ACS ES&T Air

pubs.acs.org/estair

Article

Twenty Years of High Spatiotemporal Resolution Estimates of Daily $PM_{2.5}$ in West Africa Using Satellite Data, Surface Monitors, and Machine Learning

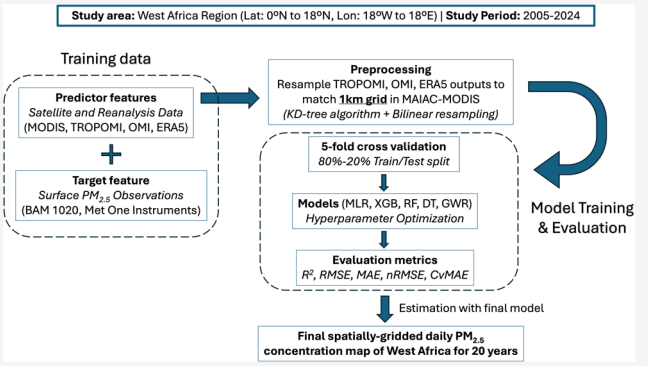
Daniel M. Westervelt,* Joe Adabouk Amooli, and Abhishek Anand

Cite This: https://doi.org/10.1021/acsestair.4c00366

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Estimates of air pollution mortality in sub-Saharan Africa are limited by a lack of observations of fine particulate matter ($PM_{2.5}$). Satellite data represents a promising solution with near-complete spatial coverage and high temporal coverage, but representativeness of surface conditions is a critical issue. Here we estimate surface $PM_{2.5}$ concentrations over West Africa at a daily, 1 km^2 spatiotemporal resolution based on satellite-derived and reanalysis inputs trained against surface $PM_{2.5}$ observations using several machine learning algorithms. Among machine learning models tested, Extreme Gradient Boosting (XGBoost) demonstrated the highest accuracy, with a 0.91 r^2, mean absolute error of 9.1 $\mu g\ m^{-3}$, and a CvMAE of 0.1, indicating about a 10% error across all sites on aggregate. Seasonal and annual $PM_{2.5}$ patterns were well captured, revealing severe air quality challenges via near-universal exceedances of World Health Organization air quality guidelines and interim targets. The data set's long-term perspective (2005−2024) highlights worsening air quality trends in both rural and urban areas. Our findings provide actionable data to support air quality management and policy, public health, and environmental justice initiatives in a critically underserved region of the world.



KEYWORDS: *particulate matter, satellite, data fusion, Africa*

1. INTRODUCTION

Outdoor fine particulate matter, known as $PM_{2.5}$ (particles with aerodynamic diameters less than 2.5 μm), poses a significant global public health challenge, contributing to an estimated 4.7 million premature deaths annually.[1,2] Beyond public health, $PM_{2.5}$ exposure is also a critical equity concern, as its effects disproportionately burden the poorest and most vulnerable segments of the population.[3] $PM_{2.5}$ is a complex mixture with both anthropogenic sources including several types of combustion and natural sources such as desert dust and sea spray. Gaseous pollutants, including troposphere ozone (O_3), carbon monoxide (CO), nitrogen oxides (NO_x = NO + NO_2), and sulfur dioxide (SO_2) are emitted by many of the same anthropogenic sources and also play a significant role in the health impacts of air pollution. SO_2 and NO_x, in particular, are important precursors to physical and chemical formation of $PM_{2.5}$ in the atmosphere. $PM_{2.5}$ is a subset of the ambient aerosol, which exert net cooling radiative effects on climate and impact clouds and the hydrologic cycle.[4,5]

In Africa, air pollution exposure has been linked to 1.1 million premature deaths annually.[6] In many African countries, environmental policy makers are limited in their ability to craft mitigation policies due to a dearth of reliable measurements.[7] This lack of $PM_{2.5}$ exposure data, coupled with a lack of health data, results in a high degree of uncertainty in burden of disease estimates in Africa, and a major blind spot for environmental managers and public health policymakers. The use of low-cost sensors for $PM_{2.5}$ and gaseous pollutants has rapidly proliferated across the African continent and around the world.[8−17] A key issue remains the reliability of this data, and the strengths and weaknesses of these devices are becoming better characterized.[18−22] While progress on data coverage has been impressive over the last 5−10 years, it is not feasible to have complete spatial coverage with hybrid networks of low-cost air sensors and reference monitors alone, especially in resource limited settings. Satellite-derived $PM_{2.5}$ estimates that are ground-truthed by available monitor and sensor measurements can help to fill the missing air pollution data and thus pave the way for effective policy, epidemiological studies, and environmental justice studies.

Received: December 29, 2024
Revised: June 27, 2025
Accepted: June 30, 2025

Estimating surface $PM_{2.5}$ from satellite data and modeling has received extensive research focus in recent decades.[23−27] Such estimates have proven worthy for a number of application areas, including providing exposure estimates for global burden of disease mortality estimates.[28] Methodologies for converting satellite products, which do not directly measure surface $PM_{2.5}$ but instead measure columnar Aerosol Optical Depth (AOD), include geophysical modeling approaches, statistical or machine learning (ML) modeling approaches, and hybrid approaches. The geophysical approach has been pioneered by van Donkelaar et al.,[23] Van Donkelaar et al.,[26] and references therein, providing annual and monthly surface $PM_{2.5}$ data sets for 1998-present, derived from a combination of satellite AOD retrievals, the GEOS-Chem chemical transport model, and geographically weighted regression. Statistical models have ranged from simple approaches such as multiple linear regression to more complex machine learning approaches.[29] Recent studies have leveraged multiple algorithms, including a neural network, random forest, and gradient boosting, which can be combined using a generalized additive ensemble model to estimate daily $PM_{2.5}$ for the contiguous U.S. with higher accuracy.[30] While these approaches are well documented and validated in North America, Europe, and Asia[25,31−35] the situation is much different for the African continent, where ground-level validation data is much more scarce.[36] Zhang et al.[37] used a random forest model to estimate daily $PM_{2.5}$ concentrations in South Africa, but the study geography was limited to only South Africa. Additionally, Léon et al.[38] use Moderate Resolution Imaging Spectroradiometer (MODIS) AOD and ground-based AErosol RObotic NETwork (AERONET) sun photometer data to estimate $PM_{2.5}$ in West Africa, but their analysis is limited to a small number of cities in Côte d'Ivoire and Benin. Finally, Amooli et al.[39] applied the AOD-$PM_{2.5}$ machine learning methodology specifically over the city of Ouagadougou and its environs, but did not expand further into the region.

Here we present a new daily, 1 km^2 spatiotemporal resolution surface $PM_{2.5}$ data set built on input satellite retrievals trained in an ML model against sparsely available surface $PM_{2.5}$ across West Africa. The West African domain is chosen as the most populated and fastest growing region of the African continent, with an estimated population of over 420 million people.[40] For the first time, we bring a West African spotlight on a now well-established and mature data estimation methodology, evaluating the accuracy of such methods against independent validation data sets. Beyond the new geography, we expand on previous studies in several ways. First, in addition to satellite AOD, we also include satellite retrievals of several aerosol precursors as inputs to the ML model, including NO_2, HCHO (formaldehyde), O_3, and SO_2. Adding these trace gas satellite retrievals as explanatory variables has been shown to improve accuracy of satellite-derived surface $PM_{2.5}$ predictions.[41] Second, we experiment with and intercompare a variety of popular ML methods, in order to determine which approach is best suited for this application. Third, we push the limits of spatial and temporal resolution possible, aiming for and achieving daily and 1 km^2 and extending back two decades prior to present day in the satellite record. The resulting evaluated surface $PM_{2.5}$ data set in West Africa will help close a critical data gap with actionable data for air quality managers, policymakers, and epidemiologists in crafting potential mitigation strategies and accessing long-term exposure data sets.

