Corrigendum to Atmos. Meas. Tech., 15, 3315–3328, 2022
https://doi.org/10.5194/amt-15-3315-2022-corrigendum

Atmospheric Measurement Techniques
Open Access
EGU

Corrigendum to "Development and evaluation of correction models for a low-cost fine particulate matter monitor" published in Atmos. Meas. Tech., 15, 3315–3328, 2022

Brayden Nilson[1,2], **Peter L. Jackson**[1], **Corinne L. Schiller**[1,2], **and Matthew T. Parsons**[2]

[1]Department of Geography, Earth and Environmental Sciences, University of Northern British Columbia, Prince George, V2N 4Z9, Canada
[2]Air Quality Science – West, Meteorological Service of Canada, Environment and Climate Change Canada, Vancouver, V6C 3S5, Canada

Correspondence: Brayden Nilson (brayden.nilson@ec.gc.ca)

Published: 27 July 2023

The labels for two different $PM_{2.5}$ values provided by the Plantower PMS5003 sensor using different internal calibrations ("CF 1" and "CF ATM") were incorrectly swapped in the dataset used for the publication. As a result, we built and tested our models using CF ATM data instead of the intended CF 1 data. This mistake also affected the evaluation of two out of the four models from the literature. However, Models 5 and 7 were not affected, as we erroneously used the incorrect column, which ended up being correct given the mislabelling. Our overall conclusions were not affected. To correct this issue, we propose changes to the publication that primarily involve revising the presentation and discussion of the models from the literature. We also correct references to the CF 1 data to say CF ATM. An additional figure (Fig. 8) is provided here showing the CF ATM and CF 1 data from our dataset after correcting the issue. The following text shows the affected sections, with altered text underlined, along with any updated figures and tables.

We performed additional analysis to further validate our results by attempting to fit our Model 2 with the (correctly labelled) CF 1 data and found degraded performance. We fit a k value of 0.74 for the CF 1 data and then calculated the mean AQHI+ bias across the range of Federal Equivalent Method (FEM) AQHI+ (Fig. 9). We observed an average AQHI+ bias of −0.06 and an average absolute AQHI+ bias of 0.17 for the Model 2 CF ATM, compared with an average AQHI+ bias of 0.1 and an average AQHI+ absolute bias of 0.3 for Model 2 CF1. We also re-calculated the performance statistics from Fig. 6 using the CF 1 Model 2 and observed degraded performance in comparison with the raw, uncorrected CF ATM data for much of the observed range. The CF 1 data in our dataset appear to not be as impacted by humidity as the CF ATM data, reducing the fit of our Model 2 which assumes a larger humidity impact. We also applied Models 5–8 to the opposite CF column to ensure the optimal data were being used for each model, and noticed significantly degraded performance when using the columns not presented in this corrigendum (Fig. 9).

Abstract. Four correction models with differing forms were developed on a training data set of 32 PurpleAir-Federal Equivalent Method (FEM) hourly fine particulate matter ($PM_{2.5}$) observation colocation sites across North America (NA). These were evaluated in comparison with four existing models from external sources using the data from 15 additional NA colocation sites. Colocation sites were determined automatically based on proximity and a novel quality control process. The Canadian AQHI+ system was used to make comparisons across the range of concentrations common to NA, as well as to provide operational and health-related context to the evaluations. The model found to perform the best was our Model 2, $PM_{2.5-corrected} = PM_{2.5-atm}/(1+0.24/(100/RH_{\%} - 1))$ – relative humidity (RH) is limited to the range [30 %, 70 %], which is based on the RH growth model developed by Crilley et al. (2018).

Corrected concentrations from this model in the moderate to high range, the range most impactful to human health, outperformed all other models in most comparisons. Model 7 (Barkjohn et al., 2020) was a close runner up and excelled in the low-concentration range (most common to NA). The correction models do not perform the same at different locations, so we recommend testing several models at nearby colocation sites and utilizing that which performs best if possible. If no nearby colocation site is available, then we recommend using our Model 2. This study provides a robust framework for the evaluation of low-cost $PM_{2.5}$ sensor correction models and presents an optimized correction model for North American PurpleAir (PA) sensors.

