basic plots and then a second figure showing all the bias plots if you feel you need both.

Prediction Interval: I don't understand how to calculate this based on the information you have included in the text. Please cite additional references and include the equations in the SI if needed. Please include a citation for the R package if one was used. What are the 3 diagonal lines shown on Figure 4? Can you label them?

Line 186: "Through examination it was found that residual trends were best eliminated by raising both the sensor and reference data to the 0.4 power." More scientific explanation is needed. How would we repeat this method in the future would others always just use 0.4 since that is what you "examined" and found or would they need to examine their datasets and come to different conclusions? Your figure doesn't show transformed data so it is unclear how you would determine this. Please provide additional explanation.

Past work has shown that sensors may respond nonlinearly at high concentrations (**https://doi.org/10.1111/ina.12621**, https://doi.org/10.1016/j.envpol.2018.11.065). Would a prediction interval still be appropriate in a case like this? Would you be able to remove the residual trends in a dataset like this?

Line 189: "This did not change the outcome significantly". Please define what you mean by this. How did you decide to use 70 ug/m3 as your split to have equal number of data points above and below? Does Figure 4 show the full dataset or the subset? How important is having a well-balanced dataset to getting accurate results?

The need to balance a dataset seems like a limitation of this method as compared to R2 or RMSE that has not been discussed.

Figure 4: I would recommend coloring the arrows uniquely and including them next to the text in the bottom right corner as a legend so the figure can be more quickly interpreted.

Figure 5: Is this only for T640 data? If so, why not also show BAM data as shown in the other figure?

Figure 5: Did you consider whether uncertainty as a % would be more stable?

Line 219: "It allows for better comparison between sensors, as the evaluation results are not biased by the range of concentrations observed during evaluation." I think more explanation is needed here. It doesn't seem to me this is one of the findings of your analysis since you only showed the results from one dataset covering the full range. It might be interesting to show another data subset with a different concentration range to understand how the range doesn't impact the results but maybe there is another way to explain since as I said above I don't really understand how you are calculating the prediction interval.

Lines 306-308: "Two of the most popular evaluation metrics, R2 and RMSE, can be influenced by averaging time, choice of reference instrument, and the range of concentrations observed (see Fig. 3). This study shows how a prediction interval can be used as a more statistically thorough evaluation tool." Figure 5 shows that prediction interval is also influenced by averaging time. Are you saying it isn't influenced by reference instrument? If so, I think you need more results to show that. Overall, this statement seems misleading.

Figure 6: With much of the data below 5 ug/m3 did you consider how LOD of the sensor and reference influence your results?

Figure 2 seems to show that the bias is much more variable at high RH. How can you take that into account using prediction interval?

Did you consider how the precision of the sensor influences prediction interval? I'm assuming that Figure 4 is for all the sensors but if it is for a group of sensors that would be helpful to clarify in the caption/text.

It would be helpful to add the prediction interval for all of the sensors you tested not just the best sensor so that readers could compare the R2/RMSE/PI more closely across devices and understand how they could use this in the future.

Have you thought about how you could report this PI as something more easily to compare across papers than a plot (which may have different axis labels etc.)? For example, fitting a function or reporting the 95% uncertainty at various AQI breakpoints, etc?

This work is missing relevant citations. Examples: Giordano 2021 calibration review paper https://doi.org/10.1016/j.jaerosci.2021.105833, Zheng 2018 similar discussion of

averaging interval and the precision of the reference https://doi.org/10.5194/amt-11-4823-2018, some others included in my other responses.

Minor:

Figure 1: Could you include the averaging interval you are using for exclusion in the figure caption?

Line 45: “The root of a calibration for low-cost particulate matter sensors is simple: sensors and reference instruments measure the same mass of air for a period and then adjustments are made to better align sensor measurements.”. I’m not sure “root” is the clearest way to express this.

Line 255: US EPA recommends at least 30 days for their PM2.5 sensor evaluations. https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=542106&Lab=CEMM