2. METHODS

Our geographic focus is West Africa, specifically from 0°N to 18°N latitude and 18°W to 18°E longitude, covering the majority of the Economic Community of West African States (ECOWAS).

2.1. Satellite and Reanalysis Data. Input satellite data is derived primarily from NASA and ESA products. For aerosol optical depth, we use the 550 nm wavelength AOD from MODIS using the MAIAC (Multi-Angle Implementation of Atmospheric Correction) retrieval algorithm on the Aqua/Terra satellites.[42] Currently, the native horizontal resolution of MAIAC-MODIS is 1 km × 1 km. MODIS has a daily overpass of 10:30 am local time for Terra and 1:30 pm local time for Aqua, with data extending back to about 2000 through present-day. We also leverage satellite retrievals of several aerosol precursors, including NO_2, HCHO, SO_2, and tropospheric O_3, from the Tropospheric Ozone Monitoring Instrument (TROPOMI). TROPOMI data comes at a 5.5 km × 3.5 km native resolution.[43] For data prior to TROPOMI launch (2017), we rely on OMI (Ozone Monitoring Instrument), flying on the NASA Aura satellite, which has a coarser spatial resolution of 13 km × 24 km.[44] Both OMI and TROPOMI have a 1:30 pm overpass time which aligns well with MAIAC. Finally, we used the European Center for Medium-Range Weather Forecasts Reanalysis data set v5 (ERA5) for input temperature (t2m), relative humidity (rh), total precipitation (tp), surface pressure (sp), and U−V wind components (u10, v10). U−V wind components and precipitation, as predictors of dust emissions, account for the large influence of Saharan dust on $PM_{2.5}$ in West Africa. The ERA5-Land daily aggregates have a spatial resolution of 11 km × 11 km and are available from 1950 to present.[45] All of the satellite and reanalysis data were retrieved from 2005 to 2024, with 2005 marking the first full year of satellite-based measurements of $PM_{2.5}$ precursors with OMI.

To address the mismatched spatial resolution in each satellite or reanalysis product, we created uniform gridded geospatial data sets over West Africa by resampling the ERA5-Land, TROPOMI, and OMI data sets to align with the 1 km resolution grid of MAIAC-MODIS. A *k*-dimensional tree algorithm (KD-tree) was used to efficiently locate the nearest neighbors in the original data sets relative to the target grid points. The KD-tree algorithm organizes spatial data into a hierarchical tree structure, facilitating rapid nearest-neighbor searches,[46] which is particularly advantageous when processing large geospatial data sets. After identifying the nearest neighbors, bilinear resampling was used to compute values on the finer grid. This method calculates the value at each target grid point as a weighted average of the four closest grid points from the original data set, with weights determined by the inverse distances between the target point and its neighbors. We combined the latitude and longitude grid information from all data sets into a single, uniform data set to facilitate machine learning analysis. Missing data is omitted and the machine learning models were trained using only those spatial grid cells and daily time steps where all independent variables and the dependent variable ($PM_{2.5}$) were available.

2.2. Surface Observations. The surface $PM_{2.5}$ data, which serves as training and testing data for the ML algorithm, was obtained from the AirNow US Department of State database, which has 10 locations in our domain at the time of writing (see Figure 5). Data is provided at an hourly time resolution

https://doi.org/10.1021/acsestair.4c00366
ACS EST Air XXXX, XXX, XXX−XXX

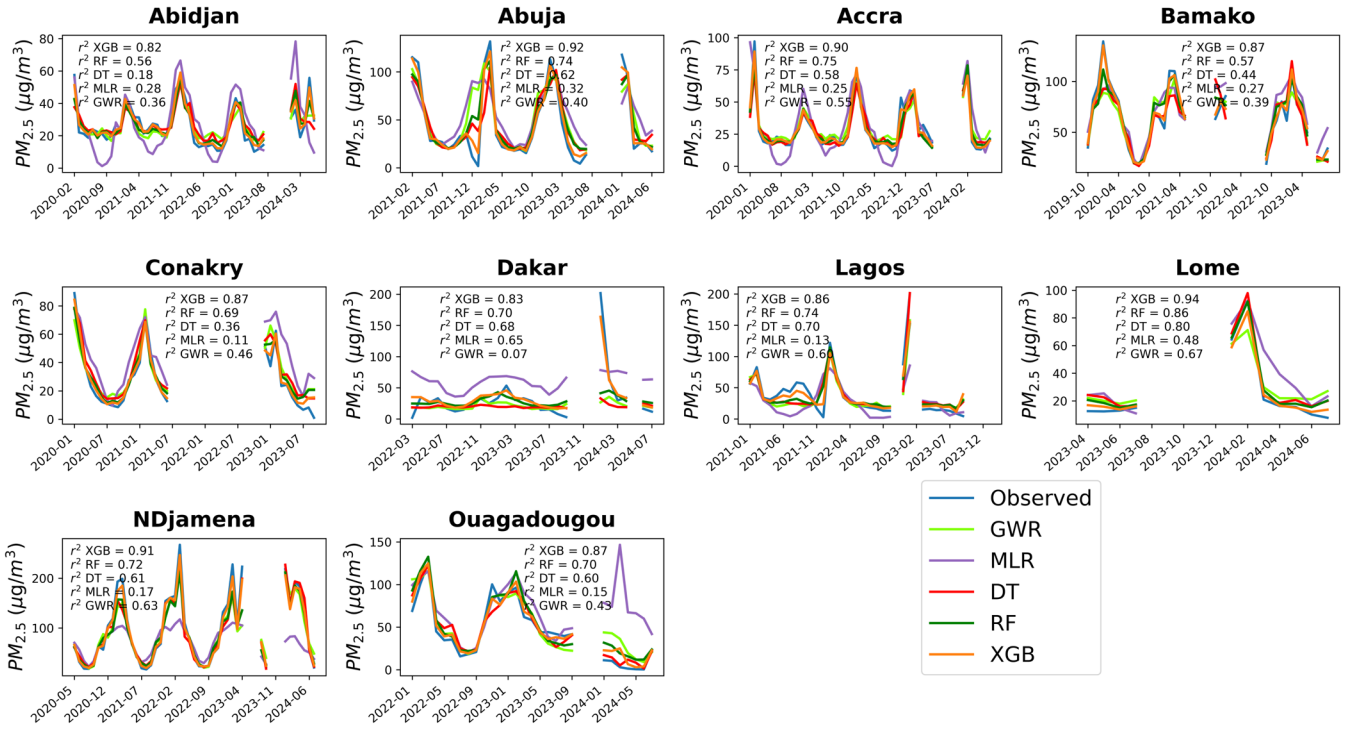


Figure 1. Monthly mean $PM_{2.5}$ concentrations from AirNow surface observations (blue), multiple linear regression prediction (purple), decision tree (red), random forest (green), XGBoost (orange), and GWR (bright green), for 10 major West African urban areas.

beginning in some locations as early as 2019, with newer locations flowing data beginning as late as 2023. All sites are equipped with a Met One Beta Attenuation Monitor (BAM 1020), which is a certified US EPA Federal Equivalent Method monitor for ambient $PM_{2.5}$ concentrations. The instrument determines $PM_{2.5}$ concentration by utilizing beta ray attenuation through a filter tape loaded with size-selected particles sampled from ambient air.[47] Only quality-assured data is used for the analysis. A map of the domain and cities with ground-based data is shown in Figure S1.