1 Introduction

PurpleAir (PA) monitors are targeted at citizen scientists and air quality professionals alike as small, low-cost, and easy-to-install devices and as such have proliferated to form a global network of monitors with thousands of devices in North America (NA) alone. These monitors contain two Plantower PMS5003 nephelometer sensors, which each detect $PM_{2.5}$ (named "A" and "B"), as well as separate sensors for measuring relative humidity (RH) and temperature. $PM_{2.5}$ concentration is reported by the monitor using two different proprietary correction factors ("$PM_{2.5}$ CF 1" and "$PM_{2.5}$ CF ATM") that convert the measured particle scattering amplitudes into the reported concentrations. The "CF ATM" correction factor is derived from Beijing atmospheric conditions, whereas "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; Johnson Yang, personal communication, 2019). The CF 1 data were found to correlate marginally better (−3 % to 6 %) with Federal Equivalent Method (FEM) observations in our dataset, however, the root-mean-square error (RMSE) and mean bias (MB) were significantly worse at most sites (RMSE: 1 % to 202 %; MB: −85 % to 149 %). $PM_{2.5}$ concentrations from the PA monitors have shown promising results when a colocation-study-derived correction model is applied but tend to overestimate FEM readings otherwise (Kim et al., 2019; Malings et al., 2019; Li et al., 2020; Feenstra et al., 2020; Tryner et al., 2020).

2.1 Colocation site selection and data retrieval

An automated algorithm to detect potential PA and FEM monitor colocation sites and apply quality control (QC) methods was developed to identify the sites that were co-located, defined as one or more outdoor PA monitors being within 50 m of each other (as of November 2020), and remove any periods of invalid data. Monitor coordinates were provided through the AirNow (for FEM monitors) and PA databases. Based on this, 86 sites were identified; however, further analysis was necessary to remove sites or periods of time where the PA monitors were likely not co-located or were located indoors. FEM monitor detector types were retrieved from the US AQS database (EPA, 2020) for the US stations; Canadian station information was collected through contact with the monitor operators.

FEM observation data were obtained from the AirNow database (AirNow, 2021), which provided hourly $PM_{2.5}$ concentration observations from sites across North America. PA data were retrieved from their ThingSpeak repository as hourly averages for comparison with the FEM monitors (PurpleAir 2021). Sensor A and sensor B "CF = ATM" data from each PA monitor were averaged together to produce a single record. Historically, the "CF=1" and "CF=ATM" were erroneously mislabelled in the PA data; this has since been resolved and it was ensured that the actual "CF = ATM" data were being used here.

3.1 Correction development

Eight correction models were selected for evaluation (Table 2) including four developed in this study from regressing the training data (Models 1–4), as well as four others available from the literature (Models 5–8). Models 1–5 use the "PM2.5 CF ATM" column, while Models 6–8 use the "PM2.5 CF 1" column. Model 1 is a multiple linear model including an RH term. Model 2 uses an RH growth adjustment ($k = 0.24$) to reduce the $PM_{2.5}$ concentration as RH increases (see Eq. 5). Model 3 is a second-degree polynomial with an RH term included. Model 4 is a three-breakpoint piecewise model with breakpoints selected to visually fit the data best over multiple iterations. The last four equations are provided from studies by other parties and consist of two simple linear models (Models 5 and 6), a multiple linear model including RH (Model 7) and a multiple linear model including both RH and temperature (Model 8).

