ARTICLE

Utilizing Reanalysis Datasets to Improve the Performance of Low-Cost Air Sensors in the Global South

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Michael R. Giordano*[,a,b], Matthias Beekmann[c,d], Emmanuel Appoh[e], Allison F. Hughes[e,f], Michael J. Gatari[g], James Nimo[e,f], Moses N. Njeru[g], Stuart Piketh[h], Albert Presto[a], Nyaga Waiguru[c,d], Daniel M. Westervelt[i,j], R. Subramanian[k,*]

A Low-cost sensors for particulate matter can provide high spatiotemporal resolution monitoring of air quality, especially in much of the Global South, and sub-Saharan Africa (SSA) in particular, where reference-grade instrumentation is often not available. However, ensuring high-quality data from low-cost sensor (LCS) platforms is essential. Until now, LCS required calibration by collocation with a reference-grade monitor to be used for more than qualitative studies of air quality, but reference-grade monitors are not available in many countries of the Global South. Since a key artifact in optical PM sensors is aerosol hygroscopic growth, we explore the viability of an alternative LCS calibration method: a hygroscopic growth correction factor using particle composition data from the MERRA-2 reanalysis dataset. We compare 3 different LCS located in 3 different areas of SSA – Kenya, Ghana, and South Africa - with 3 different calibration techniques: traditional linear calibrations with a reference-grade monitor, a κ-Köhler-derived correction with MERRA-2 data, and a random forest machine learning regression utilizing MERRA-2 and the regulatory-grade monitor. Random forest regressions using MERRA-2 particle composition data and collocation with a reference-grade monitor improve sensor performance to near that of regulatory-grade monitors. But even without collocation, a hygroscopic growth correction based on MERRA-2 particle composition alone can improve LCS $PM_{2.5}$ performance by reducing mean-normalized bias to near-zero and reducing error by up to 40%.

Introduction

Poor air quality is one of the leading causes of premature mortality across the globe. In sub-Saharan Africa (SSA), the problem of poor air quality is especially acute and is not only linked to high morbidity and mortality but also to billions of dollars of lost economic output.[1,2] Addressing the issue of poor air quality, however, first requires that pollution concentrations are measured. In much of SSA, regulatory-grade, high-fidelity monitoring of air pollution is not widely available (if at all), generally due to the price of such monitors (on the order of USD 40,000 after duties and taxes), which creates a barrier for policy and decision-makers to take action against air quality issues.[3]

Low-cost sensors (LCS), with costs on the order of USD 10-1000, offer an attractive alternative to high fidelity monitoring (i.e. regulatory- or research-grade monitors) due to their low(er) initial and maintenance costs and ease of deployment. Unfortunately, out-of-the-box performance of LCS is generally low and unsuitable for anything more than qualitative assessments of air quality, especially when deploying LCS in areas very different from factory-calibration conditions, i.e. the Global South in general and SSA in particular.[4] Out-of-the-box performance can, however, be corrected to near-regulatory grade measurements and much work has been done on the subject.[5]

A common and effective method to correct LCS $PM_{2.5}$ performance is through collocations with regulatory- or research-grade instruments (hereafter "reference instruments") in proximately the same area where the sensors are to be deployed. Collocation-based corrections empirically account for the varied factors that affect LCS performance, such as usual mass loadings, size distributions, particle shape, particle composition, source mixtures, and ambient temperature and humidity ranges.[e.g. 6–8] These methods, whether they use simpler linear regression methods or more complex machine learning tools, however, necessitate the presence of a reference instrument. In much of SSA, access to such instrumentation is simply unavailable or inaccessible, so alternative methods to ensure higher-accuracy data from LCS are needed. Fully theoretical approaches to LCS calibration have recently been developed utilizing Mie Theory but these

[a.] *Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA*
[b.] *AfriqAir, Kigali, Rwanda.*
[c.] *Univ Paris Est Creteil, CNRS UMS 3563, Ecole Nationale des Ponts et Chaussés, Université de Paris, OSU-EFLUVE – Observatoire Sciences de L'Univers-Envelopes Fluides de La Ville à L'Exobiologie, F-94010 Créteil, France*
[d.] *Université de Paris and Univ Paris Est Creteil, CNRS, LISA, F-75013 Paris, France*
[e.] *Afri-SET, Accra, Ghana*
[f.] *Department of Physics, University of Ghana, Accra, Ghana*
[g.] *Department of Electrical and Information Engiineering, University of Nairobi, Nairobi, Kenya*
[h.] *Unit for Environmental Sciences and Management, Northwest University, Potchefstroom, South Africa*
[i.] *Lamont-Doherty Earth Observatory of Columbia University, New York, NY, USA*
[j.] *NASA Goddard Institute for Space Studies, New York, NY, USA*
[k.] *Center for Study of Science, Technology and Policy (CSTEP), Bengaluru, Karnataka 560094, India*
Supplementary Information available: See DOI: 10.5281/zenodo.11537585

approaches can be computationally expensive, fairly opaque to all but the most advanced of users, and require a priori knowledge of the aerosols being measured by the LCS.[9,10] A different approach is to use semi-empirical relationships, such as κ-Köhler theory.[11] This approach still requires some a priori knowledge of aerosol size distribution and composition but is much less computationally expensive and much more transparent to lay-users. Previous work has already demonstrated that by using κ-Köhler theory LCS accuracy can be improved to within 33% of reference instruments and can result in lower mean absolute errors.[12–14] However, information about aerosol hygroscopicity, which depends on aerosol size distribution and chemical composition, is still needed. Aerosol chemical composition is generally measured with filters or aerosol mass spectrometers, but collecting such data can be logistically complicated and difficult, especially on an ongoing basis and at high time resolution (~hourly), which usually requires the use of aerosol mass spectrometers.

