ACS EARTH AND SPACE CHEMISTRY

http://pubs.acs.org/journal/aesccq

Article

Application of Gaussian Mixture Regression for the Correction of Low Cost $PM_{2.5}$ Monitoring Data in Accra, Ghana

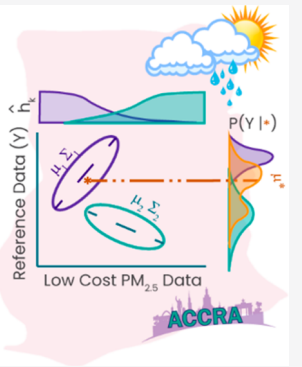
Celeste McFarlane, Garima Raheja, Carl Malings, Emmanuel K. E. Appoh, Allison Felix Hughes, and Daniel M. Westervelt*

Cite This: https://doi.org/10.1021/acsearthspacechem.1c00217

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Low-cost sensors (LCSs) for air quality monitoring have enormous potential to improve air quality data coverage in resource-limited parts of the world such as sub-Saharan Africa. LCSs, however, are affected by environment and source conditions. To establish high-quality data, LCSs must be collocated and calibrated with reference grade $PM_{2.5}$ monitors. From March 2020, a low-cost PurpleAir $PM_{2.5}$ monitor was collocated with a Met One Beta Attenuation Monitor 1020 in Accra, Ghana. While previous studies have shown that multiple linear regression (MLR) and random forest regression (RF) can improve accuracy and correlation between PurpleAir and reference data, MLR and RF yielded suboptimal improvement in the Accra collocation (R^2 = 0.81 and R^2 = 0.81, respectively). We present the first application of Gaussian mixture regression (GMR) to air quality data calibration and demonstrate improvement over traditional methods by increasing the collocated $PM_{2.5}$ correlation and accuracy to R^2 = 0.88 and MAE = 2.2 μg m^{-3}. Gaussian mixture models (GMMs) are a probability density estimator and clustering method from which nonlinear regressions that tolerate missing inputs can be derived. We find that even when given missing inputs, GMR provides better correlation than MLR and RF performed with complete data. GMR also allows us to estimate calibration certainty. When evaluated, 95% confidence intervals agreed with reference $PM_{2.5}$ data 96% of the time, suggesting that the model accurately assesses its own confidence. Additionally, clustering within the GMM is consistent with climate characteristics, providing confidence that the calibration approach can learn underlying relationships in data.



KEYWORDS: *low-cost sensors, particulate matter, air quality, Africa, Gaussian mixture regression*

1. INTRODUCTION

In 2019, ambient air pollution ranked fourth globally out of mortality risk factors, causing nearly 6.75 million premature deaths, with over 4 million of these deaths attributed to ambient $PM_{2.5}$, particulate matter mass concentrations for particles with diameters less than 2.5 μm.[1] Sparse air pollution monitoring, however, creates high uncertainty for estimates of exposure and impact.[2] Low-cost sensors (LCSs) have great potential for the improvement of air quality data coverage in resource-limited parts of the world such as sub-Saharan Africa.[3] LCSs, however, are sensitive to many factors including temperature, relative humidity, emissions source, and sensor aging.[4−7] By placing a LCS directly adjacent to a reference grade monitor, or collocating, calibration models can be used to develop correction factors for LCSs and establish high-quality data.

Various methods have been used to perform in-field calibrations for a variety of LCSs. The most studied method has been multiple linear regression (MLR), which is advantageous due to its simple, transparent nature.[8−12] Due to the complexity of atmospheric chemistry, however, it is important that the calibration method used can capture nonlinear relationships, which MLR is not able to do. For example, hygroscopicity, which causes particles to grow by absorbing water, is a nonlinear function of humidity.[13] Thus, nonlinear approaches such as random forest (RF) and k-nearest neighbors models have also been used to calibrate LCSs.[11,12,14,15] While both of these methods are able to capture complicated nonlinear relationships, application of the model beyond the training data is difficult. Another method that has been applied to LCS calibration is Gaussian process regression (GPR) which is advantageous due to its ability to successfully handle missing or incomplete data. While Nowack et al. found the application of GPR to LCS data to be successful, Malings et al. found that the high dimensional nature of the model created risks of overfitting.[12,16] Artificial neural networks have also been studied as a method to calibrate LCS.[12,17] Although neural networks are extremely versatile and are able to capture almost any nonlinear relationship, they require large amounts of training data

Special Issue: Mario Molina Memorial

Received: June 17, 2021
Revised: August 11, 2021
Accepted: August 12, 2021

which is not always accessible, particularly when working in the context of the air quality data gap in sub-Saharan Africa.

Gaussian mixture models (GMMs) are a method that models the joint probability density of input and output data as a mixture of Gaussian distributions.[18] GMMs have been shown to smoothly approximate almost any continuous probability density, including non-Gaussian data.[19,20] This is ideal when working with LCS data which can contain many distinct, complex relationships between variables, such as climate, aerosol type, and season, that cannot be represented by a single Gaussian. GMMs may be able to detect these heterogeneous relationships and appropriately cluster model components so they can be represented by normal distributions. GMMs have been successfully implemented in a variety of fields including speech recognition and robotic learning.[20,21] Recently, GMMs have also been used in geophysical data analysis to accompany atmospheric Lagrangian particle dispersion models.[22]

Gaussian mixture regression (GMR), first proposed in Ghahramani and Jordan (1993), works by modeling the probability density of the output data conditional to the input data as a GMM.[23] GMR does not directly model the regression function but rather derives the regression from the joint density model described by an initial GMM. The probabilistic nature of this method is what allows it to capture complex, nonlinear relationships. Given an input value, the regressed output value is a linear combination of nonlinear regressions performed on the mixture of Gaussians generated by the GMM. Each regression performed within a Gaussian component is predicated by a weight that describes the probability that the given input value belongs to the respective distribution. A key advantage of GMR is that it can handle missing data from unobserved explanatory variables. Regression can be performed on any subset of input-output data, which is ideal given the lack of air pollution data available in sub-Saharan Africa.

While GPR can also handle missing data, GMR contrasts GPR by treating the distribution of the data as multiple independent Gaussians. This allows GMR to better capture heterogeneous relationships when compared to GPR. These models also differ with respect to their treatment of time. In GMR, time-correlations between successive measurements are neglected, whereas GPR attempts to account for these correlations. Additionally, with GPR, the model user must specify their own priors based on existing understandings of relationships between explanatory variables. This contrasts GMR where priors are calculated with an expectation maximization algorithm. This makes GMR more accessible for applications without pre-existing knowledge of relationships within the data.

GMR has been used in the area of robot programming.[26−28] Newer applications of GMR include soft sensor technology as well as breast cancer prognosis predictions.[29,30] To our knowledge, GMR has yet to be used in air quality data calibration.

Here, we present a GMR calibration model for a LCS (PurpleAir monitor) to a United States Environmental Protection Agency Federal Equivalent Method (FEM) reference monitor (Met One beta attenuation monitor 1020) during collocation in Accra, Ghana. This collocation showed moderate initial correlation and moderate accuracy between reference and PurpleAir $PM_{2.5}$ data, indicating a need for calibration. Both MLR and random forest methods provided suboptimal improvement in correlation and accuracy, motivating the need for alternative calibration methods. Given GMR's ability to capture complex nonlinear relationships, handle missing data, and employ metrics able to reduce the risk of overfitting, it was chosen for this collocation. Here we first develop a GMR model using PurpleAir $PM_{2.5}$, temperature, and relative humidity (RH) data. We then evaluate the regressor's effectiveness with missing data. Additionally, we evaluate the GMM components in terms of its explanatory variables and seasonality to better understand LCS condition sensitivities. Finally, we evaluate the GMR model performance against MLR and RF performance.

2. METHODS

2.1. Sampling Location and Technology. A PurpleAir PA-II-SD low-cost $PM_{2.5}$ device (https://www.purpleair.com/) was deployed at the U.S. embassy in Accra, located at 5.58 N latitude and 0.17 W longitude, in the Cantonments neighborhood of Accra. PurpleAir monitor sampling began on March 1, 2020 and continues through the present. In addition, in January of 2020, a MetOne beta attenuation monitor (BAM) 1020 was located at the U.S. Embassy in Accra, providing a reference point for the PurpleAir monitor (https://www.airnow.gov/international/us-embassies-and-consulates/, last accessed 2021-05-28). For our calibration, we use the overlapping data set of BAM-1020 and PurpleAir from March 2020 to March 2021.

A PA-II sensor costs approximately $250 USD, which is nearly 100 times cheaper than reference $PM_{2.5}$ monitors. Their cost makes PurpleAir monitors an attractive tool to address the air pollution data gap. PurpleAir $PM_{2.5}$ has been shown to strongly correlate with reference grade monitors ($R^2 > 0.8$), subject to biases at high temperatures and relative humidity.[5,31,32] Accordingly, to achieve high quality data, PurpleAir monitors require careful field calibration.