2.3. Machine Learning (ML) Techniques. We employ three different tree-based ML algorithms, Geographically Weighted Regression (GWR), and multiple linear regression as techniques for regressing the surface observations against satellite retrievals. In each case, the satellite data in the pixels where surface monitoring sites exist were first matched. Eighty percent of the data were randomly selected for model training and 20% was set aside for evaluation. Using a grid search technique, hyperparameters were set to obtain the optimal performance of the models. The models were evaluated using average performance metrics from a 5-fold cross-validation, including the average coefficient of determination (R^2), the root-mean-square error (RMSE), and mean absolute error (MAE), as well as the normalized RMSE (nRMSE) and coefficient of variation of MAE (CvMAE), as defined in Giordano et al.[8] These reported statistics are based on the 20% of data set aside and therefore represent a level of independence from the model training data. The trained models were used to estimate $PM_{2.5}$ concentrations in all other pixels, generating a spatially gridded resolution $PM_{2.5}$ data set over West Africa. Shapley Additive exPlanations (SHAP) analysis is performed to estimate feature importance in the machine learning models.[48]

Our implementation of multiple linear regression (MLR) uses Python scikit-learn to model the relationship between satellite retrievals and surface $PM_{2.5}$. The Linear Regression class leverages the least-squares algorithm to minimize the sum of squared residuals between the observed and predicted values.[49] We also use the scikit-learn library implementation of Decision Tree Regressor (DT), a nonparametric supervised learning algorithm suitable for both classification and regression tasks.[50] The model builds a tree structure where decision nodes represent feature-based decisions, and leaf nodes provide the predicted output. The tree recursively splits the data based on binary questions, ultimately predicting values at the leaf nodes. We fine-tuned the model using a hyperparameter grid search, setting a maximum depth of 10, minimum sample split and leaf of 2, verbosity of 2, random state of 100 for reproducibility, and n_jobs set to −1 for parallel computation.

We used the Random Forest Regressor from the scikit-learn library, a robust supervised ensemble learning algorithm for regression and classification tasks.[51] The algorithm constructs multiple decision trees using bootstrap sampling, where subsets of the training data are sampled with replacement. For regression, the final prediction is the average of the individual tree predictions, helping to reduce variance and improve model accuracy. This ensemble approach enhances prediction stability by leveraging the strength of multiple trees. To further optimize model performance, we fine-tuned the model through hyperparameter grid search to 100 estimators, maximum depth of 11, minimum sample split and leaf of 2, verbose of 2, random state of 100, and *n*-jobs of −1. We used the XGB Regressor class of the XGBoost library in Python to model the data. XGBoost is a powerful ensemble learning algorithm based on gradient boosting, renowned for its computational efficiency and scalability in large-scale machine learning tasks.[52] It optimizes decision tree models using advanced approximation techniques, enabling the identification of highly accurate models. XGBoost combines both linear and

https://doi.org/10.1021/acsestair.4c00366
ACS EST Air XXXX, XXX, XXX−XXX

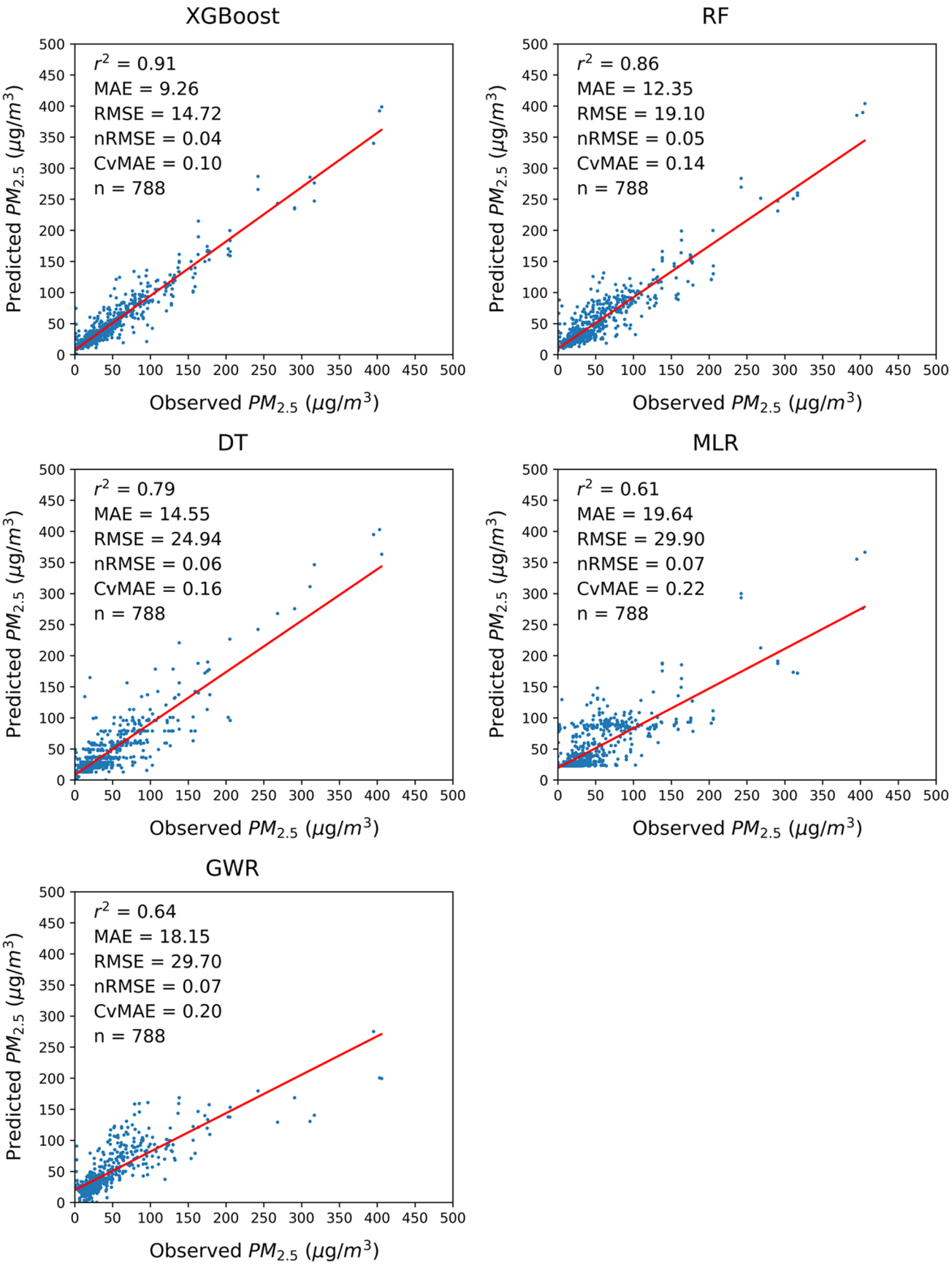


Figure 2. Sitewide aggregation of ML-predicted $PM_{2.5}$ versus AirNow Observations using XGBoost (upper left), Random Forest (upper right), Decision Tree (middle left), multiple linear regression (middle right), and Geographic Weighted Regression (lower left).