To overcome the need for reference monitors and measured aerosol composition, we build upon, first, recent work showing that most cheap particle sensors (e.g Plantower, Sensirion, etc.) are relatively insensitive to aerosol size distribution[15] and, second, the finding by Malings et al. (2020) that the hygroscopic correction factor (as applicable to LCS) is not particularly sensitive to the exact aerosol composition; an approximation of the regional composition may be sufficient (this latter finding may well be a result of the former). A relatively new, freely available source of spatially-resolved global aerosol composition is NASA's Modern-Era Retrospective analysis for Research and Applications-2 (MERRA-2) reanalysis data set, obtained from satellite observations combined global chemistry transport modelling. In this paper, we evaluate how MERRA-2 $PM_{2.5}$ composition outputs can be used with κ-Köhler theory to generate correction factors for LCS deployed in SSA. We performed this analysis at sites in East Africa, West Africa, and South Africa, where reference-grade instruments were available for verification. While typical correction models are built on linear regressions, the complexity of aerosol hygroscopic behavior may not be captured well by linear parametric relationships. Hence, we also examine how machine learning algorithms can be applied to the combined MERRA-2 and LCS data to improve LCS performance in data-sparse regions.

Experiment

Sites, Instrumentation, and Data Availability

This paper covers data from three deployments of LCS collocated with reference instruments from AfriqAir (www.afriqair.org) and AfriqAir partners across sub-Saharan Africa. The first set of instruments includes 2 Real-time, Affordable, Multi-Pollutant (RAMP) monitors deployed with attached PurpleAir PA-II (Plantower PMS5003) units and a Met One Beta Attenuation Monitor (BAM-1020). These instruments were deployed at the Theha Setjhaba Primary School (26°51'09.5"S 27°51'23.0"E) in Zamdela, South Africa. Zamdela is a semi-urban informal settlement approximately 80km SSW of Johannesburg with approximately 90000 residents. Major $PM_{2.5}$ sources include vehicles and combustion from residential cooking, heating, and waste disposal. The RAMPs were deployed from November 2019 to April 2021. In this study, we focus on data from November 2019 to April 2020 due to instrument maintenance issues caused by COVID-19 lockdowns. We use an average of the (cf = atm) A and B $PM_{2.5}$ channels from the Plantower sensors, provided they are within 10% of each other (data are not included if this check is not passed).

The second set of instruments covered here are 5 Clarity Node-S monitors (with internal Plantower PMS 6003 sensors for PM measurements) deployed with another Met One BAM 1020 at the University of Nairobi (UoN; 1°16'44.2"S 36°49'02.9"E) from February 2021 to June 2021. The UoN campus is located in central Nairobi and previous work has shown that mineral dust and traffic are the dominant $PM_{2.5}$ sources though other combustion and industrial and "mixed" sources are also important.(Gaita et al., 2014) Data from the Clarity nodes were downloaded from the Clarity API. We use the "Raw PM2.5 Mass Concentration" (pm2_5ConcMass.raw) data. $PM_{2.5}$ concentrations of less than 1 μg/m^3 were removed from the dataset but no other filtering of the data was necessary.

The final set of instruments used in this study was a Quant-AQ Modulair-PM unit deployed with a US Dept. of State Met-One BAM-1020 at the US Embassy in Accra, Ghana (5°34'46.2"N 0°10'14.6"W). We use data from July 2021 to January 2022. The Modulair-PM uses two light scattering-based particle sensors (Plantower PMS5003 and Alphasense OPC-N3) to measure PM_1, $PM_{2.5}$, and PM_{10} concentrations and size distributions. Here we use the $PM_{2.5}$ output of the instrument on which Quant-AQ applies a calibration incorporating assumptions of chemical composition. The Modulair-PM calibration utilizes the dual optical sensor outputs along with size-dependent hygroscopicity (κ) and density (ρ) assumptions to calibrate the output data.(Hagan and Cross, 2022) In Accra, the κ and ρ values were taken from the instrument data portals and range from $0.02 < \kappa < 0.3$ and $1.65 < \rho < 2.5$.

MERRA-2

MERRA-2 is an atmospheric reanalysis of the modern satellite era produced by NASA's Global Modeling and Assimilation Office (GMAO), utilizing satellite observations combined with the Goddard Earth Observing System model.(Gelaro et al., 2017) We utilize the Modern Era Retrospective analysis for Research and Applications Aerosol Reanalysis (MERRAero) reanalysis which simulates, through the Goddard Chemistry, Aerosol, Radiation, and Transport aerosol module, the surface concentrations of five types of aerosols: dust ($PM_{2.5}$ segregated; $DUST_{2.5}$), sea salt ($PM_{2.5}$ segregated; $SS_{2.5}$), black carbon (BC), organic carbon (OC) and sulfate (SO_4). These are treated as non-interacting external mixtures and are derived from surface wind-speeds (dust and sea salt) and standard emissions inventories such as EDGAR 4.2.(Buchard et al., 2016) MERRA-2 has a native spatial resolution of 0.5° lat x 0.625° lon (nominal ~50km latitudinal resolution) and 1-hour temporal resolution.

Correction Models

Three different methods of calibrating the LCS are presented here: multilinear regressions, composition dependent hygroscopicity corrections, and a machine learning random forest regression. These three methods are respectively the "traditional" (and simplest) method, a novel application of a semi-empirical method, and a novel application of machine learning algorithms for LCS calibration. The linear calibrations all follow the same form:

$$PM_{2.5,cal} = \alpha PM_{2.5,LCS} + \beta RH_{LCS} + \gamma \quad (1)$$

where $PM_{2.5,\,cal}$ is the calibrated $PM_{2.5}$ from the LCS (the $PM_{2.5}$ measurements from the collocated reference instrument when calculating the regression coefficients), RH_{LCS} is the relative humidity measurement from the LCS, and α, β, and γ are regression coefficients solved by a multilinear regression implemented in Python. As noted in Giordano et al.(Giordano et al., 2021), temperature measurements (from the LCS or external) are generally not necessary for many LCS collocations. This was confirmed for the collocations presented here by including a temperature term in Eq. 1 which resulted in standard errors on the temperature coefficients showing that the terms are not statistically significant in this case.

