Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC1
https://doi.org/10.5194/amt-2021-425-RC1, 2022

Comment on AMT-2021-425

Anonymous Referee #1

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-RC1, 2022

It's not clear what the objectives of this study were or what this manuscript adds to the literature on Plantower PMS5003 sensors and PurpleAir monitors.

Were the authors trying to develop simple models to correct ambient $PM_{2.5}$ concentrations reported by PurpleAir monitors across a range of locations and seasons in Canada and the United States? If so, how does this work build upon and differ from that of Barkjohn et al. [DOI: 10.5194/amt-14-4617-2021]?

It sounds like the authors might have been more interested in developing correction models that would help PurpleAir monitors predict "high" and "very high" AQHI+ levels correctly and expected wildfire smoke to be the most common cause of high or very high AQHI+ levels in Canada and the US. If so, the manuscript would benefit from discussions of (a) the existing literature describing the PurpleAir response to wildfire smoke [for example, Delp and Singer, DOI: 10.3390/s20133683 and Holder et al., DOI: 10.3390/s20174796] and (b) how the PurpleAir response to smoke is affected by the physics governing the PMS5003 sensor operation. Results from multiple field studies indicated that PMS5003 sensors overestimate smoke concentrations, and recent work by Ouimette et al. [DOI: 10.5194/amt-2021-170] indicated that this overestimation might be due to the small size of the particles produced during combustion. It is unlikely that the PMS5003 overestimates PM concentrations associated with wildfire smoke because of hygroscopic growth alone, so I'm a bit concerned by the authors' conclusion that Model 2, which was designed to account primarily for RH effects, is best for correcting PurpleAir data reported when $PM_{2.5}$ concentrations are "high" or "very high". I guess the k-value that was fit empirically helps account for some of the other factors.

Regardless of the authors' objectives, it would be great to see more discussion of the physics that govern PMS5003 sensor operation and parameters that can affect the accuracy of $PM_{2.5}$ concentrations reported by these sensors. Considering that the

relationship between the ambient $PM_{2.5}$ concentration and the light scattering signal received by the PMS5003 depends on particle size distribution, shape, refractive index, and density, and that these parameters can vary with location, season, and day-to-day weather, why do the authors conclude that a simple correction such as Model 2 or Model 7 is likely to produce acceptable results across a wide range of locations and seasons? Are there data on ambient aerosol properties in Canada and the US that the authors can cite to support this argument? *I appreciate simple correction approaches because they contribute to the goal of making $PM_{2.5}$ monitors and monitoring data accessible to a wide range of stakeholders*, but I also think readers will have reasons to be skeptical of these correction models, and the authors should acknowledge and discuss those concerns.

The authors argue that it's most important for PurpleAir monitors to predict "high" and "very high" AQHI+ values correctly, but I'm not sure I agree. When $PM_{2.5}$ concentrations are high or very high, it's usually clear to the public that air quality poses a health hazard. I think it might be more important for the PurpleAir to predict AQHI+ values between 2 and 6 correctly, so that they can alert the public to pollution levels that are elevated above the baseline AQHI+ of 1 but that individuals might not readily perceive as hazardous using their senses. This comment is just for the authors' consideration.

The comments that follow refer to specific lines, figures, and tables in the manuscript:

- Lines 72-73: "$PM_{2.5}$ concentration is reported by the sensors using two different proprietary correction factors ($PM_{2.5}$ CF 1 and $PM_{2.5}$ CF ATM) which convert the estimated particle count in size bins into the reported concentrations." This statement is speculative at best and most likely incorrect. I suggest the authors delete it. As the authors describe on lines 58-62, the PMS5003 functions as a nephelometer and not an optical particle counter. Ouimette et al. [DOI: 10.5194/amt-2021-170] have shown that the sensor output labeled "number of particles with diameter beyond 0.3 μm in 0.1 L of air" is a measure of the amount of light scattered to the detector by particles passing through the sensor. The relationship between the amount of light scattered to the detector and the $PM_{2.5}$ concentration reported by the sensor is unknown, but multiple published datasets indicate that the particle count distribution output by the PMS5003 is not accurate and largely invariant [see He et al., DOI: 10.1080/02786826.2019.1696015; Tryner et al., DOI: 10.1016/j.jaerosci.2020.105654; and Ouimette et al., DOI: 10.5194/amt-2021-170]. Tryner et al. show that PM_1, $PM_{2.5}$, and PM_{10} mass fractions calculated from the particle count data reported by the PMS5003 are not consistent with the PM_1, $PM_{2.5}$, and PM_{10} mass concentrations reported directly by the sensor [DOI: 10.1016/j.jaerosci.2020.105654, Figure 6]. Wallace et al. [DOI: 10.1016/j.atmosenv.2021.118432] also note this discrepancy in their Appendix ("Apparently, the hidden algorithms of the Plantower approach assign values to measurements that in some way depart from using the mass calculated from the numbers of particles in the size categories.").
- Lines 74-75: "The CF ATM correction factor is derived from Beijing atmospheric conditions while CF 1 was derived from a lab study using symmetrical particles of a known size and is recommended for use in industrial settings (Zhou, 2016)." Where did the authors obtain this information? The PMS5003 manual that the authors cite does not contain any such statements.
- Lines 76-77: "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).” I’m not sure why the authors chose to give this correction approach special attention in this paragraph. The approach of Wallace et al. wasn’t tested by the authors of this study and approaches to which the authors did compare their data aren’t discussed until Lines 91-97. I’m not convinced that the approach of Wallace et al. is promising, either, since it seems to be based on flawed assumptions: that the PMS5003 measures particle counts and that the $PM_{2.5}$ concentrations reported by the PMS5003 are calculated from the particle count data (see Comment 1).

- Lines 104-107: What was the domain for this analysis? Canada and the United States?
- Lines 132-133: Why not download hourly $PM_{2.5}$ concentration observations in the US from EPA AQS? There is a several-month lag between when the data are collected and when the data are available in AQS, but data in AQS have been QC’ed.
- Lines 137-138: “We further removed several sites after viewing scatter plots of their valid PA and FEM $PM_{2.5}$ observations and determining the performance to be unsatisfactory relative to the other sites.” What criterion/criteria was/were used to determine that performance was unsatisfactory? Pearson correlation < 50%? The answer is not clear based on the current phrasing of this sentence and the one that follows.
- Lines 141-143: “PA RH values were restricted to the range 30%–70% (any values above/below this were set to 30% or 70%, respectively) as these values are near the efflorescence and deliquescence points typical of fine particulate matter (Parsons et al. 2004, Davis et al. 2015). Corrections utilizing RH tended to overcorrect observations at these extreme RH values.” (a) Please specify the fraction of 1-hour average data points that were affected by this restriction. (b) What was the justification for modifying RH values below 30%? I don’t recall seeing evidence in the literature that low RH values are a problem. Did the authors do a sensitivity analysis to see whether and how their results were affected by leaving in RH values below 30%? (c) It’s unclear whether the authors decided to replace RH values above 70% based on prior results reported in the literature or based on the results of their own modeling. If this decision was based on prior results from the literature, the authors should cite relevant studies. If this decision was based on the authors’ own modeling, it would be nice to see these results presented in the form of a sensitivity analysis.
- Line 145: “Solar radiation impacts were too difficult to estimate given the variations in siting at each of the locations.” I don’t think the lack of correction for solar radiation is a big concern. The temperature and RH reported by the PurpleAir are biased high and low, respectively, even when the PurpleAir is installed indoors, due to heat generated by the electronics, so adjusting for solar radiation would still not eliminate the bias.
- Lines 190-192: How did the authors assign sites to the training and testing datasets? Was this assignment done randomly? Or did the authors try to make sure the full range of geographic areas, climates, and seasons were represented in each dataset?
- Line 206: “increases” should be “increased”.
- Line 208: “The mean testing site concentrations had similar ranges for matching monitor types between the two data sets…” I don’t understand what the authors mean by this. First the authors refer to “testing site concentrations”, but then they refer to “the two data sets”. Are the two data sets testing and training? FEM and PurpleAir? What were the “matching monitor types”? Does this phrase refer to the type of FEM monitor (e.g., beta attenuation) or to FEM vs. PurpleAir?
- Figure 2: “Distributions of the Federal Equivalent Method (FEM) and PurpleAir training/testing sites median $PM_{2.5}$ concentrations (μg m^{-3}) at dry (0%–33%), moderate (34%–66%) and humid (67%–100%) relative humidity (RH) groupings.” (a) How was each site categorised as dry, moderate, or humid? Based on the mean or median hourly RH at the site over the full data collection period? (b) Have the authors considered making the groups 0-30%, 30-70%, and 70-100% since they chose to modify RH values in the 0-30% and 70-100% ranges?
- Figure 3: Why were hours where the FEM AQHI+ was equal to 1 removed? For what fraction of hours was the FEM AQHI+ equal to 1? Most of them, right, considering the