PurpleAir monitors use dual Plantower PMS 5003 optical sensors to estimate $PM_{2.5}$ mass concentrations. Within each PurpleAir monitor, there are two sensors, labeled A and B, which measure PM mass concentrations at 120 s intervals. The $PM_{2.5}$ concentrations used in this study are the calculated averages of the concentrations reported by sensors A and B at each 120 s time interval. Previous work has suggested that these optical sensors function as nephelometers and report particle concentrations in six size bins, ranging from 300 nm to 10 μm, at 120 s intervals.[33] A proprietary algorithm converts raw sensor measurements to PM_1, $PM_{2.5}$, and PM_{10} mass using assumptions about particle shape and density. The PMS 5003 reports $PM_{2.5}$ values with correction factors (CFs). In this study, we use the "CF = ATM" PurpleAir $PM_{2.5}$ data field, which is actually the "CF = 1" data field due to mislabeling of columns by PurpleAir. While PurpleAir corrected this issue as related to its online data in late 2019, in this study the firmware updates from PurpleAir were never pushed to the offline device as we use PurpleAir data directly from the SD card due a lack of connectivity at the site. The CF = 1 data, unlike the other CF, is not transformed nonlinearly, which makes the CF = 1 data field a more appropriate raw input into regression models. Internal temperature, pressure, and RH are estimated by a Bosch BME280. Wireless connectivity is generally used to transmit data in real-time; however, in this study, due to lack of connectivity at the site, the sensors were operated in an offline mode.

2.2. Data Preparation. Data used included temperature, RH, and $PM_{2.5}$ data from a PurpleAir-II sensor. Hourly

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX

recorded $PM_{2.5}$ concentrations from a BAM-1020 were used as reference $PM_{2.5}$ values. Data were collected from March 2, 2020 to March 5, 2021. PurpleAir data from September 5, 2020 to September 24, 2020 were missing. This left 349 paired daily averaged BAM-PurpleAir data points. Models were built with the entirety of paired data points.

Data were cleaned by removing values outside the constraints of physical limitations. The maximum temperature in Accra, Ghana recorded from March 2020 to March 2021 was 32 °C and the minimum temperature was 23 °C.[34] Bosch BME280 sensors have been shown to report higher temperatures than ambient temperature values.[5,32,35] Accordingly, the data was restricted to PurpleAir temperature values between 16 and 49 °C. This removed 9 of 221 988 observations. Values with PurpleAir RH greater than 100% were also removed from the data set. This removed 4 of 221 988 observations. Observations with differences between PurpleAir sensors A and B greater than 20 $\mu g\ m^{-3}$ were removed from the data set. This was decided by visual inspection of the data. This removed 23 of 221 988 observations. Hourly reference grade data was subset to only include positive values less than 350 $\mu g\ m^{-3}$, based on the March 2018−June 2020 maximum recorded hourly $PM_{2.5}$ of 340 $\mu g\ m^{-3}$ in Kinshasa, DRC, another sub-Saharan African city.[11] This removed 24 of 9 396 hourly reference observations.

2.3. Gaussian Mixture Regression. *2.3.1. Gaussian Mixture Models.* Gaussian mixture models (GMMs) are a probability density estimator that can be used for clustering analysis.[18] Unlike other methods that group observations by similarity, GMMs model the joint probability density of the data as a mixture of normal distributions. At any given value, the probability density associated with an observation will have its own composition of the Gaussian components that make up the GMM. This composition is described with the probabilities that a given observation belongs to the distribution of each respective model component.

The observed data consist of N time points $\{x_1, ..., x_N\}$ each containing D features such as temperature, RH, and PurpleAir $PM_{2.5}$. Let X be an $N \times D$ matrix containing these data. We define dimensions of with input and output features as X^I and X^O, respectively. If each data point, x_n can be represented by a single multivariate Gaussian distribution, then the probability density function for the data matrix X is given by eq 1.

$$p(X|\mu, \Sigma) = p(X^I, X^O|\mu, \Sigma) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \Sigma) \tag{1}$$

Here, μ is the mean vector and Σ is the covariance matrix of a Gaussian distribution (denoted with $\mathcal{N}$). If X is from a multimodal distribution, as assumed in the case of GMMs, the probability density function is given by eq 2.[18]

$$p(X|\Omega) = \prod_{n=1}^{N} \sum_{k=1}^{K} h_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \tag{2}$$

Here, K is the number of components in the GMM and h_k is the probabilistic weight of the kth component (i.e., a number between 0 and 1, where $\sum_{k=1}^{K} h_k = 1$). The entirety of parameters describing the GMM is then given by eq 3.

$$\Omega = \{\{h_1, \mu_1, \Sigma_1\}, \{h_2, \mu_2, \Sigma_2\}, ...\{h_K, \mu_K, \Sigma_K\}\} \tag{3}$$

2.3.2. Expectation Maximization Algorithm. To build a GMM, we search for the most probable mixture of Gaussian distributions that represent the data.[18] We seek the maximum likelihood estimate of the model parameters, Ω^*, described by eq 4.

$$\Omega^* = \text{argmax}_{\Omega}[p(X|\Omega)] \tag{4}$$

Given an initial set of parameters, where in this case Ω is initialized by random sampling of the data series, we can apply an expectation maximization algorithm proposed by Ghahramani and Jordan.[23] By successively applying the expectation step (E-step) and maximization step (M-step), we can find Ω^*. The E-step evaluates the "responsibilities" of the model based on Ω. The responsibilities in a GMM are the likelihoods that a given point is characterized by a given model component. The M-step then updates the model parameters Ω. In this procedure, we introduce a latent variable z which is formulated in eq 5.

$$z_{nk} = \begin{cases} 1 & \text{if } x_n \text{ in component } k \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The probability that x_n is in component k, $p(z_{nk} = 1)$, is expressed by h_k. With this expression, the E-step can calculate the responsibilities of the model, γ, which is expressed by eq 6.

$$\gamma_{nk} = p(z_{nk} = 1|x_n) = \frac{p(x_n|z_{nk} = 1)p(z_{nk} = 1)}{\sum_{j=1}^{K} p(x_n|z_{nj} = 1)p(z_{nj} = 1)} \tag{6}$$

The maximization step then uses the responsibilities, γ, to calculate parameters Ω^* that maximize the likelihood function of eq 2. Briefly, we differentiate the logarithm of the likelihood with respect to each parameter and set the results equal to zero. We then solve for the model parameters to obtain eqs 7−9 for the M-step.

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma_{nk} x_n}{\sum_{n=1}^{N} \gamma_{nk}} \tag{7}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_{nk}(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma_{nk}} \tag{8}$$

$$h_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N} \tag{9}$$

We iterate between the E-step of eq 6 and the M-step of eqs 7−9 until the parameters have converged to their maximum likelihood estimates. This is the point where the model responsibilities, γ, have stopped changing in value.

2.3.3. Gaussian Mixture Regression. The fitted mixture model can now be used to predict the missing features of a data point x conditioned on the knowledge of the other features. We refer to the known features of the data point as input x^I and the unknown features as output x^O. In the present context, we seek to estimate missing output data for reference grade $PM_{2.5}$ given input data for PurpleAir $PM_{2.5}$, temperature and relative humidity. GMR uses the information obtained from modeling the joint probability distribution of the data, $p(x^I, x^O|\Omega)$, to estimate the conditional probability distribution for the output data given the input data, $p(x^O|x^I, \Omega)$. The regressed value is taken as a linear combination of the means of the posteriors from the conditional distributions of the output given the input from each component of the GMM.[23] Notably,

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX

these distributions can be conditioned over any subset of input values, allowing the model to handle instances of missing data.