The semi-empirical hygroscopic correction applied here are described in detail in Malings et al.[13]. In short, the LCS measurements are corrected such that:

$$PM_{2.5,cal} = \frac{PM_{2.5,LCS}}{f(RH,T,\kappa)} \quad (2)$$

where the hygroscopic growth factor $f(RH,T,\kappa)$ is calculated as:

$$f(RH,T,\kappa) = 1 + \kappa_{bulk}\frac{a_w(T,RH)}{1 - a_w(T,RH)} \quad (3)$$

where water activity a_w is calculated as:

$$a_w(T,RH) = RH_{LCS}exp(\frac{4\sigma_w M_w}{\rho_w RT D_p})^{-1} \quad (4)$$

where σ_w, M_w, and ρ_w are the surface tension, molecular weight, and density of water, respectively; T is the absolute temperature from the LCS; R is the ideal gas constant; and Dp is the particle diameter. κ_{bulk} is calculated from the MERRA-2 dataset such that:

$$\kappa_{bulk} = \kappa_{SO4}\varepsilon_{SO4} + \kappa_{SS}\varepsilon_{SS} + \kappa_{OC}\varepsilon_{OC} \quad (5)$$

where κ_{SO4}, κ_{SS}, and κ_{OC} are the sulfate, sea salt $PM_{2.5}$, and organic carbon κ vales set to 0.5, 1.1, and 0.15, respectively; and ε_{SO4}, ε_{SS}, and ε_{OC} are the mass fractions of the sulfate, sea salt, and organic carbon in $PM_{2.5}$ from MERRA-2 where the total mass of $PM_{2.5}$ from MERRA-2 is calculated using the five aerosol types as described in Buchard et al.[16]

The Random Forest (RF) regression is a machine learning algorithm that predicts regression values from new input data by constructing an ensemble of decision trees using a training data set.[17] Here we construct and train the RF using the $PM_{2.5}$, RH, and T measurements from the LCS along with the 5 aerosol masses from MERRA-2 with the reference instrument $PM_{2.5}$ measurements as the known training values. A Python

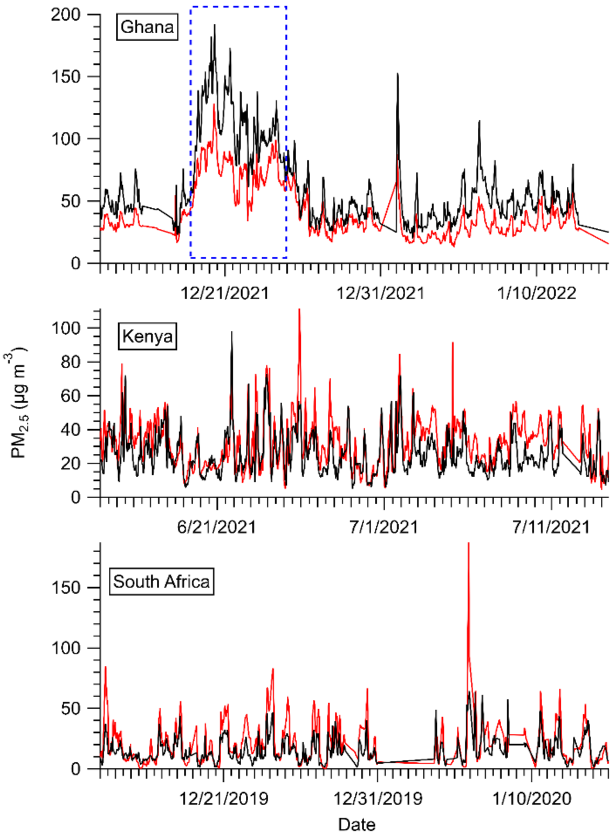

Figure 1 Representative PM2.5 concentrations from the reference instruments (black) and LCS (red) for the three measurement locations at 1-hour time resolution. The blue box indicates a Harmattan dust event captured in Ghana.

implementation of the RF regression is implemented with the Scikit-learn library.[18] The default settings for the RandomForestRegressor are used with the exceptions of increasing the number of trees to 1000 due to the density of the measurements used here and setting random_state = 42 for reproducibility. The number of trees can significantly impact the performance of the regressor but tests of 100, 250, 500, 750, 1000, 2000, and 5000 trees indicate that 1000 is a good compromise between performance and computation time for these datasets. A 5-fold cross validation is performed to help mitigate a risk of over fitting as well.

Performance Metrics

Here we focus on the Pearson correlation coefficient (r), mean-normalized bias (MNB), mean absolute error (MAE), and coefficient of variation in the mean absolute error (CvMAE). All are described in detail in Malings et al.[13]. The metrics presented in the main text well capture the accuracy and precision of the calibration methods though additional metrics are presented in the Supplementary Information.

Results and Discussion

$PM_{2.5}$ Concentrations

Figure 1 shows a representative subset of the $PM_{2.5}$ data measured by both the reference instruments and the LCS (uncorrected data) from each of the three measurement locations.

Over the entire collocation period in Accra, Ghana (195 days) $PM_{2.5}$ concentrations (from the BAM; hourly measurements) range from 5-192 µg/m³ with an average of 31.7 µg/m³. If the Harmattan dust event in Dec. 2021 (confirmed as a dust event from on-the-ground observers) is excluded from averaging it will be 24.3 µg/m³. In Nairobi, Kenya, the range is not as extreme as Accra but still large at 2-98 µg/m³ and a mean of 19.9 µg/m³ (from the BAM). Zamdela, South Africa is similar to Nairobi with a range of 1-79 µg/m³ and an average of 17.3 µg/m³. Figure 1 emphasizes an important fact about $PM_{2.5}$ measurements in Africa: the range and average values are generally much greater than those measured in the Global North, irrespective of the sampling heights from ground level. This, combined with the fact that the PM sources are also very different between the Global North and Africa, means that applying calibrations made outside the specific deployment location in Africa (since PM sources differ between E, W, and S Africa) will generally yield poor results (see SI Figure S1).

