

Atmos. Meas. Tech. Discuss., referee comment RC2
<https://doi.org/10.5194/amt-2021-425-RC2>, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on amt-2021-425

Anonymous Referee #2

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-RC2>, 2022

This is an interesting paper looking at a North American correction for PurpleAir sensors. It builds to an extent on similar past work (Barkjohn 2021 <https://doi.org/10.5194/amt-14-4617-2021>) by evaluating higher time resolution data (hourly averages), considering a different correction and limiting the range of RH from 30-70%, including sensors from Canada, and analyzing in terms of the 1-hr Canadian AQHI. The paper is missing a summary of the hourly data used and more details on the QC methods applied. I have a number of other specific comments listed below.

Specific:

- It seems like you are missing a summary of the dataset. What is the range of hourly concentrations? How many points are there per AQHI category? Did you see nonlinearity in the high concentration data (<https://doi.org/10.1111/ina.12621>)?
- Model 5 should also be applied to the cf_atm data. This is the AQ&U equation from the PurpleAir map. Although a cf isn't listed on the PurpleAir map you can check which cf it is by checking the calculation at a high concentration site (since we know the "raw" outdoor data is cf_atm) (I did this today Feb 8th and it seems to still be applied to the cf_atm data). The Kelly paper was published in 2017 long before PurpleAir flipped the labels to reflect Plantower's labels. Further confirmation: cf_atm is used in this equation in this recent study: <https://amt.copernicus.org/articles/14/4617/2021/>
- Did you consider whether RH and PM₅ are correlated at hourly averages in your dataset? In multiple linear regression independent variables should be independent.
- Lines 75-77: The "CF 1" data were found to correlate better with FEM observations in our data set. A recent study has developed a particle count correction factor using US based sites which shows promise however that was not tested here (Wallace et al., 2021)." This belongs in results/discussion not introduction. Sharing the correlations for the CF_1 vs CF_atm data would be helpful.
- Lines 85-87: "In addition, we and others have found the PA temperature observations to be biased high (and in turn RH biased low) because of internal heat produced by the electronics as well as incoming solar radiation (which has varying impacts depending on

the physical location and placement of each monitor)" your results should go in the results & discussion. It would be good to include citations here of past work showing warmer and dryer (e.g. <https://doi.org/10.3390/s20174796>, <https://doi.org/10.1080/02786826.2019.1623863>)

- Update the Barkjohn 2020 AMTD preprint article to the final published AMT article.
- Lines 104-109: Can you clarify did you download all nearby sensors or only sensors that were labeled as outdoor sensors?
- Line 125: How many months were flagged as invalid for temperature and RH? And what is the fraction of months removed from sensors where this is an issue? It would be good to understand the break down by sensor to understand are these sensors that were labeled as outdoor but are always running indoors or are they just being brought indoors for a month and then returned to the outdoors? Did you check whether this worked correctly with sensors that were marked by the user as indoor sensors?
- Line 128: Would you need to provide someone with a cut off level for a Hampel identifier (or other input variable) for them to recreate this method? Did you use a software package to complete this analysis?
- How much data was removed with each QC step?
- Line 129: How many months had <72 hours of data?
- Line 131: How many hours or what percent were removed by manual inspection?
- Line 133: How many sensors had <2 months of data?
- Line 128: Can you clarify what you mean by "3 units"
- What software was used for this analysis?
- Line 132: Why not use the QC'd data from AQS?
- "Sites with multiple colocated PA monitors were averaged together to produce a single data record for each site after flagging and removing any invalid data." Did you consider the precision of PurpleAir sensors in places where multiple sensors were close by?
- Line 149-150: Could you clarify what you mean here? "Piecewise models which were built starting from the second segment tended to perform better in the mid-range PM2.5 concentrations than those built starting from the first segment."
- ~Line 175: The US EPA performance targets for PM2.5 sensors may be valuable for this work ("Performance Testing Protocols, Metrics, and Target Values for Fine Particulate Matter Air Sensors: Use in Ambient, Outdoor, Fixed Site, Non-Regulatory Supplemental and Informational Monitoring Applications"https://cfpub.epa.gov/si/si_public_record_Report.cfm?dirEntryId=350785&Lab=CEMM)
- Line 175: Could you provide calculations for these metrics here or in the SI? (since there has been some discrepancy on calculation method especially for RMSE in the sensor literature)
- Figure 2 only shows the site medians. Could you also add a figure showing the full dataset of hourly points? Maybe a scatter plot of hourly FEM PM2.5 vs PurpleAir PM2.5?
- Table 1: Did you try looking for these in AirNow Tech by method_code? (https://aqs.epa.gov/aqswb/documents/codetables/methods_all.html)
- Figure 3: Did you consider whether just the 3 known types were significantly different?
- Table 2: It may be clearer to use letters to represent the sources so that they are not easily confused with the model numbers.
- The conclusion would be easier to interpret if when referring to the models by number it also described the model type.
- Line 305: "It should be noted that the average performance across the testing sites and over time was evaluated here; performance at colocation sites and across time was not the same." Did you consider whether there were regional or factors that could explain this?