The conditional distribution, $p(x^{O}|x^{I}, \Omega)$, for the output x^{O} given the input x^{I} is related to the joint distribution $p(x^{I}, x^{O}|\Omega)$ and the marginal distribution $p(x^{I} | \Omega)$ by the product rule of probability theory as formulated in eq 10.

$$p(x^{O}|x^{I}, \Omega) = \frac{p(x^{I}, x^{O}|\Omega)}{p(x^{I}|\Omega)} = \sum_{k=1}^{K} \hat{h}_k \mathcal{N}(\hat{\mu}_k, \hat{\Sigma}_k) \tag{10}$$

Importantly, this conditional distribution is also a Gaussian mixture of K components with probabilistic weights $\hat{h}_k$, mean vectors $\hat{\mu}_k$, and covariance matrices $\hat{\Sigma}_k$. The parameters of the conditional distribution, $\{\hat{h}_k, \hat{\mu}_k, \hat{\Sigma}_k\}$, are related to those of the GMM, $\Omega = \{h_k, \mu_k, \Sigma_k\}$, by eqs 11−13.[23]

$$\hat{\mu}_k = \mu_k^{O} + \Sigma_k^{OI}\Sigma_k^{I^{-1}}(x^{I} - \mu_k^{I}) \tag{11}$$

$$\hat{\Sigma}_k = \Sigma_k^{O} + \Sigma_k^{OI}\Sigma_k^{I^{-1}}\Sigma_k^{IO} \tag{12}$$

$$\hat{h}_k = \frac{h_k \mathcal{N}(x^{I}|\mu_k^{I}, \Sigma_k^{I})}{\sum_{k=1}^{K} h_k \mathcal{N}(x^{I}|\mu_k^{I}, \Sigma_k^{I})} \tag{13}$$

Here, the D-dimensional mean vector μ_k and the $D \times D$ covariance matrix Σ_k for the kth mixture component are decomposed in parts corresponding to the input and output features as

$$\mu_k = \begin{bmatrix} \mu_k^{I} \\ \mu_k^{O} \end{bmatrix} \tag{14}$$

and

$$\Sigma_k = \begin{bmatrix} \Sigma_k^{I} & \Sigma_k^{IO} \\ \Sigma_k^{OI} & \Sigma_k^{O} \end{bmatrix} \tag{15}$$

Our final predicted output value is then represented by the expectation and the covariance given by eqs 16 and 17.

$$\mathbb{E}(x^{O}) = \sum_{k=1}^{K} \hat{h}_k \hat{\mu}_k \tag{16}$$

$$\mathrm{cov}(x^{O}, x^{O}) = \sum_{k=1}^{K} \hat{h}_k(\hat{\Sigma}_k + \hat{\mu}_k \hat{\mu}_k^{T}) - \mathbb{E}(x^{O})\,\mathbb{E}(x^{O})^{T} \tag{17}$$

where the covariance is derived by the law of total covariance. Importantly, this is no longer the covariance of a Gaussian distribution but of the posterior conditional GMM.

A sketch of the GMR process is presented in Figure 1. The model responsibilities, $\hat{h}_k$, represented by eq 13, can be seen on the plot at the top of Figure 1. The purple and teal distributions in the plot on the far right of Figure 1 represent the conditional distributions of the output as expressed by eq 10. These conditional distributions are derived based on mean and covariance information gained from the expectation maximization algorithm (described in section 2.3.2) which characterizes the joint probability distribution of the input and output data. These joint distributions are represented by eq 1 and are shown in the bottom left plot in Figure 1. The peak of the conditional distributions in the far-right plot correspond to the mean of the respective conditional distributions.

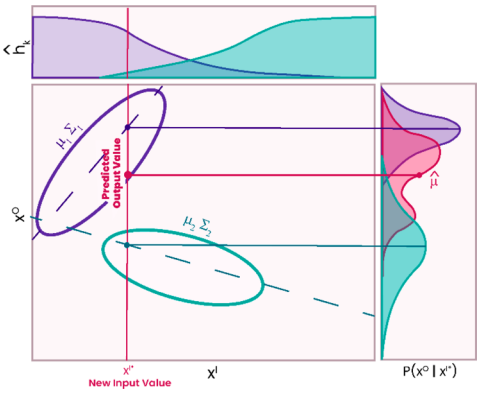


Figure 1. Sketch of a GMR performed on a GMM with two components, represented by teal and purple. The purple and teal ovals represent the joint probability distribution of the input and output data modeled by each component of the GMM. The probability that the specified input, $x^{I}*$, is in either model component is represented by $\hat{h}_k$. The probability of the output given the new input value, $x^{I}*$, for each component is shown in the plot of $p(x^{O}|x^{I}*)$ in purple and teal. The pink distribution on the plot of $p(x^{O}|x^{I}*)$ represents the conditional distribution of the output at point $x^{I}*$ from which we can derive the final predicted value, $\hat{\mu}$, shown by a pink dot.

Combining the mean of the conditional distributions with the model responsibilities, as formulated as in eq 16, gives us our predicted output value. This value is represented by the pink dot in the plot on the far left of Figure 1.

2.3.4. Model Implementation. An important consideration when using GMR is selecting the number of model components. Selecting too few components results in underfitting, while selecting too many creates a risk of overfitting the data. Various methods, including the Bayesian information criterion (BIC) and the minimum message length (MML), have been proposed to determine the ideal number of components in a model without succumbing to overfitting.[24,25]

Here, the Bayesian information criterion (BIC) was used to select the number of model components. The BIC evaluates a model's likelihood while introducing a penalty term for increasing parameters, helping address concerns of overfitting as well as model efficiency.[25] The number of components which minimizes the BIC is the most appropriate number of components for the model. This method of model selection has been used for GMMs in both Calinon et al. and Yuan et al.[27,29] Using the Sci-kit learn library in Python, the BIC was evaluated for GMMs built on training data with 1−10 components. These GMMs were built with daily averaged PurpleAir $PM_{2.5}$, reference grade $PM_{2.5}$, temperature, and RH data.

k-Fold cross validation is a method of cross validation that has previously been employed across LCS applications.[36,37] k-Fold cross validation works by splitting the sample into a number of folds and iteratively training the data on all but one fold. While k-fold cross validation has been suggested to introduce less bias than a simple training and testing data split when evaluating a model, it provides unique challenges when working with GMR. The first challenge is how to develop a singular set of model parameters as it is difficult to confirm the alignment of clusters generated from different training folds. To resolve this challenge, we employed an 80/20 training/testing data split for the purpose of generating model parameters and within the training data employed a 10-fold cross validation for model evaluation. It is important to choose

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX

sufficient folds for cross validating a GMR as the smaller the training set or the less the overlap between training sets, the greater the difference in the probability distributions of each training set. This would lead to a poor evaluation of the larger training set that generated the model parameters used for analysis. Performing a 10-fold cross validation process within the 80% training split resulted in each training fold compromising 72% of the entire data set. This is comparable to performing a 5-fold cross validation on the entire data set, where each training fold would compromise 80% of the data set, as done in previous studies.[37]

The second challenge is how to determine the number of model components to use within each training set. As each training set is best modeled by different probability distributions, the BIC for each training set differs. Accordingly, the results generated using a uniform number of model components during cross validation may be poor and misleading. To address this issue, we evaluated the BIC within each training fold and performed a GMR using the number of components suggested by the fold's criterion.

Multiple regressions were performed using daily averaged reference grade $PM_{2.5}$ as the output and using various subsets of daily averaged PurpleAir $PM_{2.5}$, temperature, and RH as inputs. Previous studies have shown that PurpleAir $PM_{2.5}$ correlates well with reference data motivating the use of PurpleAir $PM_{2.5}$ as an explanatory variable.[11] Relative humidity was selected because it has been shown to influence LCS performance due to hygroscopic growth of particles.[31] Temperature was selected because particulate concentrations may vary seasonally and temperature may provide insight into seasonality.[6,7,11,12]

The GMM was built and the regression was performed with Alexander Fabisch's gmr library in Python (https://github.com/AlexanderFabisch/gmr). Training a GMM with 279 points and performing a regression on new data can be written in as little as three lines of code and executed with a computational time of 0.56 s. This can be compared to the 0.11 s it takes to train and apply a MLR model on the same data. GMMs can be reproduced with this library by either selecting the same random state or storing and reusing the mean, covariance, and weight parameters generated by the GMM. Similar to MLR, model application has a computational time independent of the number of points used to train the model. If training a significantly large GMM that requires more notable computation time, one can simply adjust the convergence condition of the expectation maximization process.

The GMR output was evaluated based on the coefficient of determination (R^2), mean absolute error (MAE), bias-corrected mean normalized absolute error (cvMAE), bias, mean normalized bias (MNB), and 95% confidence intervals. Formulations for model evaluation metrics can be found below in eqs 18−22.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |\mathrm{PM}_{2.5,\mathrm{estimated},i} - \mathrm{PM}_{2.5,\mathrm{reference},i}|}{n} \tag{18}$$

$$\mathrm{cvMAE} = \frac{\sum_{i=1}^{n} |\mathrm{PM}_{2.5,\mathrm{estimated},i} - \mathrm{bias} - \mathrm{PM}_{2.5,\mathrm{reference},i}|}{\sum_{i=1}^{n} \mathrm{PM}_{2.5,\mathrm{reference},i}} \tag{19}$$

$$\mathrm{bias} = \frac{\sum_{i=1}^{n} \mathrm{PM}_{2.5,\mathrm{estimated},i} - \mathrm{PM}_{2.5,\mathrm{reference},i}}{n} \tag{20}$$

$$\mathrm{MNB} = \frac{\sum_{i=1}^{n} \mathrm{PM}_{2.5,\mathrm{estimated},i} - \mathrm{PM}_{2.5,\mathrm{reference},i}}{\sum_{i=1}^{n} \mathrm{PM}_{2.5,\mathrm{reference},i}} \tag{21}$$

$$95\%\mathrm{CI} = \mu \pm 1.96\ \mathrm{SE} \tag{22}$$

where SE represents the standard error of the population or the square root of the population variance, which can be derived from the covariances generated by the GMR in eq 17. Analyses performed here were done on daily averaged data since FEM/FRM designation is only applicable to daily averaged $PM_{2.5}$.[38] To provide a more thorough evaluation of GMR, however, we have also included an hourly model in the Supporting Information.