The overall performance of MERRA-2 in terms of surface-level $PM_{2.5}$ mass estimates is also important to note. Figure 2 shows the $PM_{2.5}$ concentrations measured from the 3 reference monitors compared to the overall $PM_{2.5}$ mass estimated by MERRA-2, calculated as in Buchard et al.[16] Included in the figure

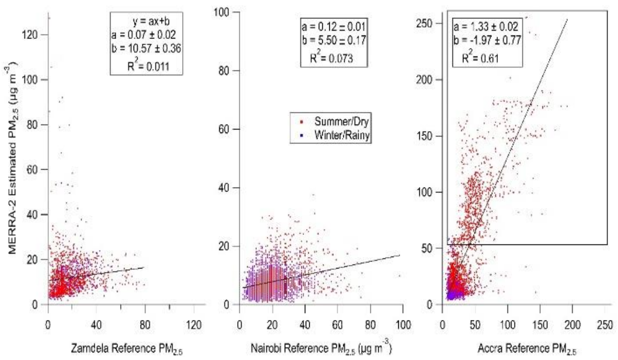

Figure 2 Comparison of MERRA-2 estimated PM2.5 with surface reference measurements in Zamdela, South Africa; Nairobi, Kenya; and Accra, Ghana. The

is an approximate note of the Harmattan period in Accra, Ghana (noted as a black box) and a simple linear best-fit regression for all 3 locations. Overall, the agreement between estimated and measured concentrations is quite poor for both Zamdela, South Africa and Nairobi, Kenya with R^2 correlation values <0.1. If the data is segmented into winter/summer and rainy/dry periods, respectively, neither location shows improved agreement between measured and estimated $PM_{2.5}$. In Accra, Ghana, the apparent agreement between the reference monitor and MERRA-2 is quite high, with a R^2 correlation of 0.61. However, if the Harmattan period is removed from the data, R^2 falls to 0.01 (see SI figure S2). As shown in the next section, MERRA-2 agrees with intuition and attributes most of the PM loading in the dry season to dust but, in general, over predicts overall PM loadings in Accra in the dry seasons by at least 50% and under predicts in the rainy seasons by the same (CvMAE = 0.64 and 0.5, respectively; see Table S1).

MERRA-2-derived $PM_{2.5}$ Composition

As previously discussed, LCS are sensitive to aerosol physiochemical characteristics, especially aerosol composition and size. As most countries in the Global South, and Africa especially, have little or no direct composition measurements, we must turn to other tools to obtain composition estimates. Figure 2 shows the average composition for the three measurement locations obtained from MERRA-2, segregated into dry and rainy seasons for Nairobi and Accra (where rainy seasons are March-May and October-November for Nairobi,

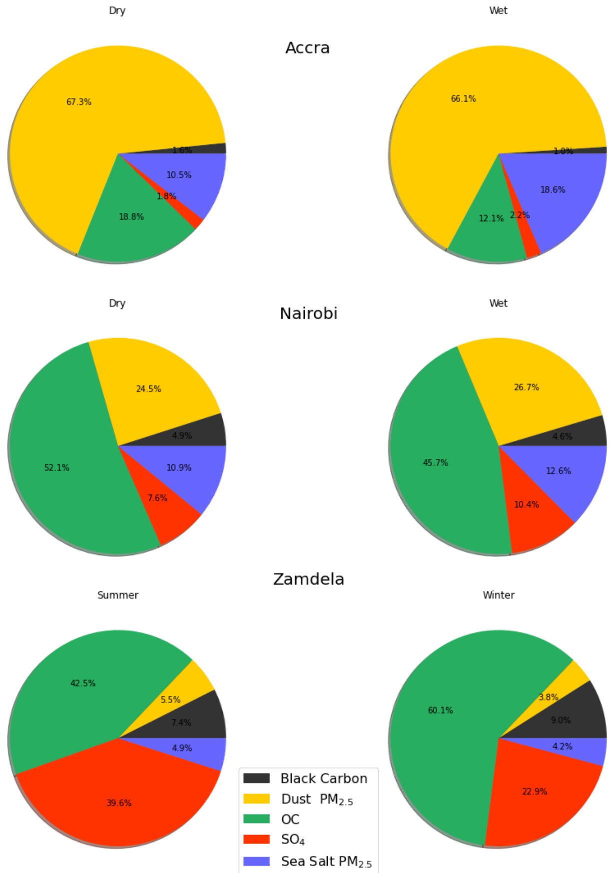

Figure 3 Average aerosol composition estimates from MERRA-2 for Accra (top), Nairobi (mid), and Zamdela (bottom) for the dry/wet (left) and summer/winter (right) seasons

and March-July Sep.-Nov. for Accra) and winter/summer in the case of South Africa.

For the most part, these MERRA-2 aerosol composition predictions make sense for both Nairobi and Zamdela. Given the proximity of the measurement sites to anthropogenic PM sources (industry, vehicles, biomass burning), OC making up the largest fraction in each site and season is reasonable and agrees with previous literature.[19,20] The high sulfate loadings in Zamdela generally agree with the work of Muyemeki et al (2021) which showed high year-round loadings with higher loadings present in the summer months. The MERRA-2 results also generally agree with Muyameki et al.(2021) results for dust

https://doi.org/10.26434/chemrxiv-2024-j5dcn ORCID: https://orcid.org/0000-0002-6820-6668

and sea salt. In Nairobi, dust making up about a quarter of the total $PM_{2.5}$ loading in both the rainy and dry seasons is in general agreement with the work of Gaita et al. [21]. Gaita et al. did show, however, that there are large seasonal differences even between the short and long rainy periods for mineral dust concentrations which do not appear to be captured in MERRA-2 (see SI Figure S3). For black carbon and sea salt, not enough work has been done in Nairobi to determine if either the overall loadings and/or their temporal distribution are correct, but the results from MERRA-2 seem reasonable.