3.2 Correction evaluation

Raw PA observations were biased positively at FEM AQHI+ levels between 1 and 9, peaking at an AQHI+ of 4 to 7, as shown in Fig. 5. An AQHI+ bias at or near zero is preferred, especially at the higher FEM AQHI+ levels most impactful to human health. Models 2, 5, and 7 minimized this bias the most on average. Model 2 was biased slightly high at AQHI+ of 6 or lower and slightly low onwards with a minimum at 10 AQHI+. Models 5 and 7 performed similarly; however, they were biased slightly low throughout except at an AQHI+ of 1 for Model 7 and AQHI + values less than 6 for Model 5. These two models had the worst performance at 8 and 10 AQHI+. Model 8 was the next best; however, it was positively biased in the moderate to high FEM AQHI+ levels. Models 1, 3, 4, and 6 had relatively large negative biases across most AQHI + levels. Further comparisons

Table 2. PurpleAir correction models selected for evaluation. The "Min" column indicates the minimum corrected value at a RH of 70 % (* and a temperature of 20 °C for Model 8).

Correction	Form	Source	Min	Formula
No model	–	–	0	pm 25_atm
Model 1	Linear (+ RH)	–	−2	$0.708 * \text{pm25\_atm} - 0.115 * \text{rh} + 5.78$
Model 2	RH growth	–	0	$\text{pm25\_atm}/(1 + 0.24/(100/\text{rh} - 1))$
Model 3	Polynomial (+RH)	–	0.3	$0.53 * \text{pm25\_atm} + 0.000952 * \text{pm25\_atm\_2} - 0.0914 * \text{rh} + 6.3$
Model 4	Piecewise at 2.5/40/300	–	1.9	pm25_atm ≤ 2.5: $0.92 * \text{pm25\_atm} + 1.86$ 2.5 < pm25_atm ≤ 40: $0.42 * \text{pm25\_atm} + 3.1$ 40 < pm25_atm ≤ 300: $0.87 * \text{pm25\_atm} - 14.8$ pm25_cf1 > 300: $1.16 * \text{pm25\_atm} - 100.6$
Model 5	Linear	A	2.6	$0.778 * \text{pm25\_atm} + 2.65$
Model 6	Linear	B	−0.7	$0.50 * \text{pm25\_cf1} - 0.66$
Model 7	Linear (+RH)	C	−0.2	$0.534 * \text{pm25\_cf1} - 0.0844 * \text{rh} + 5.71$
Model 8	Linear (+RH +T)	D	1.5*	$(\text{pm25\_cf1} + 3.04 + 0.07 * \text{temp} - 0.02 * \text{rh})/1.55$

A. Kelly et al. (2019) B. LRAPA (2019) C. Barkjohn et al. (2020) D. Ardon-Dryer et al. (2019).

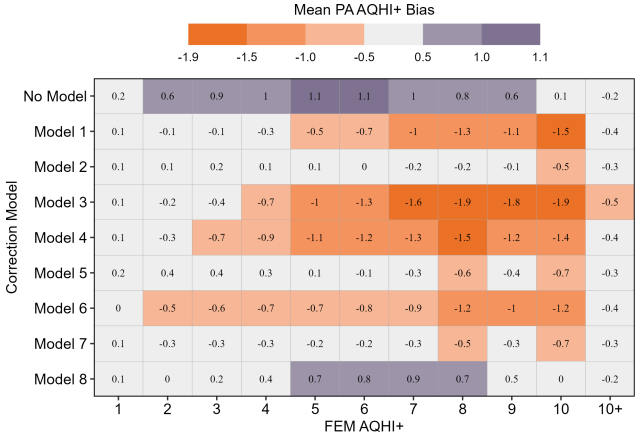

Figure 5. Mean PurpleAir AQHI+ bias for each correction model (including the raw data) at Federal Equivalent Method (FEM) AQHI+ levels. A value at or near 0 is preferred, especially for higher AQHI+.