2.4. Multiple Linear Regression. For comparison with the GMR method outlined above, a multiple linear regression (MLR) approach and a random forest approach (described in section 2.5) were performed using Python's SciKit Learn package. Explanatory variable selection and cross validation methodology mirror that of the GMR approach described in section 2.3. The formulation for the model can be found below in eq 23, which is similar to the methodology of Malings et al. 2020. The MLR model was evaluated based on R^2, MAE, cvMAE, bias, and MNB; as this is a deterministic model, confidence intervals cannot be evaluated. Models were generated using daily averaged data.

$$\mathrm{PM}_{2.5} = \beta_0 + \beta_1\ \mathrm{PurpleAir\ PM}_{2.5} + \beta_2 T(^\circ\mathrm{C}) + \beta_3\ \mathrm{RH}(\%) \tag{23}$$

2.5. Random Forest. A random forest approach was also used to correct PurpleAir $PM_{2.5}$ data toward reference grade $PM_{2.5}$. A key advantage of random forest regression, in contrast to MLR, is the model's ability to capture nonlinear relationships between variables.[14] RF models work by employing a large number of decision trees, each constructed with a random bootstrapped sample from the training data set. Decision trees are constructed with a set of hierarchical rules that group inputs based on thresholds (specific conditions which stratify the data). The thresholds on these trees are called nodes. During the training, the tree nodes, the inputs to which they are applied and the associated order of application are calibrated. The final groupings of the input variables in a given decision tree, the "leaves", are then associated with the mean of the respective output variables. To predict an output given a new input, the rules developed during the training phase are applied to the new data. The output value associated with the resulting leaf becomes the prediction for that specific tree. The final predication is the mean output of all the trees in the model.

Here, we use a RF model constructed with 500 trees and PurpleAir $PM_{2.5}$ concentrations, temperature, and RH as input variables. By using a large number of trees, each constructed with a different assortment of values, the RF algorithm reduces the risk of overfitting. Additionally, the number of variables evaluated at each node can be determined by the model user. As this value increases, so does model accuracy and risk of overfitting. In this case, this value was set to one. We employ the same cross validation methodology as explained in section 2.3. Analyses were conducted on daily averaged data. This model was developed using Python's Scikit Learn package. The

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX

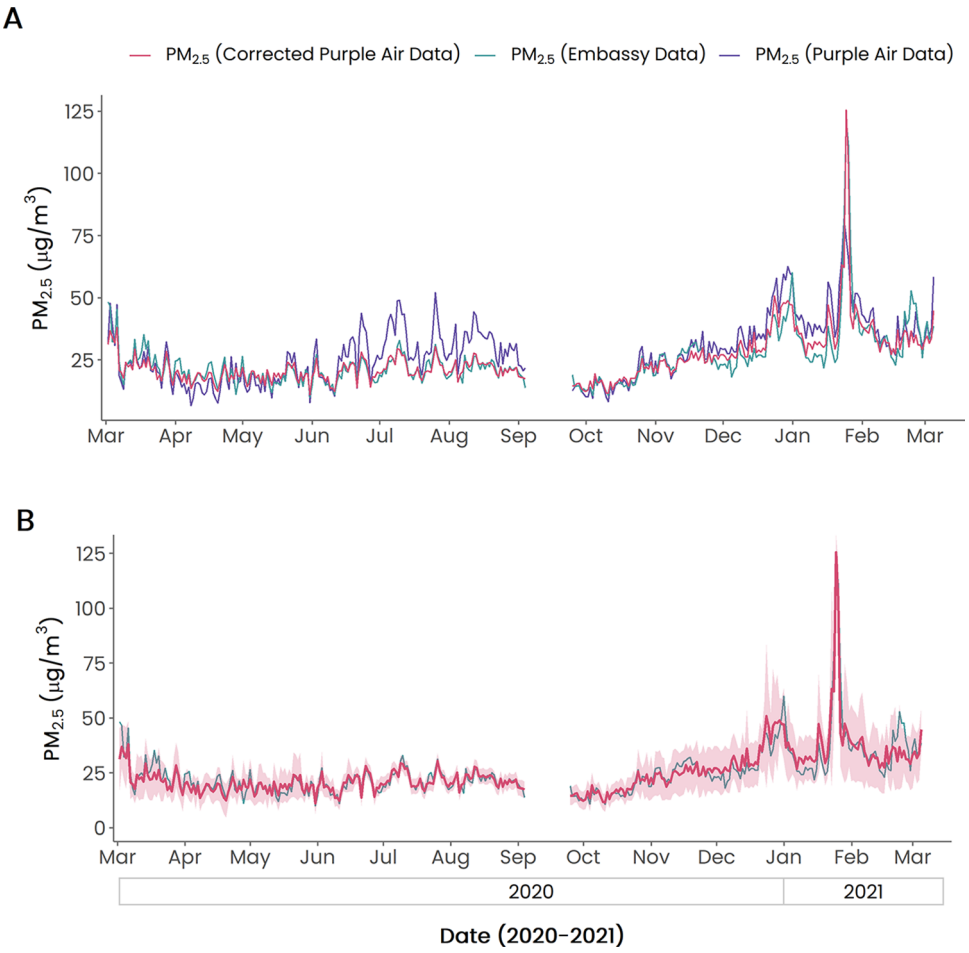


Figure 2. Performance evaluation and calibration of PurpleAir $PM_{2.5}$ data versus Federal Equivalent Method (FEM) $PM_{2.5}$ data between March 2020 and March 2021 at the U.S. embassy in Accra, Ghana. In part A, raw daily data are shown in purple, FEM data in teal, and GMR corrected low-cost sensor data in pink. Part B shows FEM data in teal, GMR corrected low-cost sensor data in pink, and 95% confidence intervals of daily corrected $PM_{2.5}$ values shaded in pink.

RF model was evaluated based on R^2, MAE, cvMAE, bias, and MNB, as with the MLR.

3. RESULTS AND DISCUSSION

3.1. Raw Data. Daily averaged PurpleAir $PM_{2.5}$ data showed moderate initial correlation with reference grade $PM_{2.5}$ data (R^2 = 0.53) and moderate initial error (MAE = 6.2 μ m^{-3}, μ = 28.5 μg m^{-3}). When visualized, the daily averaged raw data displayed distinct periods of the PurpleAir monitor over-predicting and underpredicting $PM_{2.5}$ when compared to reference grade data at the U.S. embassy in Accra. The PurpleAir monitor consistently underpredicts from March to May 2020 and then overpredicts from June to July 020. This is visualized in Figure 2a, which presents the raw PurpleAir $PM_{2.5}$ data, reference $PM_{2.5}$ data, and GMR corrected $PM_{2.5}$ data. This observation of different performance regimes is a key motivator to use a mixture model, as they are often applied to help capture distinct, heterogeneous relationships in data.[11]

3.2. Daily Averaged Gaussian Mixture Regression Fit. In order to select the ideal number of components for the GMM, the BIC was evaluated for GMMs built on training data with 1 to 10 components. The results of this analysis are presented in Figure 3. The number of components which minimize the BIC is considered ideal, supporting our choice of building a GMM with four components.

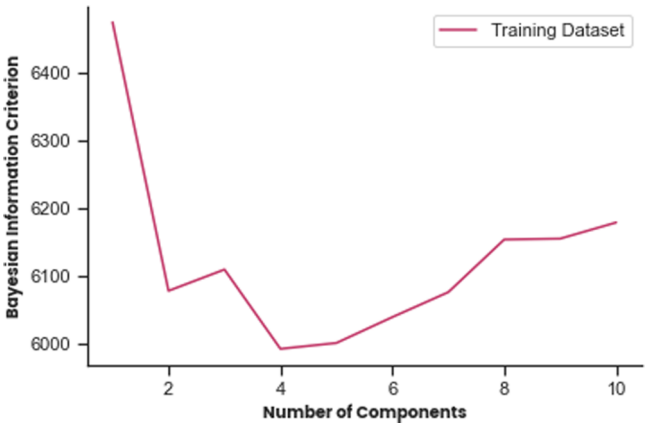


Figure 3. Bayesian information criterion evaluated for Gaussian mixture models with 1 to 10 model components built with daily averaged training data.