The largest question mark is, however, in Accra where dust constitutes the most abundant species in both the rainy and dry seasons. Accra does suffer from high dust outflows from the Sahara due to the Harmattan winds caused by seasonal changes in the Inter-Tropical Convergence Zone, but this generally only occurs in December and January. MERRA-2 calculates extremely high dust loadings, showing a sinusoidal pattern with elongated peaks that top out at over 80% of the total $PM_{2.5}$, for most of the year except from May to September where the contribution drops to near 0 (see SI Figure S4). Even examining the overall $PM_{2.5}$ loadings, MERRA-2, does not seem to perform well over Accra. The reanalysis simply has high overall loadings from approximately September to March and completely misses the March-May rainy season and the short summer dry season (see SI Figure S5 or US Embassy BAM-1020 data from airnow.gov). Therefore, this dry and wet demarcation of the MERRA-2 data for Accra is not applicable. Due to the lack of recent speciation measurements, it is unclear if the BC, OC, SO_4, and sea salt estimated contributions are realistic but it can be said that the dust contributions are unrealistic with the exception of the Harmattan months.

It is important to remember that we use MERRA-2 because it seems to be the best estimates for aerosol composition that we have available. It is also important to remember that this usage of MERRA-2 may not be ideal due to both the emissions inventories that MERRA-2 uses and the fact that a nominal ~50 km horizontal spatial resolution may not well capture the urban emissions setting that the Accra monitor is in. Still, using these estimates could be a good starting point to develop calibration models for low-cost sensors. More work can and should be done comparing MERRA-2 data for more sites in West Africa specifically.

Low-cost sensor calibrations

In total, there are 4 data streams from the low-cost sensors that need to be examined: raw (uncalibrated/uncorrected) data, corrected data using collocation-based linear corrections, corrected data using the aerosol hygroscopicity *f(RH,T,κ)* linear correction, and corrected data using the aerosol composition-based random forest calibration. Note that the first and third corrections, the linear model and random forest, require the presence of a reference monitor while the *f(RH,T,κ)* correction as implemented here does not. Figures 4, 5, and 6 show the four LCS data streams for Accra, Nairobi, and Zamdela, respectively, plotted against the reference monitor $PM_{2.5}$ data for those locations. These figures should be viewed in conjunction with Table 1 to provide a quantifiable context to the visuals.

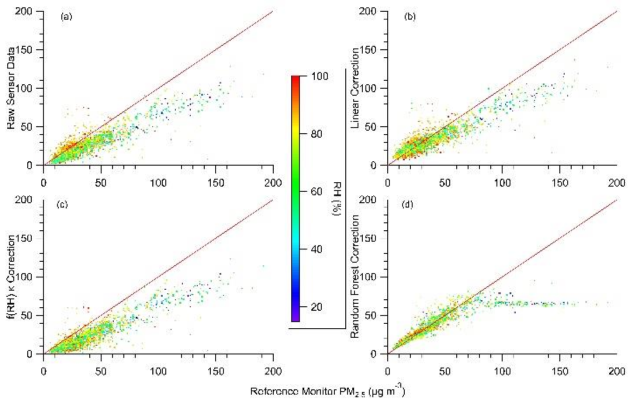

Figure 4 Raw low-cost sensor PM2.5 data (a), linear correction applied to LCS (b), f(RH) correction applied to LCS (c), and random forest correction applied to LCS (d) vs reference monitor PM2.5 data in Accra, Ghana with a 1:1 line added for visual aid. All points colored by ambient relative humidity.

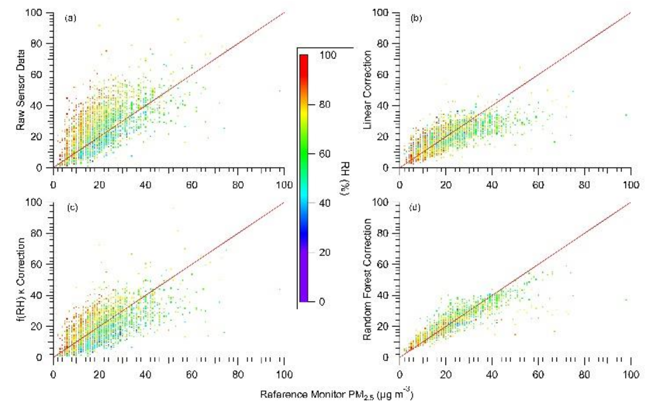

Figure 5 Raw low-cost sensor PM2.5 data (a), linear correction applied to LCS (b), f(RH) correction applied to LCS (c), and random forest correction applied to LCS (d) vs reference monitor PM2.5 data in Nairobi, Kenya with a 1:1 line added for visual aid. All points colored by ambient relative humidity.

