Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., author comment AC2
https://doi.org/10.5194/amt-2021-425-AC2, 2022

Reply on RC2

Brayden Nilson et al.

Author 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-AC2, 2022

Thank you for the time you put into reviewing our manuscript and the very useful and helpful feedback which has led to improvements in the paper. Please see our following responses and proposed alterations which we believe will resolve your individual comments.

Comment 1:

- It seems like you a 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)?

Response 1:

We added the following text and a new figure after Figure 1 in the results section.

"Hourly concentrations of PM2.5 ranged between 0 – 837 µg m-3 and 0 – 986 ug/m3 across all sites during this period for PA and FEM monitors, respectively (Figure 2). PA monitors at most sites tended to be within a factor of 2 of FEM, typically biased higher. For most sites this bias appears to be linear as concentrations increase. PA PM2.5 concentrations across all sites were categorised as "Low AQHI+" (0 – 30 ug/m3) for 91.1% of observations, "Moderate AQHI+" (30 – 60 ug/m3) for 7.7%, "High AQHI+" (60 – 100 ug/m3) for 0.7% and "Very High AQHI+" (100+ ug/m3) for 0.6% of observations. In the same order for the FEM monitors at all sites: 97.5%, 1.9%, 0.3% and 0.4% of observations were in the four AQHI+ categories."

Comment 2:

- 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/

Response 2:

Changed "Kelly et al. (2019)" to "Kelly et al. (2017)" in Table 2 and swapped CF_1 for CF_atm. Regenerated figures/results with this adjusted model. This model now performs comparably to the Barkjohn et al. 2020 model. Adjusted the results and conclusions to reflect this (Model 5 now discussed in favour of Model 8).

Comment 2:

- Did you consider whether RH and PM5 are correlated at hourly averages in your dataset? In multiple linear regression independent variables should be independent.

Response 2:

Yes, they were not strongly correlated at most locations. site-wide Pearsons correlation coefficients ranged from -0.13 to 0.48 with a median correlation of 0.18.

Comment 3:

- 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.

Response 3:

Removed discussion of Wallace et al. in response to another comment from Referee 1.

Comment 4:

- 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)

Response 4:

We altered the phrasing here so it was more a statement of potential error and less a comment on results we have noticed as considering biases in T and RH were not a purpose of our study.

Comment 5:

- Update the Barkjohn 2020 AMTD preprint article to the final published AMT article.

Response 5:

We updated the reference and the intext citations.

Comment 6:

- Lines 104-109: Can you clarify did you download all nearby sensors or only sensors

that were labeled as outdoor sensors?

Response 6:

We clarified in the revised text that we only selected sensors labelled as outdoor sensors.

Comment 7:

- Line 125: a) 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? b) Did you check whether this worked correctly with sensors that were marked by the user as indoor sensors?

Response 7:

- a) Added the following paragraph to the beginning of the results detailing the % of data removed by each QA/QC step and the number of sites affected by each.

"The colocation site selection metric we used detected 86 potential colocation sites during this period in Canada and the United States. All sites had missing data, five sites had PA sensors with manually flagged invalid data, 65 had months where the temperature or RH were deemed too invariable to be outdoors, 67 had hours flagged invalid from differences between the A and B sensors within the PAs, and six sites had monitors with less than two months of valid data. Across all of these sites, 40.1% of the PM2.5 observations were missing (either from the FEM or a PA), <0.0001% were manually flagged as invalid, 3% were flagged as months where the PA was likely indoors, 2.3% were flagged by our PA A/B sensor comparison, and 1.3% were removed from PA monitors with less than two months of valid data."

- b) We did not - but we confirmed it worked at sites we knew were outdoors.

Comment 8:

- 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?

Response 8:

Added to the manuscript that the standard cutoff was used (outliers exist >3 median absolute differences from the median).

See response 11

Comment 9:

- 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?

Response 9:

See response 7

Comment 10:

- Line 128: Can you clarify what you mean by "3 units"

Response 10:

Added "(ie. 3 °C for temperature or 3% for RH)" to this sentence.

Comment 11:

- What software was used for this analysis?

Response 11:

Added a statement of the software used (R) plus several key R packages.

Comment 12:

- Line 132: Why not use the QC'd data from AQS?

Response 12:

Our data flows were setup for Airnow already, as the Canadian monitors would not be available from the AQS site.

Comment 13:

- "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?

Response 13:

They correlate well with each other and tend to be within a few μg m^{-3} from my experience, but no we did not analyse that specifically for this dataset.

Comment 14:

- 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."

Response 14:

Normally piecewise models are built starting from the first segment, then you fit the next segment from the end of the first, and so on. Since we wanted to focus on the moderate-high ranges we tried fitting this middle range first, then fitting the first and last segments to the ends of the middle.

We clarified this in the revised text (adding the underlined text in the following sentence)

"Piecewise models which were built starting from the second segment (fitting the mid-range data first) tended to perform better in the mid-range PM2.5 concentrations than those built starting from the first segment (fitting the low-range data first)"

Comment 15:

- 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)

Response 15:

We added the following discussion of this to the paper in section 2.4.

"Duvall et al. (2021) outline several key metrics to consider for small sensor performance: precision, bias and error, linearity, effects of RH and temperature, sensor drift, and accuracy at high concentrations. Evaluating precision is not viable in this study given that many sites only had a single PA installed. We will evaluate bias, error, and linearity through our analysis, as well as the effects of RH. We found temperature impacts to be minimal for our dataset, especially when the impacts of RH were already considered. Sensor drift is outside of the scope of our study, and accuracy at high concentrations is less of a concern given our use of the AQHI+ scale and focusing on the moderate to high concentrations."

Comment 16:

- 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)

Response 16:

We added these metrics to the revised manuscript in section 2.4 (Eq. 6 – 11).

Comment 17:

- 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?

Response 17:

See response 1

Comment 18:

Table 1: Did you try looking for these in AirNow Tech by method_code? (https://aqs.epa.gov/aqsweb/documents/codetables/methods_all.html)

Response 18:

Yes, we pulled our information on the US sensors from the AQS database.

Comment 19:

- Figure 3: Did you consider whether just the 3 known types were significantly different?

Response 19:

The 3 known types were not significantly different; we tested this, and the p value went from 0.15 to 0.11 after removing the unknown category.

Comment 20:

- Table 2: It may be clearer to use letters to represent the sources so that they are not easily confused with the model numbers.

Response 20:

We modified the table to do this.

Comment 21:

- The conclusion would be easier to interpret if when referring to the models by number it also described the model type.

Response 21:

We added clarification to the conclusion briefly describing each model. I.e. "... Model 7, the multiple (RH) regression from Barkjohn et al."

Comment 22:

- 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?

Response 22:

Yes, one of our future recommendations is to develop a seasonally and regionally specific model as more data/sites become available.