Table 1 presents the correlation and resulting error associated with GMRs conditioned on a four component GMM built with daily averaged training PurpleAir $PM_{2.5}$, reference grade $PM_{2.5}$, RH, and temperature data. GMR models were conditioned on various subsets of inputs. The use of all possible explanatory variables for which we had data as

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX

Table 1. Statistics of Daily Averaged $PM_{2.5}$ Data, MLR, RF, and GMR Calibration Models

model	model inputs	R^2	MAE ($\mu g\ m^{-3}$)	cvMAE	bias ($\mu g\ m^{-3}$)	MNB
raw data		0.53	6.2	0.24	3.4	0.14
GMR	PurpleAir $PM_{2.5}$, temperature, RH	0.88	2.2	0.09	0.43	0.02
	PurpleAir $PM_{2.5}$, temperature	0.85	2.5	0.10	0.85	0.04
	PurpleAir $PM_{2.5}$, RH	0.83	2.4	0.10	0.35	0.02
	PurpleAir $PM_{2.5}$	0.74	3.2	0.13	0.29	0.01
MLR	PurpleAir $PM_{2.5}$, temperature, RH	0.81	2.8	0.12	0.71	0.03
random forest	PurpleAir $PM_{2.5}$, temperature, RH	0.81	2.7	0.11	0.98	0.04

inputs (PurpleAir $PM_{2.5}$, RH, and temperature) yielded the best fit within the 80/20 training/testing data split (R^2 = 0.88, MAE = 2.20 $\mu g\ m^{-3}$, and bias = 0.43 $\mu g\ m^{-3}$). The results of the cross validation performed within the training data set are discussed in section 3.4, and a table of the performance metrics within each fold can be found in the Supporting Information.

In Figure 2, we plot daily averaged raw PurpleAir $PM_{2.5}$ data, reference $PM_{2.5}$ data from the Accra U.S. embassy, and $PM_{2.5}$ data corrected by the GMR model conditioned over PurpleAir $PM_{2.5}$, RH, and temperature data. In Figure 2a, we can see that the corrected $PM_{2.5}$ data have less definitive periods of underpredicting and overpredicting $PM_{2.5}$ when compared to raw PurpleAir values. We can also observe a very close fit between the corrected data and reference data in the initial six months of the timeseries. In November 2020, when corrected $PM_{2.5}$ values begin to increase, partially due to a seasonal reduction in precipitation and the onset of the Harmattan, characterized by low RH and large amounts of wind, the error between the corrected data and reference data begins to increase as well. Despite the increase in error, the overall distributions of the corrected data and reference data series closely align. The mean of both the corrected data series and the reference grade data series is 25.1 $\mu g\ m^{-3}$ with the standard deviations being 10.9 $\mu g\ m^{-3}$ and 11.6 $\mu g\ m^{-3}$, respectively. This suggests that GMR is an effective tool for providing accurate descriptive statistics of long-term variability in $PM_{2.5}$.

Another advantage to GMR is its ability to estimate the uncertainty of the model's predictions. Figure 2b presents the GMR-corrected $PM_{2.5}$ values, their 95% confidence intervals and reference grade $PM_{2.5}$ values for the entire time series. Confidence interval values range between ±3.4 $\mu g\ m^{-3}$ and ±441 $\mu g\ m^{-3}$. The median confidence interval for the data series is only ±6.9 $\mu g\ m^{-3}$. This is because the distribution of confidence interval values is right skewed by dates later in the time series. The wider confidence intervals are frequently associated with months after October 2020 when both $PM_{2.5}$ values and model error are larger. The reference grade $PM_{2.5}$ data falls within the 95% confidence interval of the GMR predictions 96% of the time, suggesting a high level of reliability for the model.

A key advantage of GMR is the model's ability to handle missing inputs. Using the GMM built with all inputs and removing all RH values from the GMR input yielded similar results to removing all temperature values from the GMR input. Both of these input subsets yielded better correlation and accuracy than both MLR and RF methods performed with all possible explanatory variables (Table 1). Using PurpleAir $PM_{2.5}$ alone as a GMR input, however, did not provide improved correlation or accuracy when compared to MLR or RF performed with all possible explanatory variables. These statistics provide a lower bound for how well GMR can handle missing inputs, as inputs were subset uniformly across the data set. In reality, there in fact may be different inputs missing across the data set when working with LCSs in real-time. Given that LCSs may be sensitive to variables other than temperature and RH, such as emission source profiles and aerosol size and composition, the model's ability to correct error despite missing data for additional potential explanatory variables is particularly notable.[5,31,39]

3.3. Daily Averaged Gaussian Mixture Model "Soft" Assignments. There may be a relationship between components within the GMR model and different climate conditions. Ghana has a tropical climate with two rainy seasons: from April through July and again in September and October. This contrasts with the Harmattan which occurs from late November to mid-March. In Figure 4, we analyze the

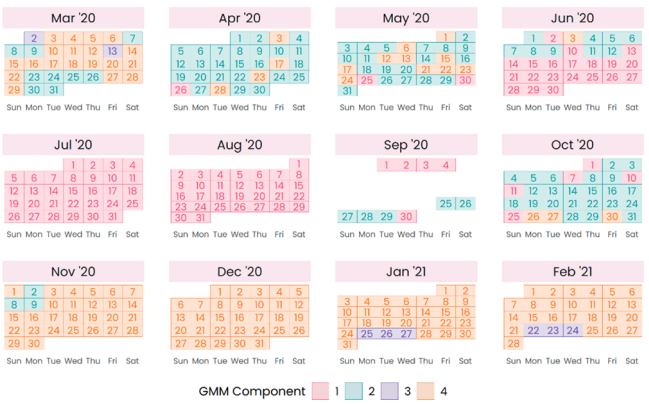


Figure 4. Calendar representation of GMM soft assignments prescribed by the component with the largest responsibility given an observation. Component 1 is shown in pink, component 2 in teal, component 3 in purple, and component 4 in orange.

GMM responsibilities in the context of these seasons. Figure 4 presents the dates of the timeseries represented by the Gaussian component to which they most probably belong, i.e., the observation's GMM soft assignment. These soft assignments were made based on the original GMM responsibilities conditioned on both input and output values. These assignments are not that of the GMR calibration and are not the responsibilities used in the regression. These assignments, however, can be helpful to understand the physical meaning behind the model's components.

Model soft assignments roughly correlate with seasons in Ghana. It is important to note that GMMs do not directly group observations based on similarity but based on the probability that they are characterized by the same normal distribution. Therefore, it is not expected for there to be perfect overlap between seasons and model soft assignments. The organization of components, however, does help provide

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX

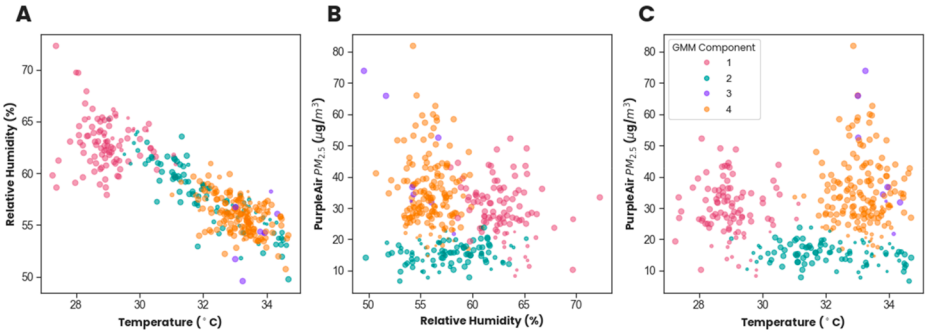


Figure 5. Gaussian mixture model soft assignments as a function of PurpleAir $PM_{2.5}$, RH, and temperature. Assigned components, displayed with color, are defined by the most probable Gaussian component to which an observation belongs. The size of each observation is scaled relative to the probability that a given point belongs to the assigned component. Component 1 is shown in pink, component 2 in teal, component 3 in purple, and component 4 in orange.

information on when in the year different LCS-reference grade correlations begin.

Observations in mid-June through July, the peak of the first rainy season, seem to have the most consistent model soft assignments, noted in Figure 4 by component 1 in pink. The November to February period, which encapsulates the Harmattan and was previously described as having a more notable error, is represented by component 4 in orange. This component also coincides with the portion of the model with the widest confidence intervals in the GMR model. The assignments in components 2 and 3, represented by teal and purple, lack a clear seasonal explanation, however.

We also analyze GMM soft assignments in the context of meteorological data and model error as shown in Figures 5 and 6. Figure 5 presents GMM soft assignments, shown by color, as a function of GMM inputs. Figure 6 presents GMM soft assignments as a function of both model error and PurpleAir $PM_{2.5}$ data error. The size of each observation is scaled relative to the probability that a given point belongs to the assigned component.