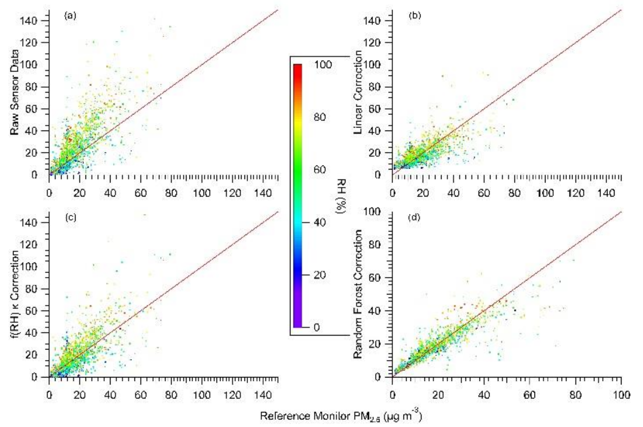

Figure 6 Raw low-cost sensor $PM_{2.5}$ data (a), linear correction applied to LCS (b), f(RH) correction applied to LCS (c), and random forest correction applied to LCS (d) vs reference monitor $PM_{2.5}$ data in Zamdela, South Africa with a 1:1 line added for visual aid. All points colored by ambient relative humidity.

	Nairobi				Accra				Zamdela			
	Raw	Linear	fRH	Random Forest	Raw	Linear	fRH	Random Forest	Raw	Linear	fRH	Random Forest
Absolute Bias (µg/m³)	55.73	32.28	40.12	14.28	33.59	26.68	48.12	12.47	78.90	47.34	55.68	22.91
Mean Normalized Bias	0.24	-0.02	-0.07	-0.01	-0.29	-0.08	-0.44	-0.07	0.45	0.00	0.07	-0.01
Mean Absolute Error (µg/m³)	8.35	5.02	6.51	2.52	10.61	7.38	14.24	4.94	10.93	5.41	7.30	2.97
Root Mean Squared Error (µg/m³)	11.32	7.05	8.77	4.57	15.28	11.49	18.45	12.78	17.26	8.20	11.77	5.39
Bias Corrected CvMAE	0.40	0.25	0.32	0.13	0.25	0.24	0.24	0.18	0.60	0.31	0.44	0.18
Pearson r	0.66	0.73	0.67	0.90	0.90	0.91	0.92	0.88	0.76	0.75	0.73	0.90

Table 1: Performance metrics for the three calibration methods and raw LCS data compared against reference monitor data at the three study locations. Metrics for each of the 4 data streams at each of the 3 locations are color coded relative to each other and best performance. Absolute and Mean Normalized bias are colored such that 0.0 is optimal performance; MAE, RMSE and CvMAE are colored so the lowest value is optimal; and r such that the highest value is optimal. The red color indicates the worst performance and green the best performance.

Unsurprisingly, raw, uncalibrated data perform the worst out of all the data streams when compared to reference monitor data, with the one exception of the r correlation in Zamdela, South Africa. However, a difference of 0.02 between r values (or 0.004 for r^2) is negligible and it can be said that neither the linear correction nor the *f(RH)* correction improves the linearity of the Zamdela data with respect to the reference monitor. The fact that these corrections do decrease bias and error significantly suggests that the correlation metrics are being thrown off by some systemic error or property of the Plantower low-cost sensor.[10] In general though, the low-cost sensors perform fairly well for linearity with out-of-the-box corrections applied by manufacturers (c_f = atm for the Plantower sensors, Quant-AQ's κ and ρ curves) which is encouraging as it continues to suggest that the use of these sensors is viable in the multitude of different meteorological, geographical, and anthropogenic and biogeogenic sources of pollution that comprise Africa as a continent (e.g. Giordano et al., 2021 and references therein; Raheja et al., 2022, 2023).

The performance of the linear correction models will not be discussed in detail here as the body of literature on that topic, even for the areas of study in this work, is well saturated (see Giordano et al., 2021 and references therein). An interesting fact to note from the collocation-based linear correction models is that these models are excellent at reducing mean-normalized bias to near-zero. Instead, we will compare the performance of the f(RH) and RF correction models with that of both the raw data and the linear models. While collocation-based linear correction models are quickly becoming seen as the minimum level of data processing that should be expected when presenting or analyzing data from these types of low-cost $PM_{2.5}$ sensors, most of the Global South simply does not have access to the required reference/research-grade instrumentation to construct these models. Therefore, the collocation-based linear model performance presented here acts as a "corrected baseline" against which the other two correction models can be compared.

Before discussing the f(RH) correction, it should again be stressed that this method *does not use*, *nor requires*, any collocation with a reference/research-grade instrumentation. Instead, the f(RH) model is based solely on freely-accessible data from EarthData (www.earthdata.nasa.gov) and the low-cost sensors themselves. This is important to reiterate as it contextualizes the relatively worse performance of the f(RH) models as compared against the linear models. In Nairobi and Zamdela, the f(RH) model performs worse than the linear model but only marginally so. The f(RH) model significantly decreases both bias (absolute and mean-normalized) and error (MAE, RMSE, CvMAE) compared against the raw/uncorrected LCS data. The f(RH) correction does not have any impact on linearity/correlation, which makes sense given the form of the correction and the fact that the model uses locally-unverified composition data from MERRA-2. For both Zamdela and Nairobi, this suggests that the $PM_{2.5}$ composition estimates are likely within reason, as discussed earlier, at least with respect to the method of calculating a bulk κ used here. There exist infinitely many combinations of $PM_{2.5}$ composition which would yield the same bulk κ so while this work cannot verify the performance of MERRA-2 reanalysis over Nairobi or Zamdela, it at least suggests the overall effects of bulk $PM_{2.5}$ hygroscopicity effects are well-enough-captured in the reanalysis dataset. For

Accra, however, the performance of the f(RH) model is 25-50% worse than simply using the raw data. The performance of the f(RH) model even slightly degrades if separate models are made for the dry and rainy seasons. Given MERRA-2's dust-dominated composition estimates over Accra for the entirety of the year, however, poor performance of a correction model driven entirely by the composition estimates is not unexpected (e.g. see SI Figure S4). Pointing out an exact reason for the poor performance is not possible with the data available but the same analysis centering around bulk κ may yield some insights. For the time period of this study, bulk κ averages around 0.4 for July-September, 0.2 for September and October, increases from 0.1 to 0.2 in November, and sharply decreases to less than 0.1 for December and January (SI Figure S6). This appears to be driven completely by the dust component although a full analysis of the input emissions inventories and regional emissions estimates that go into MERRA-2 would have to be performed to say for certain, which is well beyond the scope of this manuscript. It is also unclear whether this issue would affect all desert and desert-adjacent regions in Africa or if the Gulf of Guinea, or simply Accra itself, is uniquely affected. Other public data products that provide aerosol composition (e.g. GEOS-CF) may be useful where MERRA-2 is currently falling short but more work is needed to confirm this. Regardless of the reason for the poor performance of this model in Accra, however, it can be said that this methodology for creating a reference instrument-independent correction model shows promise for much of Africa and likely over the Global South as a whole - as long as MERRA-2 is approximately correct for bulk κ.