https://doi.org/10.1021/acsestair.4c00366
ACS EST Air XXXX, XXX, XXX−XXX

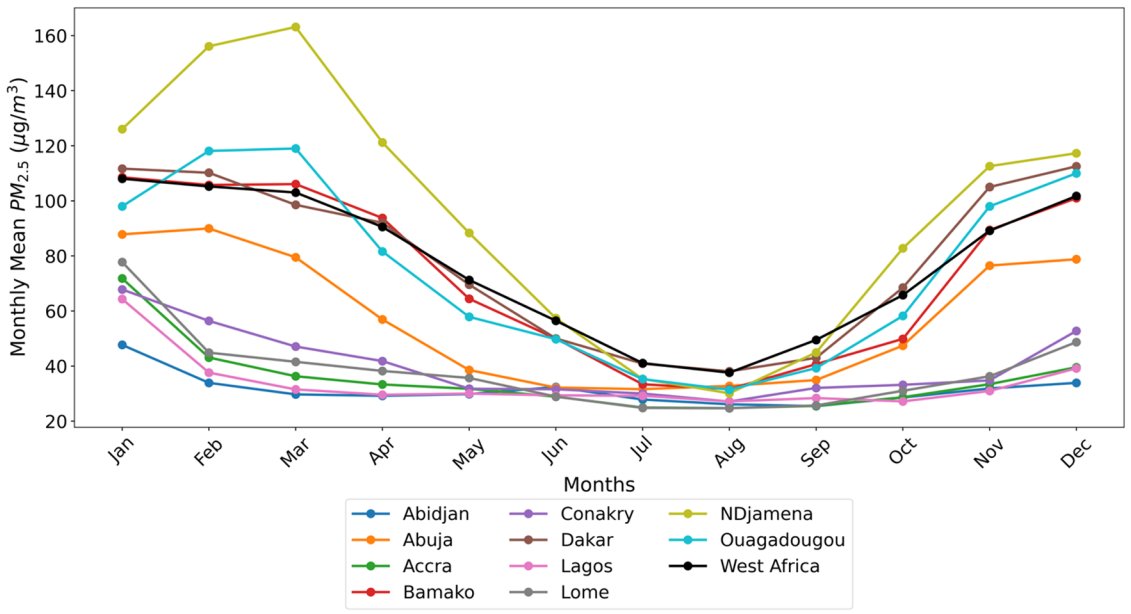


Figure 3. Monthly mean XGB-predicted $PM_{2.5}$ averaged over the entire data set (2005−2024) for each of the 10 locations and a spatially weighted domain average (black line).

tree-based learning algorithms, making it particularly effective for modeling complex, nonlinear relationships. We fine-tuned the model through a hyperparameter search, setting the following optimal values: 250 estimators, a learning rate of 0.02, a colsample_bytree of 0.9, a subsample of 0.5, a maximum depth of 14, a minimum child weight of 5, a γ of 0.5, a lambda of 0.1 (regularization parameters), minimum sample split and leaf values of 2, verbose of 2, a random state of 100, and n-jobs of −1.

The GWR is implemented in Python using the Multiscale Geographically Weighted Regression (MGWR) package in Python. The data were scaled using the StandardScaler package to ensure uniform scaling. The optimal spatial bandwidth was selected using the Sel_BW package, which determines the spatial influence each observation has on others. The GWR model was then fitted using the scaled data and their spatial coordinates.

3. RESULTS

Figures 1 and 2 show the model predicted monthly mean $PM_{2.5}$ concentrations for multiple linear regression (MLR, purple), decision tree (DT, red), random forest (RF, green), eXtreme Gradient Boosting (XGB, orange), and Geographically Weighted Regression (GWR, bright green) versus the AirNow observations (blue) at 10 urban locations within the domain. The best performing model at all sites is XGB, with r^2 ranging from 0.82 (Abidjan, Côte d'Ivoire) to 0.94 (Lomé, Togo), indicating a high degree of correlation with the observations. The XGB $PM_{2.5}$ predictions also exhibit low error, with MAE of 9.14 $\mu g\ m^{-3}$ (Figure 2) and RMSE of 14.1 $\mu g\ m^{-3}$ aggregated across all 10 sites. With mean concentrations of 89.6 $\mu g\ m^{-3}$, this represents about a 10% error aggregated across all sites or a CvMAE of 0.10 and an nRMSE of 0.03, where numbers closer to zero indicate higher accuracy. Generally, the next most accurate model is RF, followed by DT, followed by GWR, and finally followed by MLR. This order of performance is consistent with the nature of each of the tree-based methods, as XGB is an extension of RF by having new trees learn from the errors of previous trees (boosting), while RF is an improvement over DT, by combining multiple decisions trees over the single-tree approach of DT. MLR, on the other hand, is the most simple and interpretable method, and has been successfully implemented in some specific cases,[39] but is less generalizable and can fail when the relationship between satellite and ground observations vary substantially in space and time,[29] which is the case in our large West African domain. GWR is an improvement over MLR, which uses linear regression but allows for spatial nonstationary, which explains why it performs better than MLR. Figure 2 shows that while XGB is the best performing model by all metrics, each of the other tree-based models (RF and DT) have acceptable r^2 (0.86, 0.82), MAE (12.4, 14.2 $\mu g\ m^{-3}$), RMSE (18.8, 22.6 $\mu g\ m^{-3}$), nRMSE (0.05, 0.05), and CvMAE (0.14, 0.16) when all sites are aggregated together.

Figure 1 also provides an assessment of the seasonal and annual variability of the observations and the model-predicted $PM_{2.5}$ data sets. Missing data at several sites is due to instrument recalibration and maintenance downtime. Each model generally captures the seasonality, with large peaks (greater than 60 $\mu g\ m^{-3}$) common between the months of November and March, and local minima during the wet seasons of March-April and September-October. The large peaks are associated with the Harmattan circulation, in which dry, dusty air is transported from the Sahara toward the Gulf of Guinea and out to the Atlantic Ocean.[10] Sites further from the coast (e.g., Abuja, Ndjamena, Ouagadougou) generally have higher Harmattan peaks than coastal sites (Abidjan, Accra, Lomé).

Figure S2 shows the feature importance of each explanatory variable. In all three tree-based models, surface relative humidity (RH) is an important feature. RH can be impactful for PM2.5 prediction, especially in humid tropical environments such as West Africa where aqueous-phase formation of $PM_{2.5}$ and uptake of aerosol water especially for hygroscopic aerosol species is common.[53−55] Aerosol optical depth is the

https://doi.org/10.1021/acsestair.4c00366
ACS EST Air XXXX, XXX, XXX−XXX

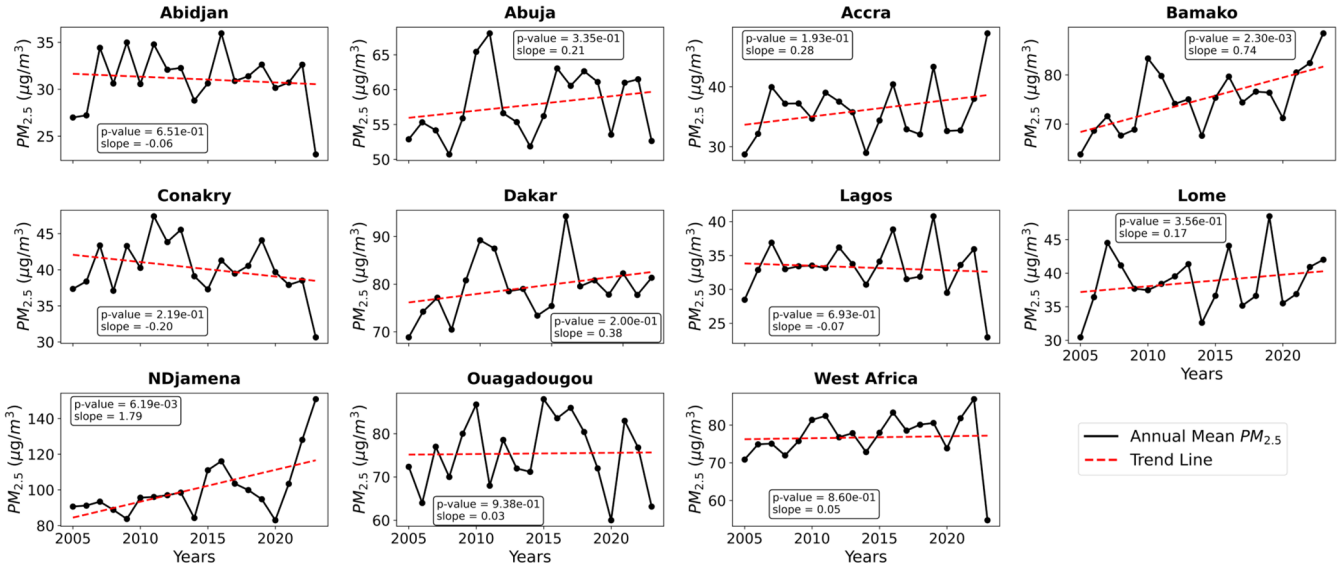


Figure 4. Annual mean trends in XGB-predicted $PM_{2.5}$ concentrations at the 10 urban centers and a regional average. Red lines indicate linear trendline.