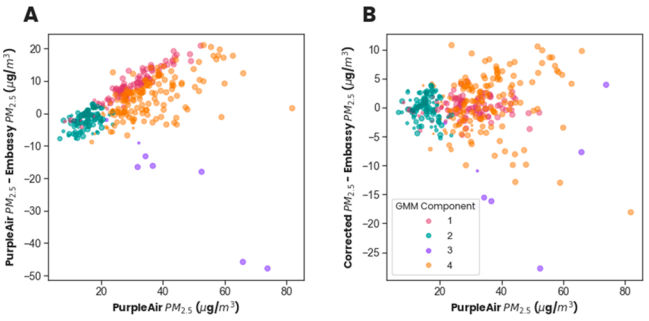


Figure 6. Gaussian mixture model soft assignments as a function of PurpleAir $PM_{2.5}$ and GMR-corrected $PM_{2.5}$ error. Assigned components, displayed with color, are defined by the most probable Gaussian component to which an observation belongs. The size of each observation is scaled relative to the probability that a given point belongs to the assigned component. Component 1 is shown in pink, component 2 in teal, component 3 in purple, and component 4 in orange.

Component 1, shown in pink, is associated with mid-June through August of 2020 and is characterized by high RH, low temperatures, and a positive PurpleAir $PM_{2.5}$ error. The positive error seen here is consistent with previous studies which have shown that PurpleAir sensors overpredict $PM_{2.5}$ concentrations at high RH due to hygroscopic growth of particles.[31] As PurpleAir monitors do not have a particle dryer on the inlet, this is expected. In Figure 6b, we can see that this positive bias was corrected to near zero. Component 2, shown in teal, is characterized by low PurpleAir $PM_{2.5}$ values, low PurpleAir $PM_{2.5}$ error, and low GMR-corrected $PM_{2.5}$ error. As component 2 solely consists of low PurpleAir $PM_{2.5}$ values, it may be acting as a "background" component for the model.

Component 3, shown in purple, is the smallest Gaussian distribution represented in this model and correlates to dates scattered throughout January and February of 2021. While this component exists at relatively high temperatures and PurpleAir $PM_{2.5}$ concentrations, when looking at Figure 5, component 3 is not visually distinctive from other components when considering meteorological variables. All of the observations in this component, however, are PurpleAir $PM_{2.5}$ values with a large negative error as shown in Figure 6a. This suggests that component 3 corresponds with very large reference grade $PM_{2.5}$ values which were underpredicted by the PurpleAir monitor. This could possibly be attributed to a local source generating small particles below the size detection limits of the PurpleAir.

Another hypothesis for the negative bias associated with this cluster is the impact of the Harmattan. During the Harmattan, heavy winds contribute dust to ambient particulate matter. It has been shown that the optical sensor within PurpleAir monitors, PMS-5003, performs poorly when detecting particles larger than approximately 0.8 μm.[40] Additionally, during a dust storm in Utah, Sayahi et al. found that PurpleAir monitors failed to report high particulate matter concentrations.[41] In this study, it was hypothesized that larger particles from windblown dust may have difficulty making the multiple 90-degree turns between the PMS-5003 inlet and the optical sensor. Given the 90-degree turns between the PMS-5003 inlet and the light sensor, we can also consider Mie theory's proposition that larger particles scatter less light per unit mass at 90 deg. Component 3 may correspond to days when the PurpleAir monitor struggled to detect high concentrations of windblown dust with a larger count median diameter than that of particulate matter measured under other conditions.

In Figure 6b, we can see that the model only incrementally corrected this large negative PurpleAir $PM_{2.5}$ error. This is consistent with previous descriptions of the data series in January and February of 2021, where the precision of the confidence intervals and accuracy of fit were low. The fourth component, shown in orange, is characterized by high

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX

temperatures, low to moderate RH, and moderate to high Purple Air $PM_{2.5}$. This component also correlates to the aforementioned period with lower model precision and accuracy (November 2020 to February 2021). From Figure 6b, we can see that while the model was able to correct the error for many of the observations in component 4, observations with moderate to high error exist, even after calibration.

The high error between PurpleAir $PM_{2.5}$ data and reference grade $PM_{2.5}$ data in components 3 and 4 suggest that PurpleAir monitors are subject to high errors at high temperatures and high $PM_{2.5}$ concentrations. This finding is consistent with previous studies.[5,32] While components 3 and 4 look similar climate-wise, a clear distinction is that component 3 consists of only observations that largely underpredicted $PM_{2.5}$, which is not the case for component 4. This distinction is significant enough for each component's observations to be better characterized by a differing distribution. The distinction between components 3 and 4, however, could also be driven by another variable for which we have no data, like particle size distribution. Additionally, the model responsibilities generated by the original GMM are made with both input and output data, differing from responsibilities formulated by the GMR. This is because the GMR is conditioned solely with input data. Since visually component 3 is better characterized by a relationship between the input and output data rather than relationships between input data, the GMR, in comparison to the GMM, may assign a higher responsibility to component 4 than to component 3. This would be because component 4 has similar meteorological characteristics to component 3. This may cause GMM characterized component 3 values to demonstrate a weaker fit when GMR is applied, as seen in Figure 6b.

Given PurpleAir sensitivities at both high RH and high temperatures, the model was better able to correct the data with high RH values compared to its correction of the data with higher temperatures. There may be additional unmeasured explanatory variables, such as wind speed or direction during the Harmattan, disturbing the model's effectiveness at higher temperatures.

3.4. Model Comparison. Due to their well-documented nature, MLR and random forest regression were initially selected as methods of improving data correlation and correcting error. The results of these efforts can be seen in Table 1. Within the 80/20 training/testing data split, MLR improved $PM_{2.5}$ correlation to $R^2 = 0.81$ and accuracy to MAE = 2.8 $\mu g\ m^{-3}$. The random forest model provided a similar level of correlation and accuracy ($R^2 = 0.81$ and MAE = 2.7 $\mu g\ m^{-3}$). These values contrast the GMR performance noted in section 3.2, where correlation from the 80/20 training/test split was improved to $R^2 = 0.88$ and accuracy to MAE = 2.2 $\mu g\ m^{-3}$. While all three models reduced bias to nearly zero, the bias of the GMR model was less than 60% of that of the MLR and RF models. A full time series comparison of all three models can be found in the Supporting Information.

For a more in-depth comparison, we can look at the results of each model's cross validation as shown in Table 2. Table 2 compares evaluation metrics of GMR to that of MLR and RF and reports the percentage of training folds with superior model performance. A full table of cross validation performance metrics can be found in the Supporting Information.

While the GMR had notably better correlation than MLR and RF with respect to the 80/20 training/test split, within the cross validation it only superseded MLR and RF correlation in approximately half of the folds. Additionally, the range of R^2 values within the GMR cross validation was greater than the range of those for MLR and RF. This indicates inconsistencies in the GMR models generated in the cross validation. This can possibly be explained by each model's mechanism. MLR and RF work by trying to directly generalize the relationship between variables, whereas GMR works by indirectly modeling the relationship between variables by evaluating probability density. Decreasing sample size increases deviations between the sample and population probability density. With smaller samples, the probability density models generated become more sensitive to outliers skewing the distribution. The inconsistences in the cross validation of the GMR are likely due to the presence of so few observations from component 3 (the smallest component) in the training fold. Without sufficient examples of component 3 in the initializing GMM, the resulting GMR may characterize what the larger model described with component 3, as members of other, dissimilar components. While increasing the sample size of training data improves the performance of most machine learning models, the results of the cross validation imply that the benefit of increasing training sample size is more notable for GMR than that for MLR and RF.

Table 2. Comparison of Model Performance during a 10-Fold Cross Validation of Daily Training Data

	% CV iterations	
evaluation metric	* = MLR	* = RF
GMR $R^2 \geq *R^2$	60%	50%
GMR MAE $\leq$ *MAE	90%	50%
GMR cvMAE $\leq$ *cvMAE	90%	70%
|GMR bias| $\leq$ |*bias|	60%	70%

Despite similar model correlation performances in the cross validation, GMR consistently outperformed MLR in terms of accuracy and bias as shown in Table 2. GMR also outperformed RF in terms of accuracy and bias but to a less notable extent than its comparison to MLR.