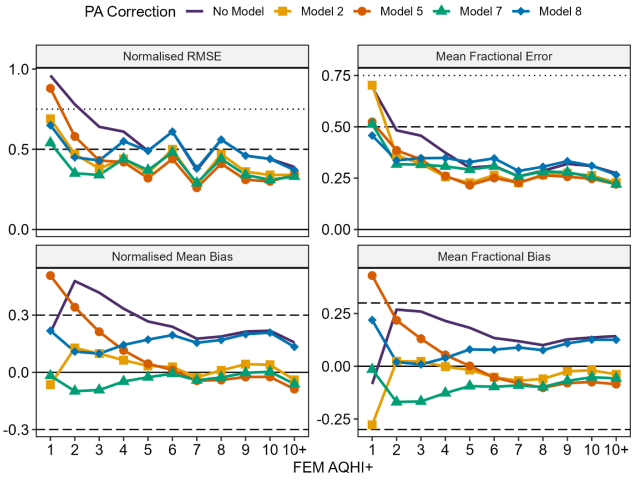

Figure 6. Comparison statistics across Federal Equivalent Method (FEM) AQHI+ levels for select correction models. Goal and acceptable levels are displayed where possible for RMSE (0.75 & 0.5), MFE (0.75 & 0.5), NMB (±0.3 & ± 0.6), and MFB (±0.3 & ± 0.6).

were only made on Models 2, 5, 7, and 8 as they showed the best performance here.

The normalized mean bias (NMB), normalized root-mean-square error (NRMSE), mean fractional error (MFE) and mean fractional bias (MFB) for the PA monitors were worse at lower concentrations and improved as concentrations increased (Fig. 6). We saw similar performance between models. All models improved upon the raw PA observations at most concentrations. Model 8 worsened the mean fractional bias measurement at an AQHI+ of 1 and had only marginal improvements over the raw data at ∼ 5 AQHI+ and above. Model 5 performed poorly in the low range but performed well in the moderate and higher ranges. Models 2 and 7 had the best performance across the concentration range. Model 7 was best for the very low observations (AQHI+ of 1); however, Model 2 tended to perform better starting at AQHI+ of 2–3.

Goal and acceptable levels for MFE and MFB are suggested in Boylan and Russell (2006). Raw PA data meet the goal level of 50 % for all but the lowest AQHI+ for MFE, where it still meets the acceptable level of 75 %. Only Models 7 and 8 bring these lowest concentrations into or near the goal level. Both uncorrected and corrected observations were within the goal range for MFB (±30 %), except for the AQHI+ level of 1 for Model 5. We assumed goal and acceptable levels for NRMSE of 50 % and 70 %, respectively, to align with the levels defined for MFE. Using these standards, the uncorrected PurpleAir data

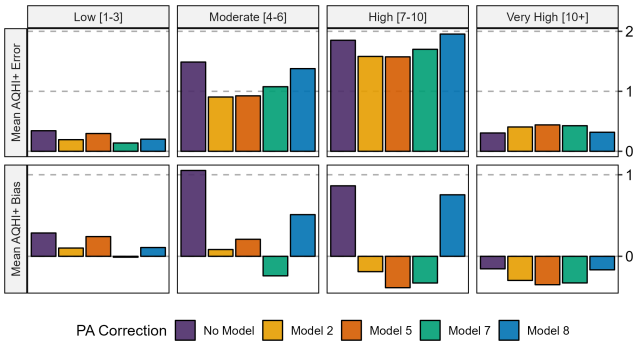

Figure 7. Mean PurpleAir (PA) AQHI+ error and bias for low, moderate, high, and very high Federal Equivalent Method (FEM) AQHI+ levels for selected correction models.