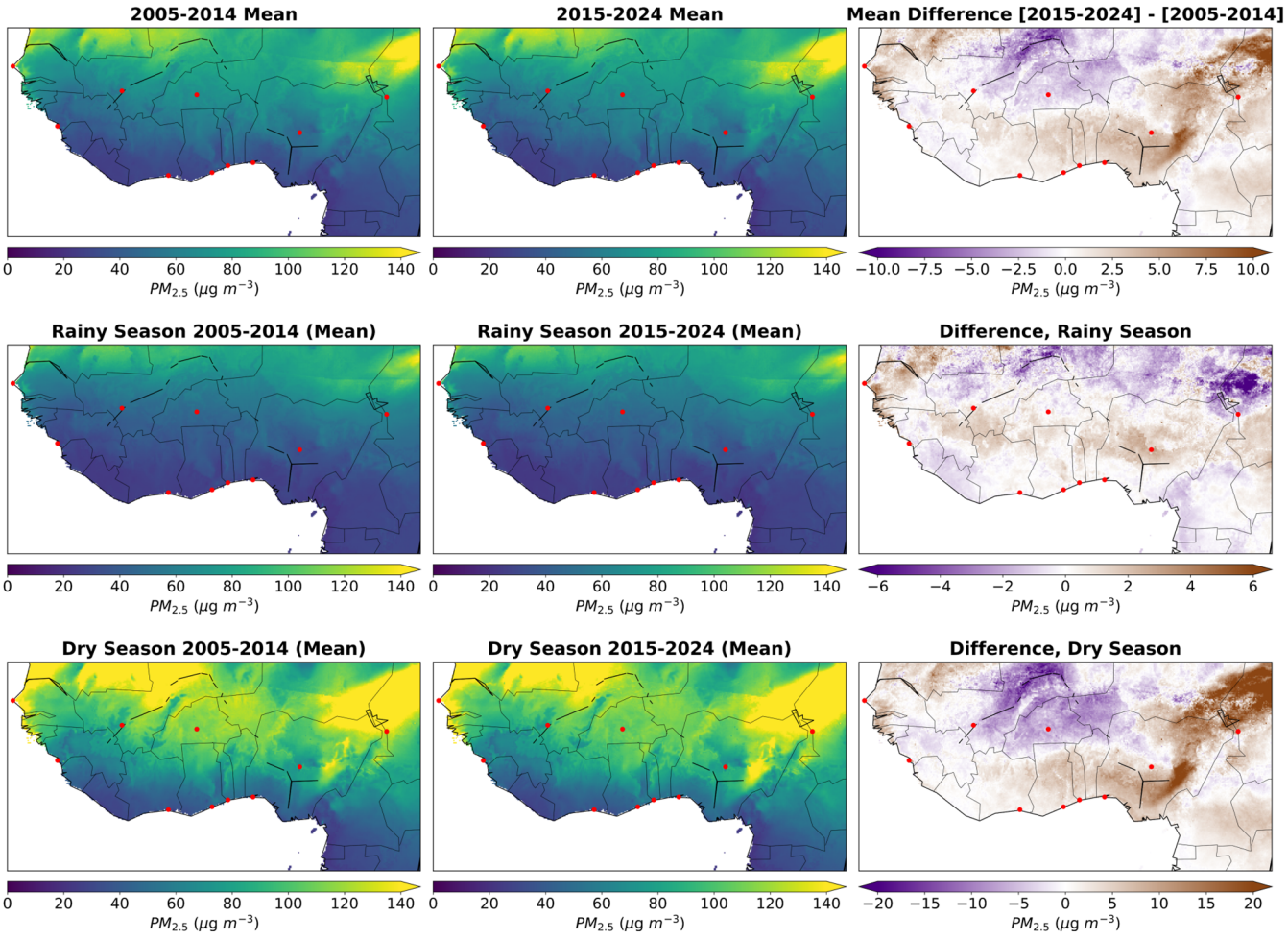


Figure 5. Spatial distribution of XGB-predicted $PM_{2.5}$ in the annual (top row), rainy season (middle row), and dry season (bottom row), for 2005−2014 mean (left column), 2015−2024 mean (middle column), and the difference (right column). Red dots indicate ground monitoring sites.

second most important feature in XGB and RF models, but is less important for DT. Trace-gas satellite retrievals have moderate importance, with SO_2 as the most important, likely due to its status as an important precursor to $PM_{2.5}$ with several sources in West Africa from high-sulfur content fuels for vehicles and a burgeoning oil and gas industry (especially in

https://doi.org/10.1021/acsestair.4c00366
ACS EST Air XXXX, XXX, XXX−XXX

Nigeria). Wind speed (either the u or v component), which is generally negatively associated with $PM_{2.5}$, and precipitation (tp), which is strongly negatively associated with $PM_{2.5}$,[53,54] are also of moderate importance with variability across the different tree-based methods. Given the higher accuracy of the XGB model compared to the others, further results are shown only for XGB.

Figure 3 shows the monthly means of $PM_{2.5}$ at all 10 cities (plus a domain-wide average, black line) for the years 2005−2024. Generally, $PM_{2.5}$ concentration minima occur around August (wet season) with values around 25−40 $\mu g\ m^{-3}$. Maximum concentrations are reported between November and March due to the Harmattan, with values ranging as low as 50 $\mu g\ m^{-3}$ (Abidjan) and as high as 160 $\mu g\ m^{-3}$ (Ndjamena). Exact months of minima and maxima can vary by 1 or 2 months due to location of the cities (coastal or far inland) with respect to the progression of the tropical rain belt. The monthly mean concentrations greatly exceed World Health Organization annual air quality guideline of 5 $\mu g\ m^{-3}$ and 24-h guideline of 15 $\mu g\ m^{-3}$ year-round, even in the "cleaner" wet season. Concentrations also exceed the more feasible WHO Interim Targets (annual means of 10−35 $\mu g\ m^{-3}$), with the exception perhaps of Interim Target 1 (35 $\mu g\ m^{-3}$), which would only not be exceeded during the wet season months in major West African urban areas. These results indicate severely impaired air quality not only due to Harmattan dust but also in non-Harmattan seasons where anthropogenic emissions are the driving factor for poor air quality.

One particular advantage of this data set is the ability to extend the data set back through the satellite record, allowing for insight into longer-term trends of air quality, which are extremely scarce in West Africa. Since this extension requires temporal stationarity assumptions, we first evaluate our historical $PM_{2.5}$ predictions against a widely used, validated historical $PM_{2.5}$ data set from van Donkelaar et al.,[26] and references therein, a satellite-derived $PM_{2.5}$ data set hosted at Washington University. Figure S3 shows the historical comparison between our predicted PM2.5 from 2005 to 2022 across 10 cities and a regional average. Regionally averaged across West Africa, the historical predicted $PM_{2.5}$ using our methodology has high correlation and low bias (r^2 = 0.73, MAE = 7.39 $\mu g\ m^{-3}$) with the widely used data product of the Atmospheric Composition Analysis Group.[26,56] Figure 4 shows the annual mean concentration at each of the 10 major cities plus the West African regional average beginning in 2005 and extending to 2024. In general, slope values are positive ranging from −0.2 $\mu g\ m^{-3}\ yr^{-1}$ (Conakry) to 1.79 $\mu g\ m^{-3}\ yr^{-1}$ (Ndjamena), indicating slightly worsening air quality over the last 20 years. Seven of the 10 cities studied exhibit worsening air quality going back to 2005, including major economic and population centers of Accra (0.28 $\mu g\ m^{-3}\ yr^{-1}$), Dakar (0.38 $\mu g\ m^{-3}\ yr^{-1}$), and Abuja (0.21 $\mu g\ m^{-3}\ yr^{-1}$). *P*-values are shown for each slope, with the most statistical significance apparent for Bamako (0.74 $\mu g\ m^{-3}\ yr^{-1}$ at p = 0.0023) and Ndjamena (1.79 $\mu g\ m^{-3}\ yr^{-1}$ at p = 0.0062). Slight decreases in $PM_{2.5}$ trends are observed for Lagos and Abidjan; however, each of these slopes have the least statistical significance (p = 0.69, p = 0.65, respectively). $PM_{2.5}$ concentrations are apparently decreasing by −0.20 $\mu g\ m^{-3}\ yr^{-1}$ in Conakry with p = 0.22, which would not be considered statistically significant by conventional thresholds. There is a substantial amount of interannual variability, which is likely tied to the strength of the Harmattan season in each of the cities.[10,16,17] The Harmattan is highly variable on a yearly basis and strongly dependent on meteorological and climatological factors.