The random forest correction model represents the combination of both reference instrumentation and MERRA-2 composition estimates. The former provides overall mass corrections for the LCS whereas the latter refines and helps correct systemic biases inherent in the optical sensing used in LCS $PM_{2.5}$ sensors. Overall, the RF models perform extremely well. In all 3 study locations, bias is reduced to near-zero, error metrics are significantly reduced (50-80%), and even correlations are improved (although Accra remains an exception on this last point for the reasons discussed above). The tail that can be seen in Figure 3 on the RF correction is indeed the Harmattan dust event noted in Figure 1. This agrees with previous experience with random forest models for low-cost sensors that these calibrations can perform poorly at the upper extremes of the data ranges.[24] Forcing a correction on this data during model development could be possible but may also lead to over-constraining the model. Other techniques such as layering correction models (both machine learning and "simpler" methods) may improve performance but in this case it is difficult to disambiguate the effects of the manufacturer-implemented calibration for the Accra sensors (which includes hygroscopicity and density and therefore is an aerosol composition surrogate) and the MERRA-2 composition estimates. In summary, although this technique of combining machine learning and reanalysis data does yield excellent results, the caveats should not be ignored. Not only does this method require a reference instrument in the same way that the linear correction models do, but it also requires some expertise in both machine learning and utilizing reanalysis datasets. Combined with the fact that reference instrumentation is unavailable in much of the Global South; the machine learning techniques are easy to over-constrain; and the reanalysis sets have not been verified over much of the Global South; it is difficult to recommend this technique for general use before much more verification and validation has been done.

Conclusions

Overall, the results presented here show that utilizing reanalysis datasets to correct for aerosol composition effects can improve the performance of low-cost sensors in the Global South.. Even though the reanalysis dataset is not suitable for use in predicting overall PM loadings, the aerosol composition estimates lead to improved low-cost sensor performance over three very different areas in Africa when reference instruments are not available, and even more so if reference instruments are available. If we can better understand and potentially correct for the apparent overestimation of dust in Accra (and similar cities) by MERRA-2, we could significantly improve the usability of LCS in all regions where no reference instruments are available. Even now, the implications here for air quality monitoring across the Global South are very promising. Reanalysis datasets such as MERRA-2 are freely available and we show that correction methods based on these data (and established aerosol science) can be used to improve the performance of popular cheap PM sensors. Random forest regression using collocation and MERRA-2 aerosol composition can further significantly improve LCS performance. Unfortunately, this method and classical collocation-based linear regression require access to expensive and often unavailable/inaccessible reference instruments. In places where that access is limited, we show that using just the aerosol composition from MERRA-2 can be used to make real gains in LCS data quality by correcting for aerosol hygroscopicity which is a known artifact with light scattering-based PM sensors.

However, there is a major potential hurdle in this methodology: the accuracy of MERRA-2. MERRA-2 does seem to capture the Harmattan in Accra, Ghana but outside of the known Harmattan season, the composition data are questionable. Additionally, the MERRA-2 reanalysis dataset is only available 20 days after each month. For real-time air quality monitoring (which could improve forecasting and aid in proactive air quality controls), we could explore the corresponding monthly composition from previous years given the relative insensitivity of LCS hygroscopic corrections to exact aerosol composition. Improving the performance of MERRA-2 over Africa (e.g. with more aerosol composition measurements and better emissions inventories) would go a long way toward ensuring better quality data from rapidly proliferating LCS networks on the continent, which can then be used to make real, substantial progress towards clean air for all.

Author contributions

MRG, RS: conceptualization and project management, RS, MB: supervision and secured funding. MRG: curated data, developed methodology, conducted formal analysis, validated results, wrote the manuscript. EA, AFH, MJG, JN, MNN, SP, NW facilitated and supported local collocations. All authors reviewed and contributed to the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article, including low-cost sensor, reference data, and MERRA-2 data used as well as all codes are available at Zenodo at https://doi.org/DOI/10.5281/zenodo.11537585