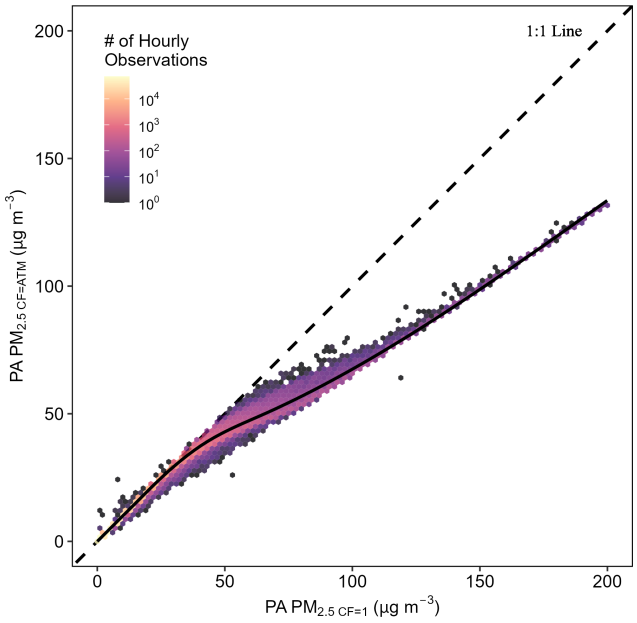

Figure 8. Comparison of PurpleAir (PA) $PM_{2.5}$ CF = ATM and CF = 1 following the correction of the mislabelling issue. CF = 1 should be larger than CF = ATM at all concentrations above $\sim 50\,\mu g\,m^{-3}$. CF = 1 observations greater than $200\,\mu g\,m^{-3}$ were removed here.

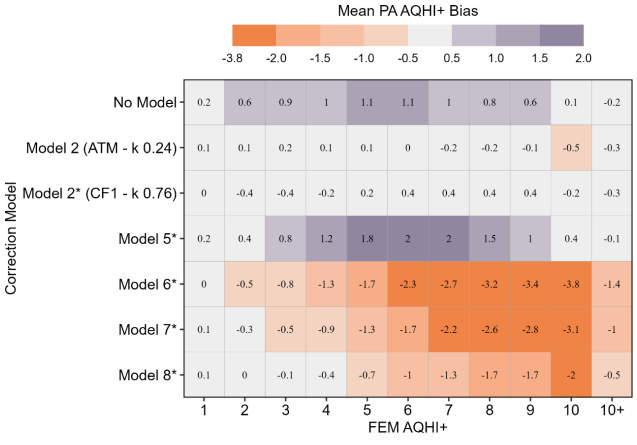

Figure 9. Mean PA AQHI+ bias for Models 5–8 when applied to the alternate CF column and for Model 2 developed for both the ATM and CF1 columns. Here, Model 5 uses CF 1, and Models 6–8 use CF ATM. Note the differences with Fig. 5 where the correct CF columns are used.

are unacceptable at AQHI+ equal to 1 and get increasingly better until reaching the goal level at high concentrations. Each correction model (except Model 5) brings the data into the goal level, except at AQHI+ equal to 1, where it is only acceptable. A goal level of ±30 % was assumed for the NMB like that for the MFB and the level defined for mean bias (MB) defined in Chang and Hanna (2004). Only the uncorrected data for AQHI+ values between 2–4 and the Model 5 data for an AQHI+ of 1–2 exceeded this goal level across our sites.

These statistical comparisons were also made on coarser AQHI+ groupings (Table 3). A ranking score was calculated for each model using the mean average of the ranks (from 1 to 5) for each statistic within an AQHI+ group. A lower score indicates better relative performance within that AQHI+ range. Model 2 performed the best in the high and very high AQHI+ categories, followed closely by Models 5 and 7. The models did not perform as consistently well in the low and moderate ranges. Models 7 and 8 tended to be the best for the low range, while Models 2 and 5 were better for the moderate range.

The error and bias in AQHI+ from the uncorrected PA monitors were greatest in the moderate to high range (Fig. 7). The selected corrections produced similar improvements to each other, all improving upon the PA's ability to correctly determine the hourly AQHI+ except for at the most extreme concentrations ($> 100\,\mu g\,m^{-3}$). This was not the case for the mean AQHI+ bias; however, as Model 1 becomes increasingly negatively biased as FEM concentrations increase. Model 2 was the best overall performer for both AQHI+ error and bias, being marginally outperformed only in the low-concentration range. This was followed by Model 7, which performed well for all but the mean AQHI+ error in the moderate-concentration range. Model 5 performed similarly to Model 7, except in the low range where Model 7 had a lower mean bias and error. Model 8 performed sufficiently; however, it did not perform as well in the moderate and high AQHI + categories.