Figure 5 shows the spatial distribution of 2015−2024 mean, 2005−2014 mean, and mean difference $PM_{2.5}$ concentrations across the whole domain, for the annual mean, rainy season, and dry season. $PM_{2.5}$ levels are much higher in the dry season and generally higher toward the northern part of the spatial domain. Highest concentrations are seen in the Sahel region, including northern Nigeria, Mali, Niger, and Burkina Faso, compared to the lower concentrations of the coastal areas. The third column of Figure 5 shows the differences in $PM_{2.5}$ concentrations between the most recent 10 years (2015−2024) and the 10-year period before that (2005−2014). These decadal averages ideally limit the impact of natural variability in weather which can have especially strong impact on Harmattan dust and thus air quality in this region. We find large increases in $PM_{2.5}$ throughout almost all of Nigeria, Africa's most populated country, and also in Chad (∼20 $\mu g\ m^{-3}$ increase). Increases are not necessarily limited to major urban population centers, highlighting the importance of air quality monitoring in both rural and urban areas in West Africa. Annual mean changes are dominated by changes in the dry season, and in some subregions that show substantial dry season increases also can exhibit rainy season decreases that are offset in the annual mean. Mali and Burkina Faso, in particular, have the most substantial decreases in $PM_{2.5}$ over this period, with values as high as −15 $\mu g\ m^{-3}$. Overall, however, concentrations appear to be increasing between these two decadal periods. Figure S4 shows the same information but zoomed into city level. Here we find increases of 33.8% in Accra but decreases in Bamako, Mali and Ouagadougou, Burkina Faso of −5.9% and −6.2% respectively.

4. CONCLUSIONS

We present, validate, and analyze a high spatiotemporal resolution (1 km^2, daily resolution) gridded $PM_{2.5}$ data set for West Africa extending back two decades. The methodology takes as input satellite retrievals of air pollutants from a variety of NASA and ESA products and trains a machine learning model against a network of ground-based monitors throughout the region. Using Extreme Gradient Boosting (XGB) as the ML model results in high accuracy when compared to surface observations, with r^2 of 0.91, MAE of 9.14 $\mu g\ m^{-3}$, RMSE of 14.1 $\mu g\ m^{-3}$, nRMSE of 0.03, and CvMAE of 0.10. The resulting data set reveals high levels of $PM_{2.5}$ exposure throughout West Africa that exceed WHO air quality guidelines and most interim targets, peaking during Harmattan season and extending in both rural and urban centers. $PM_{2.5}$ levels are increasing slightly, but trends are limited in statistical significance. A caveat to this analysis is that the training and validation data locations at U.S. embassies and consulates are largely located in major cities and not rural areas. Future work could potentially improve model performance especially in rural areas by leveraging well-calibrated low-cost air sensors, which have proliferated across West Africa. Additionally, higher resolution satellite estimates from geostationary satellites are also desperately needed across the African continent to support higher time resolution $PM_{2.5}$ concentration data. This type of data set has several useful applications including policy evaluation studies, environmental justice studies, and perhaps most critically as exposure data to examine associations with health outcomes such as cardiovascular mortality, maternal-child health, and all-cause mortality.

https://doi.org/10.1021/acsestair.4c00366
ACS EST Air XXXX, XXX, XXX−XXX

Such health cohort data is available in certain areas of West Africa,[57] and more studies are underway, but obtaining accurate exposure data, especially throughout the entire duration of the cohort, remains challenging. This method provides a potential solution to the missing exposure data and makes a substantial step forward in closing the air pollution data gap in an extremely data-scare region of the world.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestair.4c00366.

Map of domain and cities with ground-based data (Figure S1); feature importance of explanatory variables in the machine learning models (Figure S2); comparison of annual mean predicted PM2.5 in this study to the product of the Atmospheric Composition Analysis Group at Washington University (Figure S3); city-level spatial distributions of PM2.5 change in West Africa (Figure S4) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Daniel M. Westervelt − *Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, United States; Columbia University, Department of Earth and Environmental Sciences, New York, New York 10025, United States;* orcid.org/0000-0003-0806-9961; Email: danielmw@ldeo.columbia.edu

Authors

Joe Adabouk Amooli − *Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, United States; Columbia University, Department of Earth and Environmental Sciences, New York, New York 10025, United States*

Abhishek Anand − *Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, United States;* orcid.org/0000-0001-8263-192X

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestair.4c00366

Author Contributions

D.M.W.: Writing—original draft, data curation, formal analysis, methodology, software, validation, visualization, supervision, funding acquisition, conceptualization, project administration, resources, writing—review and editing. J.A.A.: Writing—review and editing, methodology, software, visualization, validation. A.A.: Writing—review and editing, conceptualization, project administration

Funding

The authors acknowledge support from National Science Foundation Grant OISE 2020677 and Clean Air Fund Grant 001519.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors declare no conflicts of interest. US Embassy data used in this study is no longer publicly available due to government funding cuts, but an archived version of the data is available at reasonable request to the authors. NASA and ESA data is publicly available on several databases, including Google Earth Engine, NASA Worldview, and ESA Earth online.

■ REFERENCES

(1) Southerland, V. A.; Brauer, M.; Mohegh, A.; Hammer, M. S.; van Donkelaar, A.; Martin, R. V.; Apte, J. S.; Anenberg, S. C. Global Urban Temporal Trends in Fine Particulate Matter (PM2.5) and Attributable Health Burdens: Estimates from Global Datasets. *Lancet Planet. Health* **2022**, *6* (2), e139−e146.

(2) Health Effects Institute. *State of Global Air Report 2024* Health Effects Institute: Boston, MA; 2024.

(3) Liu, J.; Clark, L. P.; Bechle, M. J.; Hajat, A.; Kim, S. Y.; Robinson, A. L.; Sheppard, L.; Szpiro, A. A.; Marshall, J. D. Disparities in Air Pollution Exposure in the United States by Race/Ethnicity and Income, 1990−2010. *Environ. Health Perspect.* **2021**, *129* (12), No. 127005, DOI: 10.1289/EHP8584.

(4) Westervelt, D. M.; Horowitz, L. W.; Naik, V.; Golaz, J. C.; Mauzerall, D. L. Radiative Forcing and Climate Response to Projected 21st Century Aerosol Decreases. *Atmos. Chem. Phys.* **2015**, *15*, 12681−12703.

(5) Szopa, S.; Naik, V.; Adhikary, B.; Artaxo, P.; Berntsen, T.; Collins, W. D.; Aas, W.; Akritidis, D.; Allen, R. J.; Kanaya, Y.; Prather, M. J.; Kuo, C.; Zhai, P.; Pirani, A.; Connors, S.; Péan, C.; Berger, S.; Caud, N.; Chen, Y.; Goldfarb, L.; Gomis, M.; Huang, M.; Leitzell, K.; Lonnoy, E.; Matthews, J.; Maycock, T.; Waterfield, T.; Yelekçi, O.; Yu, R.; Zhou, B. Short-Lived Climate Forcers. In *Climate Change 2021 − The Physical Science Basis*; Cambridge University Press pp 817−922.

