Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC3
https://doi.org/10.5194/amt-2021-425-RC3, 2022

Comment on amt-2021-425

Anonymous Referee #3

Referee comment on "Development and evaluation of correction models for a low-cost fine particulate matter monitor" by Brayden Nilson et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-425-RC3, 2022

This study tested several calibration models, both original and previously published, on PurpleAir PM2.5 sensors collocated throughout North America. More focused on a health-related context, they used AQHI+ levels as a comparison between the PurpleAir sensors and FEM monitors, in addition to targeting improved performance at high AQHI+ values. Additionally, by pooling together data from many geographic locations they are able to suggest a possible model for use when a collocation is not possible.

Specific Comments:

Line 123: Please state the reasoning behind your choice of 5ug/m^3 as the absolute error cut-off for identifying failures in either sensor. Is this a recommendation from PA?

Line 138: "The final set of colocation sites (47 in total) were then selected as those with at least half a year (4380 hours) of valid data from both PA and FEM and a minimum correlation of 50% for all valid hourly observations over the period of record." What is the reasoning for setting the minimum correlation to 50%? While this would remove any non-collocated sensors wouldn't this also possibly remove any poorly performing collocated sensors?

Line 148: "A temperature term was also tested; however, its impact was found to be minimal." Does this statement refer to the pooled together training dataset? Was this ever tested for individual sites?

Line 159: In equation 3 please clarify that the correction factors of a and b for PM2.5 < x are different than the factors a and b for x< PM2.5< x2.

Line 191: How were the training and testing sites divided up? Rather than dividing by site did you consider dividing by time (randomly dividing to ensure similar conditions between testing and training datasets)? Please clarify the reasoning between choosing 32 training sites and 15 testing sites (~2/3 training).

Line 246: "Further comparisons were only made on Models 1, 2, 7, and 8 as they showed the best performance here." It would be informative to include the results from the other models in a supplementary information section.

Line 294: "The concentrations of PM2.5 reported from the PA monitors were biased high compared to the FEM monitors at most colocation
sites, especially for the lower concentration range." While you evaluated model bias at different PM2.5 concentrations did you consider looking at the bias over a RH range?

Line 324: For scenarios where testing models on individual locations is not an option, such as applying a correction in an area without a nearby PA-FEM colocation site, we recommend using our Model 2." Rather than use a model that has been trained on a variety of locations/conditions and therefore is pretty generalized, would it not be more prudent to use a model that has been trained on conditions similar to those you expect to encounter?