4 Conclusions

The selected corrections discussed here improve the performance of PA similarly; however, Model 2 (our "RH Growth" model) had consistently better performance, especially at

Table 3. PurpleAir (PA) normalized mean bias (NMB), normalized root mean square error (NRMSE), mean fractional error (MFE) and mean fractional bias (MFB) at low, moderate, high, and very high Federal Equivalent Method (FEM) AQHI+ levels for each PA correction model. A crude score is calculated by averaging each statistic's integer rank (from 1 to 5) for the models within that AQHI+ group. The top-performing models are highlighted.

FEM AQHI+	Model	NMB	NRMSE	MFE	MFB	Score
Low [1–3]	No model	0.32	0.96	0.65	**−0.01**	3.8
	Model 2	**0.02**	0.62	0.63	−0.22	3
	Model 5	0.42	0.77	0.49	0.38	4.3
	Model 7	−0.05	**0.49**	0.47	−0.05	**1.8**
	Model 8	0.17	0.61	**0.43**	0.18	2.3
Moderate [4–6]	No Model	0.30	0.59	0.35	0.20	5
	Model 2	0.05	0.44	**0.25**	**−0.01**	1.9
	Model 5	0.08	**0.40**	**0.25**	0.03	**1.8**
	Model 7	**−0.03**	0.44	0.30	−0.11	2.6
	Model 8	0.16	0.56	0.34	0.05	3.8
High [7–9]	No model	0.19	0.47	0.28	0.12	4.6
	Model 2	**0.00**	0.38	0.25	**−0.06**	**1.8**
	Model 5	−0.04	**0.33**	**0.24**	−0.09	**1.8**
	Model 7	−0.03	0.36	0.27	−0.09	2.5
	Model 8	0.17	0.47	0.30	0.09	4.4
Very high [10+]	No Model	0.16	0.39	0.28	0.14	5
	Model 2	**−0.04**	**0.34**	0.23	**−0.04**	**1.8**
	Model 5	−0.08	**0.34**	**0.22**	−0.08	2.3
	Model 7	−0.06	**0.34**	**0.22**	−0.06	2
	Model 8	0.14	0.38	0.27	0.13	4

moderate to high concentrations that are important to health. This was followed closely by Model 7, the multiple (RH) linear regression from Barkjohn et al. (2021), and Model 5, the simple linear regression from Kelly et al. (2017). Model 8, the multiple (RH & temperature) linear regression from Ardon-Dryer et al. (2019) performs well, especially in the low to moderate range; however, it did not perform as well at higher concentrations for the normalized statistics we presented when compared with the raw data. Models 1 and 3 through 6 tended to overcorrect the PA data in the moderate to high range. 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. In addition, while our correction model focuses on correcting the impacts of humidity, other characteristics like refractive index and particle shape, density, and size distribution may account for differences in $PM_{2.5}$ estimates.

We recommend testing the performance of several models at specific sites of interest and selecting the best-performing model for a given site (Fig. S4 provides a breakdown for the testing sites and models evaluated here). Models 2, 5, 7, and 8 presented here are good starting points. As more colocation data become available, seasonal and area-specific correction models should be examined. Performance in the moderate- to high-concentration range should be focused on as these are the most important from a health perspective; the low concentrations are less important, while also being the most observed levels in the US and Canada. Correlation is useful for evaluating overall monitor performance at a site but not as useful for evaluating and comparing correction performance. Normalized methods such as NRMSE, MFB, or MFE are good measures, but we recommend evaluations across a range of $PM_{2.5}$ concentrations, such as using the AQHI+ framework as presented here. If one intends to develop a site-specific correction model, we recommend using the same form as our Model 2 while adjusting the k value. 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.