A key distinction between GMR, MLR, and RF is that GMR models can be trained with more explanatory variables than what is used for regression. This contrasts MLR and RF which cannot tolerate missing inputs. Given this distinction, we trained MLR models with subsets of explanatory variables to compare model performance across missing variables. If we train the MLR without temperature, we get a testing data correlation and accuracy of $R^2 = 0.78$ and MAE = 3.07 $\mu g\ m^{-3}$. This is a worse fit than removing temperature inputs from the GMR derived from the GMM built with all explanatory variables (Table 1). If we train the MLR without temperature or RH, we get a testing data correlation and fit of $R^2 = 0.71$ and MAE = 3.43 $\mu g\ m^{-3}$. This is a worse fit than removing temperature and RH inputs from the GMR derived from the GMM built with all explanatory variables (Table 1). If we train the MLR without RH, we yield a testing data fit of MAE = 2.60 $\mu g\ m^{-3}$ and $R^2 = 0.84$. This is a worse fit than removing RH inputs from the GMR derived from the GMM built with all explanatory variables (Table 1) but is a better fit than the MLR trained with all possible explanatory variables (Table 1). This suggests that RH was not an important factor in the original MLR model. This is in contrast with previous studies that have shown that RH impacts the technology used by PurpleAir to measure $PM_{2.5}$.[5,31,39] The lack of benefit brought by RH to the

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX

MLR model is likely a result of complex nonlinear relationships between variables that MLR cannot capture. Additionally, an MLR model inherently assumes that covariance between explanatory variables is insignificant, but the correlation in Figure 5a indicates that this is not a good assumption with respect to temperature and RH. This suggests that MLR is not a good physical model for the underlying relationships in the data.

4. CONCLUSIONS

Air pollution is a global public health crisis that causes millions of premature deaths in the world each year. This is a particularly challenging issue in sub-Saharan Africa where limited access to air quality monitoring creates a challenge of quantifying air pollution exposure and impact. Here we develop the first GMR model used to calibrate LCSs and subsequently apply it to an LCS-reference grade monitor collocation in Accra, Ghana. GMR is a regression method that is derived from GMMs, which model the joint probability density of input and output data. GMR proved more successful in correcting $PM_{2.5}$ data than both MLR and random forest methods with an R^2 of 0.88 compared to 0.81 and 0.81, respectively. It also proved successful in correcting the LCS $PM_{2.5}$ error, reducing MAE to 2.2 $\mu g\ m^{-3}$ from 6.2 $\mu g\ m^{-3}$. Additionally, when the GMR was applied to test data with missing inputs, it was able to successfully produce concentration estimates with correlation and accuracy similar to that of MLR or RF, whereas those latter methods cannot operate with incomplete input data. This is advantageous when working with air quality monitoring data in sub-Saharan Africa where there is limited data availability.

Due to GMR's probabilistic nature, it is able to provide confidence intervals for the model predictions, unlike MLR and RF. The range of those 95% confidence intervals were shown to include reference $PM_{2.5}$ 96% of the time. When evaluating the components of the GMM used to derive the GMR, component assignments were shown to match underlying climate characteristics, providing confidence that the model can detect underlying relationships in the data. With sufficient training data, a GMM may be able to have components for different pollutant mixes and climates and thus be applicable to multiple cities or regions. Future work, however, is needed to investigate the site transferability of the GMR model.

This work represents an important step in developing methods to provide high quality, low cost, accessible air pollution data to populations with limited access to reference grade air pollution monitors. While GMR offers improved model performance, its mechanism requires a more advanced understanding of statistics when compared to MLR. MLR remains advantageous for its simplicity and ease of understanding and may offer a more accessible understanding of LCS calibration than GMR. MLR, however, may not be suitable for data with heterogeneous correlations, like those seen here. GMR may be an ideal next step when MLR fails to meet desired performance goals or when provided a data set with distinctive seasonality, like the presence of the Harmattan in Accra. In the instance of a data set with ample missing data, GMR may also be ideal. Given that regression can be performed on any subset of inputs, GMR provides an opportunity to train models including other types of air pollution monitoring data, like emission source profiles, and build an advanced process without limiting future model applications. Future work, however, is needed to collect more diverse air quality data to build such a model.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsearthspacechem.1c00217.

Hourly models, cross validation of daily averaged models, and full time series comparison of daily models (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Daniel M. Westervelt − *Lamont-Doherty Earth Observatory of Columbia University, New York, New York 10027, United States; NASA Goddard Institute for Space Studies, New York, New York 10025, United States;* orcid.org/0000-0003-0806-9961; Email: danielmw@ldeo.columbia.edu

Authors

Celeste McFarlane − *Lamont-Doherty Earth Observatory of Columbia University, New York, New York 10027, United States;* orcid.org/0000-0003-2530-3051

Garima Raheja − *Lamont-Doherty Earth Observatory of Columbia University, New York, New York 10027, United States*

Carl Malings − *NASA Postdoctoral Program Fellow, Goddard Space Flight Center, Greenbelt, Maryland 20771, United States*

Emmanuel K. E. Appoh − *Ghana Environmental Protection Agency, Accra, Ghana*

Allison Felix Hughes − *Department of Physics, University of Ghana, Accra, Ghana*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsearthspacechem.1c00217

Author Contributions

C. McFarlane wrote the paper, created the figures, and performed all of the analysis. D.W. conceived and supervised the research. G.R. and C. Malings provided insights into data science and machine learning. E.K.A. and A.F.H. supported the device deployment in Accra and hosted D.W.

Notes

The authors declare no competing financial interest.

Data and Python scripts used to generate the models and analysis performed in this study are available at https://github.com/cmm2349/LCS-GMR.

■ ACKNOWLEDGMENTS

This work was funded by NSF OISE Grant Number 2020677. C. McFarlane acknowledges the Columbia Earth Institute's Research Opportunities for Undergraduates program. D.W. acknowledges support from the Columbia Center for Climate and Life. We wish to acknowledge two U.S. State Department employees at the U.S. embassy: Isaiah Tuolienuo and Jonathan Kelsey. The findings of this paper do not represent the views of the U.S. Department of State and are solely the views of the authors. C. Malings would like to acknowledge support by an appointment to the NASA Postdoctoral Program at the Goddard Space Flight Center, administered by the Universities Space Research Association (USRA) through a contract with

NASA. We acknowledge Professor Sylvain Calinon of Ecole Polytechnique Fédérale de Lausanne and Professor Kyle Bishop of Columbia University for helpful discussions and lecture materials regarding the GMR process.

■ ABBREVIATIONS

FEM, Federal Equivalent Method; FRM, Federal Reference Method; LCS, low-cost sensor; MLR, multiple linear regression; RF, random forest; GMM, Gaussian mixture model; GMR, Gaussian mixture regression