(6) Fisher, S.; Bellinger, D. C.; Cropper, M. L.; Kumar, P.; Binagwaho, A.; Koudenoukpo, J. B.; Park, Y.; Taghian, G.; Landrigan, P. J. Air Pollution and Development in Africa: Impacts on Health, the Economy, and Human Capital. *Lancet Planet. Health* **2021**, *5* (10), e681−e688.

(7) 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. *Atmos. Environ. X* **2019**, *3*, No. 100040.

(8) Giordano, M. R.; Malings, C.; Pandis, S. N.; Presto, A. A.; McNeill, V. F.; Westervelt, D. M.; Beekmann, M.; Subramanian, R. 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**, *158*, No. 105833.

(9) Okure, D.; Ssematimba, J.; Sserunjogi, R.; Gracia, N. L.; Soppelsa, M. E.; Bainomugisha, E. Characterization of Ambient Air Quality in Selected Urban Areas in Uganda Using Low-Cost Sensing and Measurement Technologies. *Environ. Sci. Technol.* **2022**, *56* (6), 3324−3339.

(10) Raheja, G.; Sabi, K.; Sonla, H.; Gbedjangni, E. K.; McFarlane, C. M.; Hodoli, C. G.; Westervelt, D. M. A Network of Field-Calibrated Low-Cost Sensor Measurements of PM2.5in Lomé, Togo, Over One to Two Years. *ACS Earth Space. Chem.* **2022**, *6* (4), 1011−1021.

(11) McFarlane, C.; Raheja, G.; Malings, C.; Appoh, E. K. E.; Hughes, A. F.; Westervelt, D. M. Application of Gaussian Mixture Regression for the Correction of Low Cost PM2.5Monitoring Data in Accra, Ghana. *ACS Earth Space Chem.* **2021**, *5* (9), 2268−2279.

(12) Subramanian, R.; Garlan, R. M.The Powerful Potential of Low-Cost Sensors for Air Quality Research in Africa *Clean Air J.* 2021; Vol. *31* 1.

(13) 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* (7), No. 200619.

(14) 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, DOI: 10.1016/j.envpol.2018.06.044.

(15) Njeru, M. N.; Mwangi, E.; Gatari, M. J.; Kaniu, M. I.; Kanyeria, J.; Raheja, G.; Westervelt, D. M. First Results From a Calibrated Network of Low-Cost PM2.5 Monitors in Mombasa, Kenya Show Exceedance of Healthy Guidelines. *GeoHealth* **2024**, *8* (9), No. e2024GH001049.
(16) Raheja, G.; Nimo, J.; Appoh, E. K.-E.; Essien, B.; Sunu, M.; Nyante, J.; Amegah, M.; Quansah, R.; Arku, R. E.; Penn, S. L.; Giordano, M. R.; Zheng, Z.; Jack, D.; Chillrud, S.; Amegah, K.; Subramanian, R.; Pinder, R.; Appah-Sampong, E.; Tetteh, E. N.; Borketey, M. A.; Hughes, A. F.; Westervelt, D. M. Low-Cost Sensor Performance Intercomparison, Correction Factor Development, and 2+ Years of Ambient PM2.5 Monitoring in Accra, Ghana. *Environ. Sci. Technol.* **2023**, *57*, 10708−10720.
(17) Nana, B.; Raheja, G.; Ouarma, I.; Kayaba, H.; Gounkaou, W. Y.; Daho, T.; Béré, A.; Mellouki, A.; Westervelt, D. M. Monitoring of PM2.5 Using Well-Calibrated Low-Cost Sensors over One Year in Burkina Faso. *ACS ES&T Air* **2025**, *2*, 40−48.
(18) Rueda, E. M.; Carter, E.; L'Orange, C.; Quinn, C.; Volckens, J. Size-Resolved Field Performance of Low-Cost Sensors for Particulate Matter Air Pollution. *Environ. Sci. Technol. Lett.* **2023**, *10*, 247−253.
(19) Hagan, D. H.; Kroll, J. H. Assessing the Accuracy of Low-Cost Optical Particle Sensors Using a Physics-Based Approach. *Atmos. Meas. Technol.* **2020**, *13* (11), 6343−6355.
(20) 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*, No. 117067.
(21) 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. Technol.* **2018**, *11* (8), 4883−4890.
(22) Ouimette, J. R.; Malm, W. C.; Schichtel, B. A.; Sheridan, P. J.; Andrews, E.; Ogren, J. A.; Arnott, W. P. Evaluating the PurpleAir Monitor as an Aerosol Light Scattering Instrument. *Atmos. Meas. Technol.* **2022**, *15* (3), 655−676.
(23) van Donkelaar, A.; Martin, R. V.; Park, R. J. Estimating Ground-Level PM $_{2.5}$ Using Aerosol Optical Depth Determined from Satellite Remote Sensing. *J. Geophys. Res.* **2006**, *111* (D21), No. D21201.
(24) Liu, Y.; Sarnat, J. A.; Kilaru, V.; Jacob, D. J.; Koutrakis, P. Estimating Ground-Level PM2.5 in the Eastern United States Using Satellite Remote Sensing. *Environ. Sci. Technol.* **2005**, *39* (9), 3269−3278.
(25) Xiao, Q.; Wang, Y.; Chang, H. H.; Meng, X.; Geng, G.; Lyapustin, A.; Liu, Y. Full-Coverage High-Resolution Daily PM2.5 Estimation Using MAIAC AOD in the Yangtze River Delta of China. *Remote Sens. Environ.* **2017**, *199*, 437−446.
(26) Van Donkelaar, A.; Martin, R. V.; Brauer, M.; Hsu, N. C.; Kahn, R. A.; Levy, R. C.; Lyapustin, A.; Sayer, A. M.; Winker, D. M. Global Estimates of Fine Particulate Matter Using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. *Environ. Sci. Technol.* **2016**, *50* (7), 3762−3772.
(27) Wei, J.; Li, Z.; Lyapustin, A.; Wang, J.; Dubovik, O.; Schwartz, J.; Sun, L.; Li, C.; Liu, S.; Zhu, T. First Close Insight into Global Daily Gapless 1 Km PM2.5 Pollution, Variability, and Health Impact. *Nat. Commun.* **2023**, *14* (1), No. 8349.
(28) Cohen, A. J.; Brauer, M.; Burnett, R.; Anderson, R.; Frostad, J.; Estep, K.; Balakrishnan, K.; Brunekreef, B.; Dandona, L.; Dandona, R.; Feigin, V.; Freedman, G.; Hubbell, B.; Jobling, A.; Kan, H.; Knibbs, L.; Liu, Y.; Martin, R.; Morawska, L.; Arden, C.; Iii, P.; Shin, H.; Straif, K.; Shaddick, G.; Thomas, M.; Van Dingenen, R.; Van Donkelaar, A.; Vos, T.; Murray, C. J. L.; Forouzanfar, M. H. Estimates and 25-Year Trends of the Global Burden of Disease Attributable to Ambient Air Pollution: An Analysis of Data from the Global Burden of Diseases Study 2015. *Lancet* **2017**, *389*, 1907−1918.
(29) Ma, Z.; Dey, S.; Christopher, S.; Liu, R.; Bi, J.; Balyan, P.; Liu, Y. A Review of Statistical Methods Used for Developing Large-Scale and Long-Term PM2.5 Models from Satellite Data. *Remote Sens. Environ.* **2022**, *269*, No. 112827.
(30) Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M. B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Mickley, L. J.; Schwartz, J. An Ensemble-Based Model of PM2.5 Concentration across the Contiguous United States with High Spatiotemporal Resolution. *Environ. Int.* **2019**, *130*, No. 104909.
(31) Meng, X.; Liu, C.; Zhang, L.; Wang, W.; Stowell, J.; Kan, H.; Liu, Y. Estimating PM2.5 Concentrations in Northeastern China with Full Spatiotemporal Coverage, 2005−2016. *Remote Sens. Environ.* **2021**, *253*, No. 112203.
(32) Katoch, V.; Kumar, A.; Imam, F.; Sarkar, D.; Knibbs, L. D.; Liu, Y.; Ganguly, D.; Dey, S. Addressing Biases in Ambient PM2.5 Exposure and Associated Health Burden Estimates by Filling Satellite AOD Retrieval Gaps over India. *Environ. Sci. Technol.* **2023**, *57* (48), 19190−19201.
(33) Handschuh, J.; Erbertseder, T.; Baier, F. Systematic Evaluation of Four Satellite AOD Datasets for Estimating PM2.5 Using a Random Forest Approach. *Remote Sens.* **2023**, *15* (8), No. 2064.
(34) Zhu, H.; Martin, R. V.; Van Donkelaar, A.; Hammer, M. S.; Li, C.; Meng, J.; Oxford, C. R.; Liu, X.; Li, Y.; Zhang, D.; Singh, I.; Lyapustin, A. Importance of Aerosol Composition and Aerosol Vertical Profiles in Global Spatial Variation in the Relationship between PM2.5 and Aerosol Optical Depth. *Atmos. Chem. Phys.* **2024**, *24* (20), 11565−11584.
(35) Kang, J.-G.; Lee, J.-Y.; Lee, J.-B.; Lim, J.-H.; Yun, H.-Y.; Choi, D.-R.; Kang, J.-G.; Lee, J.-Y.; Lee, J.-B.; Lim, J.-H.; Yun, H.-Y.; Choi, D.-R. High-Resolution Daily PM2.5 Exposure Concentrations in South Korea Using CMAQ Data Assimilation with Surface Measurements and MAIAC AOD (2015−2021). *Atmosphere* **2024**, *15* (10), No. 1152.
(36) Malings, C.; Westervelt, D. M.; Hauryliuk, A.; Presto, A. A.; Grieshop, A.; Bittner, A.; Beekmann, M. Application of Low-Cost Fine Particulate Mass Monitors to Convert Satellite Aerosol Optical Depth to Surface Concentrations in North America and Africa. *Atmos. Meas. Technol.* **2020**, *13* (7), 3873−3892.
(37) Zhang, D.; Du, L.; Wang, W.; Zhu, Q.; Bi, J.; Scovronick, N.; Naidoo, M.; Garland, R. M.; Liu, Y. A Machine Learning Model to Estimate Ambient PM2.5 Concentrations in Industrialized Highveld Region of South Africa. *Remote Sens. Environ.* **2021**, *266*, No. 112713.
(38) Léon, J.-F.; Akpo, A. B.; Bedou, M.; Djossou, J.; Bodjrenou, M.; Yoboué, V.; Liousse, C. PM2.5surface Concentrations in Southern West African Urban Areas Based on Sun Photometer and Satellite Observations. *Atmos. Chem. Phys.* **2021**, *21* (3), 1815−1834.
(39) Amooli, J. A.; Hackman, K. O.; Nana, B.; Westervelt, D. M. Fine Particulate Air Pollution Estimation in Ouagadougou Using Satellite Aerosol Optical Depth and Meteorological Parameters. *Environ. Sci. Atmos.* **2024**, *4* (9), 1012−1025.
(40) United Nations Population Fund (UNFPA). World Population Dashboard United Nations. https://www.unfpa.org/data/world-population-dashboard. (accessed December 29, 2024).
(41) Zheng, Z.; Fiore, A. M.; Westervelt, D. M.; Milly, G. P.; Goldsmith, J.; Karambelas, A.; Curci, G.; Randles, C. A.; Paiva, A. R.; Wang, C.; Wu, Q.; Dey, S. Automated Machine Learning to Evaluate the Information Content of Tropospheric Trace Gas Columns for Fine Particle Estimates Over India: A Modeling Testbed. *J. Adv. Model. Earth Syst.* **2023**, *15* (3), No. e2022MS003099.
(42) Lyapustin, A.; Wang, Y.; Korkin, S.; Huang, D. MODIS Collection 6 MAIAC Algorithm. *Atmos. Meas. Technol.* **2018**, *11* (10), 5741−5765.
(43) Veefkind, J. P.; Aben, I.; McMullan, K.; Förster, H.; de Vries, J.; Otter, G.; Claas, J.; Eskes, H. J.; de Haan, J. F.; Kleipool, Q.; van Weele, M.; Hasekamp, O.; Hoogeveen, R.; Landgraf, J.; Snel, R.; Tol, P.; Ingmann, P.; Voors, R.; Kruizinga, B.; Vink, R.; Visser, H.; Levelt, P. F. TROPOMI on the ESA Sentinel-5 Precursor: A GMES Mission for Global Observations of the Atmospheric Composition for Climate, Air Quality and Ozone Layer Applications. *Remote Sens. Environ.* **2012**, *120*, 70−83.