medians in Figure 2?

- Table 2: Model 1 is similar to the model fit by Barkjohn et al., but with slightly different coefficients. Did the authors fit Model 1 to PurpleAir data that had been adjusted to replace RH values < 30% with 30% and RH values > 70% with 70%? If so, did the authors also try fitting Model 1 to a dataset without adjusted RH values? It would be interesting to see how much the coefficients fit using the dataset from this study differ from the coefficients fit by Barkjohn et al.
- Table 2: Did the authors test Model 7, which was fit by Barkjohn et al., using the dataset in which RH values < 30% had been replaced with 30% and RH values > 70% had been replaced with 70%? If so, I'm not sure that's a fair test of this model because it wasn't fit using such adjusted RH values.
- Figure 4: The color scheme used here was confusing to me. My initial reaction was that overestimates should be red and underestimates should be blue. Did the authors choose red for underestimates because they view the PurpleAir incorrectly underestimating the AQHI+ value, and therefore failing to alert the public to the true extent of the health hazard posed by air pollution, to be the worse outcome?
- Lines 256-264: Did the authors consider interpreting their results using the performance targets proposed in the US EPA Air Sensor Guidebook [document ID EPA/600/R-14/159]? In Section 5, the guidebook suggests precision and bias error < 50% for educational and informational purposes, < 30% for hotspot identification and characterisation or personal exposure monitoring, and < 20% for supplemental monitoring.
- Lines 299-300: "The Canadian AQHI+ system was useful as a framework for evaluating correction models across a range of concentrations infrequent high values or numerous low values can skew performance statistics when evaluating the full range at once." Is there a word missing here? Was this supposed to be two sentences?
- Lines 301-304: The names Model 2, Model 7, etc. are not very informative to a reader who is not looking at Table 2. It would be helpful to describe the key features of Models 2 and 7 here. Please also explain that models using RH as a predictor were fit and evaluated after replacing RH values below 30% with 30% and RH values above 70% with 70%. I don't think it's necessary to note Models 3 through 6 in this paragraph.
- Lines 305-306: "…the average performance across the testing sites and over time was evaluated here; performance at colocation sites and across time was not the same." Why did the authors choose to focus on the average performance across testing sites and over time? Were there any sites or times where Models 2 and 7 performed notably better or worse? If so, what were the notable features of these sites and times (weather, unique PM sources) and what do those features say about the advantages and limitations of Models 2 and 7?
- Lines 313-315: "…the improved particle bin correction factor proposed by Wallace et al. (2021) should be implemented for these sites…" See Comment 3. I don't understand why the authors repeatedly refer the correction approach proposed by Wallace et al. I don't think it's a good approach or particularly relevant to the work presented here.