References

1. Murray CJL, Aravkin AY, Zheng P, Abbafati C, Abbas KM, Abbasi-Kangevari M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet [Internet]. 2020;396(10258):1223–49. Available from: https://www.sciencedirect.com/science/article/pii/S0140673620307522
2. Fisher S, Bellinger DC, Cropper ML, Kumar P, Binagwaho A, Koudenoukpo JB, et al. Air pollution and development in Africa: impacts on health, the economy, and human capital. Lancet Planet Heal [Internet]. 2021 Oct 1 [cited 2022 Jul 11];5(10):e681–8. Available from: http://www.thelancet.com/article/S2542519621002011/fulltext
3. Amegah AK. Proliferation of low-cost sensors. What prospects for air pollution epidemiologic research in Sub-Saharan Africa? Environ Pollut. 2018 Oct 1;241:1132–7.
4. Kelly KE, Whitaker J, Petty A, Widmer C, Dybwad A, Sleeth D, et al. Ambient and laboratory evaluation of a low-cost particulate matter sensor. Environ Pollut. 2017 Feb 1;221:491–500.
5. Giordano MR, Malings C, Pandis SN, Presto AA, McNeill VF, Westervelt DM, et al. From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. J Aerosol Sci. 2021 Nov 1;158:105833.
6. Li J, Mattewal SK, Patel S, Biswas P. Evaluation of Nine Low-cost-sensor-based Particulate Matter Monitors. Aerosol Air Qual Res [Internet]. 2020 Feb 1 [cited 2021 Mar 24];20(2):254–70. Available from: https://aaqr.org/articles/aaqr-18-12-lcs-0485
7. Tryner J, Mehaffy J, Miller-Lionberg D, Volckens J. Effects of aerosol type and simulated aging on performance of low-cost PM sensors. J Aerosol Sci [Internet]. 2020 Dec 1 [cited 2020 Sep 23];703:105654. Available from: https://doi.org/10.1016/j.scitotenv.2019.135577
8. Feenstra B, Papapostolou V, Hasheminassab S, Zhang H, Boghossian B Der, Cocker D, et al. Performance evaluation of twelve low-cost PM2.5 sensors at an ambient air monitoring site. Atmos Environ. 2019 Nov 1;216:116946.
9. Hagan DH, Kroll JH. Assessing the accuracy of low-cost optical particle sensors using a physics-based approach. Atmos Meas Tech [Internet]. 2020 Nov 26 [cited 2020 Dec 21];13(11):6343–55. Available from: https://amt.copernicus.org/articles/13/6343/2020/
10. He M, Kuerbanjiang N, Dhaniyala S. Performance characteristics of the low-cost Plantower PMS optical sensor. Aerosol Sci Technol [Internet]. 2020 Feb 1 [cited 2020 Apr 9];54(2):232–41. Available from: https://www.tandfonline.com/doi/full/10.1080/02786826.2019.1696015
11. Petters MD, Kreidenweis SM. A single parameter representation of hygroscopic growth and cloud condensation nucleus activity. Atmos Chem Phys. 2007;7(8):1961–71.
12. Crilley LR, Singh A, Kramer LJ, Shaw MD, Alam MS, Apte JS, et al. Effect of aerosol composition on the performance of low-cost optical particle counter correction factors. Atmos Meas Tech [Internet]. 2020 Mar 10 [cited 2020 Apr 9];13(3):1181–93. Available from: https://www.atmos-meas-tech.net/13/1181/2020/
13. Malings C, Tanzer R, Hauryliuk A, Saha PK, Robinson AL, Presto AA, et al. Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. Aerosol Sci Technol [Internet]. 2020 Feb 1 [cited 2020 Apr 20];54(2):160–74. Available from: https://www.tandfonline.com/doi/full/10.1080/02786826.2019.1623863
14. Patel MY, Vannucci PF, Kim J, Berelson WM, Cohen RC. Towards a hygroscopic growth calibration for low-cost PM2.5 sensors. Atmos Meas Tech. 2024 Feb 13;17(3):1051–60.
15. Kuula J, Mäkelä T, Aurela M, Teinilä K, Varjonen S, González Ó, et al. Laboratory evaluation of particle-size selectivity of optical low-cost particulate matter sensors. Atmos Meas Tech. 2020 May 15;13(5):2413–23.
16. Buchard V, da Silva AM, Randles CA, Colarco P, Ferrare R, Hair J, et al. Evaluation of the surface PM2.5 in Version 1 of the NASA MERRA Aerosol Reanalysis over the United States. Atmos Environ. 2016 Jan 1;125:100–11.
17. Breiman L. Random Forests. Mach Learn 2001 451 [Internet]. 2001 Oct [cited 2022 Jul 13];45(1):5–32. Available from: https://link.springer.com/article/10.1023/A:1010933404324
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res [Internet]. 2011 [cited 2022 Jul 13];12:2825–30. Available from: http://scikit-learn.sourceforge.net.

19. Muyemeki L, Burger R, Piketh SJ, Language B, Beukes JP, van Zyl PG. Source apportionment of ambient PM10−2.5 and PM2.5 for the Vaal Triangle, South Africa. S Afr J Sci. 2021;117(5).
20. Hughes AF. Characterization and Source Apportionment of Airborne Particulate Matter in Some Urban Neighbourhoods of Accra, Ghana. 2014 [cited 2022 Sep 15]; Available from: http://localhost:8080/handle/123456789/7138
21. Gaita SM, Boman J, Gatari MJ, Pettersson JBC, Janhäll S. Source apportionment and seasonal variation of PM2.5 in a sub-Saharan African city: Nairobi, Kenya. Atmos Chem Phys. 2014 Sep 19;14(18):9977–91.
22. Raheja G, Sabi K, Sonla H, Gbedjangni EK, McFarlane CM, Hodoli CG, et al. A Network of Field-Calibrated Low-Cost Sensor Measurements of PM2.5in Lomé, Togo, Over One to Two Years. ACS Earth Sp Chem [Internet]. 2022 Apr 21 [cited 2024 Mar 15];6(4):1011–21. Available from: https://pubs.acs.org/doi/full/10.1021/acsearthspacechem.1c00391
23. Raheja G, Nimo J, K.-E. Appoh E, Essien B, Sunu M, Nyante J, et al. Low-Cost Sensor Performance Intercomparison, Correction Factor Development, and 2+ Years of Ambient PM2.5 Monitoring in Accra, Ghana. Environ Sci & Technol [Internet]. 2023 Jul 12;0(0). Available from: https://pubs.acs.org/doi/10.1021/acs.est.2c09264
24. Zimmerman N, Presto AA, Kumar SPN, Gu J, Hauryliuk A, Robinson ES, et al. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos Meas Tech [Internet]. 2018 Jan 15 [cited 2020 Dec 21];11(1):291–313. Available from: https://amt.copernicus.org/articles/11/291/2018/