■ REFERENCES

(1) Health Effects Institute. *State of Global Air 2020*, 2020.
(2) Martin, R. v.; Brauer, M.; van Donkelaar, A.; Shaddick, G.; Narain, U.; Dey, S. No One Knows Which City Has the Highest Concentration of Fine Particulate Matter. *Atmospheric Environment: X* **2019**, *3*, 100040.
(3) Amegah, A. K. Proliferation of Low-Cost Sensors. What Prospects for Air Pollution Epidemiologic Research in Sub-Saharan Africa? *Environ. Pollut.* **2018**, *241*, 1132−1137.
(4) Hagan, D. H.; Kroll, J. H. Assessing the Accuracy of Low-Cost Optical Particle Sensors Using a Physics-Based Approach. *Atmos. Meas. Tech.* **2020**, *13* (11), 6343−6355.
(5) Tryner, J.; L'Orange, C.; Mehaffy, J.; Miller-Lionberg, D.; Hofstetter, J. C.; Wilson, A.; Volckens, J. Laboratory Evaluation of Low-Cost PurpleAir PM Monitors and in-Field Correction Using Co-Located Portable Filter Samplers. *Atmos. Environ.* **2020**, *220*, 117067.
(6) Levy Zamora, M.; Xiong, F.; Gentner, D.; Kerkez, B.; Kohrman-Glaser, J.; Koehler, K. Field and Laboratory Evaluations of the Low-Cost Plantower Particulate Matter Sensor. *Environ. Sci. Technol.* **2019**, *53* (2), 838−849.
(7) Bulot, F. M. J.; Johnston, S. J.; Basford, P. J.; Easton, N. H. C.; Apetroaie-Cristea, M.; Foster, G. L.; Morris, A. K. R.; Cox, S. J.; Loxham, M. Long-Term Field Comparison of Multiple Low-Cost Particulate Matter Sensors in an Outdoor Urban Environment. *Sci. Rep.* **2019**, *9* (1), 1−13.
(8) Holstius, D. M.; Pillarisetti, A.; Smith, K. R.; Seto, E. Field Calibrations of a Low-Cost Aerosol Sensor at a Regulatory Monitoring Site in California. *Atmos. Meas. Tech.* **2014**, *7* (4), 1121−1131.
(9) Jiao, W.; Hagler, G.; Williams, R.; Sharpe, R.; Brown, R.; Garver, D.; Judge, R.; Caudill, M.; Rickard, J.; Davis, M.; Weinstock, L.; Zimmer-Dauphinee, S.; Buckley, K. Community Air Sensor Network (CAIRSENSE) Project: Evaluation of Low-Cost Sensor Performance in a Suburban Environment in the Southeastern United States. *Atmos. Meas. Tech.* **2016**, *9* (11), 5281−5292.
(10) Malings, C.; Tanzer, R.; Hauryliuk, A.; Saha, P. K.; Robinson, A. L.; Presto, A. A.; Subramanian, R. Fine Particle Mass Monitoring with Low-Cost Sensors: Corrections and Long-Term Performance Evaluation. *Aerosol Sci. Technol.* **2020**, *54* (2), 160−174.
(11) McFarlane, C.; Isevulambire, P. K.; Lumbuenamo, R. S.; Ndinga, A. M. E.; Dhammapala, R.; Jin, X.; McNeill, V. F.; Malings, C.; Subramanian, R.; Westervelt, D. M. First Measurements of Ambient PM2.5 in Kinshasa, Democratic Republic of Congo and Brazzaville, Republic of Congo Using Field-Calibrated Low-Cost Sensors. *Aerosol Air Qual. Res.* **2021**, *21*, 200619−200619.
(12) Malings, C.; Tanzer, R.; Hauryliuk, A.; Kumar, S. P. N.; Zimmerman, N.; Kara, L. B.; Presto, A. A.; Subramanian, R. Development of a General Calibration Model and Long-Term Performance Evaluation of Low-Cost Sensors for Air Pollutant Gas Monitoring. *Atmos. Meas. Tech.* **2019**, *12* (2), 903−920.
(13) Vu, T. v.; Shi, Z.; Harrison, R. M. Estimation of Hygroscopic Growth Properties of Source-Related Sub-Micrometre Particle Types in a Mixed Urban Aerosol. *npj Climate and Atmospheric Science* **2021**, *4* (1), 1−8.
(14) Zimmerman, N.; Presto, A. A.; Kumar, S. P. N.; Gu, J.; Hauryliuk, A.; Robinson, E. S.; Robinson, A. L.; Subramanian, R. A Machine Learning Calibration Model Using Random Forests to Improve Sensor Performance for Lower-Cost Air Quality Monitoring. *Atmos. Meas. Tech.* **2018**, *11* (1), 291−313.
(15) Hagan, D. H.; Isaacman-Vanwertz, G.; Franklin, J. P.; Wallace, L. M. M.; Kocar, B. D.; Heald, C. L.; Kroll, J. H. Calibration and Assessment of Electrochemical Air Quality Sensors by Co-Location with Regulatory-Grade Instruments. *Atmos. Meas. Tech.* **2018**, *11* (1), 315−328.
(16) Nowack, P.; Konstantinovskiy, L.; Gardiner, H.; Cant, J. Towards Low-Cost and High-Performance Air Pollution Measurements Using Machine Learning Calibration Techniques. *Atmos. Meas. Tech.* **2021**, *14*, 5637−5655.
(17) Spinelle, L.; Gerboles, M.; Villani, M. G.; Aleixandre, M.; Bonavitacola, F. Field Calibration of a Cluster of Low-Cost Available Sensors for Air Quality Monitoring. Part A: Ozone and Nitrogen Dioxide. *Sens. Actuators, B* **2015**, *215*, 249−257.
(18) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, 2006.
(19) Hersch, M.; Guenter, F.; Calinon, S.; Billard, A. Dynamical System Modulation for Robot Learning via Kinesthetic Demonstrations. *IEEE Transactions on Robotics* **2008**, *24* (6), 1463−1467.
(20) Drews, P.; Núñez, P.; Rocha, R. P.; Campos, M.; Dias, J. Novelty Detection and Segmentation Based on Gaussian Mixture Models: A Case Study in 3D Robotic Laser Mapping. *Robotics and Autonomous Systems* **2013**, *61* (12), 1696.
(21) Lu, L.; Ghoshal, A.; Renals, S. Cross-Lingual Subspace Gaussian Mixture Models for Low-Resource Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2014**, *22* (1), 17.
(22) Crawford, A. The Use of Gaussian Mixture Models with Atmospheric Lagrangian Particle Dispersion Models for Density Estimation and Feature Identification. *Atmosphere* **2020**, *11* (12), 1369.
(23) Ghahramani, Z.; Jordan, M. I. Supervised Learning from Incomplete Data via an EM Approach. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann Publishers, 1994.
(24) Oliver, J. J.; Baxter, R. A.; Wallace, C. S. *Unsupervised Learning Using MML*, 1996.
(25) Fraley, C.; Raftery, A. E. How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. *Comput. J.* **1998**, *41* (8), 578−588.
(26) Cederborg, T.; Li, M.; Baranes, A.; Oudeyer, P.-Y. Incremental Local Online Gaussian Mixture Regression for Imitation Learning of Multiple Tasks. In *2010 IEEE/RSJ, International Conference on Intelligent Robots and Systems*; IEEE, 2010, DOI: 10.1109/IROS.2010.5652040.
(27) Calinon, S.; Guenter, F.; Billard, A. On Learning, Representing, and Generalizing a Task in a Humanoid Robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **2007**, *37* (2), 286−298.
(28) Tian, Y.; Sigal, L.; Badino, H.; de La Torre, F.; Liu, Y. Latent Gaussian Mixture Regression for Human Pose Estimation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin, Heidelberg, Germany, 2011; Vol. *6494*, pp 679−690, DOI: 10.1007/978-3-642-19318-7_53.
(29) Yuan, X.; Ge, Z.; Song, Z. Soft Sensor Model Development in Multiphase/Multimode Processes Based on Gaussian Mixture Regression. *Chemom. Intell. Lab. Syst.* **2014**, *138*, 97−109.
(30) Falk, T. H.; Shatkay, H.; Chan, W. Y. Breast Cancer Prognosis via Gaussian Mixture Regression. In *Canadian Conference on Electrical and Computer Engineering*; Institute of Electrical and Electronics Engineers Inc., 2006; pp 987−990, DOI: 10.1109/CCECE.2006.277570.
(31) Jayaratne, R.; Liu, X.; Thai, P.; Dunbabin, M.; Morawska, L. The Influence of Humidity on the Performance of a Low-Cost Air Particle Mass Sensor and the Effect of Atmospheric Fog. *Atmos. Meas. Tech.* **2018**, *11* (8), 4883−4890.
(32) Magi, B. I.; Cupini, C.; Francis, J.; Green, M.; Hauser, C. Evaluation of PM2.5 Measured in an Urban Setting Using a Low-Cost

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX

Optical Particle Counter and a Federal Equivalent Method Beta Attenuation Monitor. *Aerosol Sci. Technol.* **2020**, *54* (2), 147−159.
(33) Hagan, D. H.; Kroll, J. H. Assessing the Accuracy of Low-Cost Optical Particle Sensors Using a Physics-Based Approach. *Atmos. Meas. Tech.* **2020**, *13* (11), 6343−6355.
(34) *Accra, Ghana Weather Averages* https://www.worldweatheronline.com/accra-weather-averages/greater-accra/gh.aspx (accessed 2021-05-18).
(35) Holder, A. L.; Mebust, A. K.; Maghran, L. A.; McGown, M. R.; Stewart, K. E.; Vallano, D. M.; Elleman, R. A.; Baker, K. R. Field Evaluation of Low-Cost Particulate Matter Sensors for Measuring Wildfire Smoke. *Sensors* **2020**, *20* (17), 4796.
(36) Tryner, J.; Good, N.; Wilson, A.; Clark, M. L.; Peel, J. L.; Volckens, J. Variation in Gravimetric Correction Factors for Nephelometer-Derived Estimates of Personal Exposure to PM2.5. *Environ. Pollut.* **2019**, *250*, 251−261.
(37) Zimmerman, N.; Presto, A. A.; Kumar, S. P. N.; Gu, J.; Hauryliuk, A.; Robinson, E. S.; Robinson, A. L.; Subramanian, R. A Machine Learning Calibration Model Using Random Forests to Improve Sensor Performance for Lower-Cost Air Quality Monitoring. *Atmos. Meas. Tech.* **2018**, *11* (1), 291−313.
(38) United States Environmental Protection Agency. *Reference and Equivalent Method Applications: Guidelines for Applicants*, Raleigh, September 27, 2011.
(39) Westervelt, D. M.; Horowitz, L. W.; Naik, V.; Tai, A. P. K.; Fiore, A. M.; Mauzerall, D. L. Quantifying PM2.5-Meteorology Sensitivities in a Global Climate Model. *Atmos. Environ.* **2016**, *142*, 43−56.
(40) Kuula, J.; Mäkelä, T.; Aurela, M.; Teinilä, K.; Varjonen, S.; González, Ó.; Timonen, H. Laboratory Evaluation of Particle-Size Selectivity of Optical Low-Cost Particulate Matter Sensors. *Atmos. Meas. Tech.* **2020**, *13* (5), 2413−2423.
(41) Sayahi, T.; Butterfield, A.; Kelly, K. E. Long-Term Field Evaluation of the Plantower PMS Low-Cost Particulate Matter Sensors. *Environ. Pollut.* **2019**, *245*, 932−940.

https://doi.org/10.1021/acsearthspacechem.1c00217
ACS Earth Space Chem. XXXX, XXX, XXX−XXX