https://doi.org/10.1021/acsestair.4c00366
ACS EST Air XXXX, XXX, XXX−XXX

(44) Levelt, P. F.; Van Den Oord, G. H. J.; Dobber, M. R.; Mälkki, A.; Visser, H.; De Vries, J.; Stammes, P.; Lundell, J. O. V.; Saari, H. The Ozone Monitoring Instrument. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44* (5), 1093−1101, DOI: 10.1109/TGRS.2006.872333.
(45) Muñoz-Sabater, J.; Dutra, E.; Agustí-Panareda, A.; Albergel, C.; Arduini, G.; Balsamo, G.; Boussetta, S.; Choulga, M.; Harrigan, S.; Hersbach, H.; Martens, B.; Miralles, D. G.; Piles, M.; Rodríguez-Fernández, N. J.; Zsoter, E.; Buontempo, C.; Thépaut, J. N. ERA5-Land: A State-of-the-Art Global Reanalysis Dataset for Land Applications. *Earth Syst. Sci. Data* **2021**, *13* (9), 4349−4383.
(46) Huang, H.; Cui, C.; Cheng, L.; Liu, Q.; Wang, J. Grid Interpolation Algorithm Based on Nearest Neighbor Fast Search. *Earth Sci. Inf.* **2012**, *5* (3−4), 181−187.
(47) Hagler, G.; Hanley, T.; Hassett-Sipple, B.; Vanderpool, R.; Smith, M.; Wilbur, J.; Wilbur, T.; Oliver, T.; Shand, D.; Vidacek, V.; Johnson, C.; Allen, R.; D'Angelo, C. Evaluation of Two Collocated Federal Equivalent Method PM2.5 Instruments over a Wide Range of Concentrations in Sarajevo, Bosnia and Herzegovina. *Atmos. Pollut. Res.* **2022**, *13* (4), No. 101374.
(48) Lundberg; Scott, M.; Lee, Su.-In.*A Unified Approach to Interpreting Model Predictions*; Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17); NIPS, 2017; pp 4768−4777.
(49) Jobson, J. D. Multiple Linear Regression. In *Applied Multivariate Data Analysis: Springer Texts in Statistics*; Springer: New York, NY, 1991; pp 219−398.
(50) Altınçay, H. Decision Trees Using Model Ensemble-Based Nodes. *Pattern Recognit.* **2007**, *40* (12), 3540−3551.
(51) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(52) Chen, T.; Guestrin, C.et al.*XGBoost: A Scalable Tree Boosting System*; Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM, 2016; pp 785−794.
(53) Tai, A. P. K.; Mickley, L. J.; Jacob, D. J. Correlations between Fine Particulate Matter (PM2.5) and Meteorological Variables in the United States: Implications for the Sensitivity of PM2.5 to Climate Change. *Atmos. Environ.* **2010**, *44* (32), 3976−3984.
(54) 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.
(55) Fu, T. M.; Jacob, D. J.; Heald, C. L. Aqueous-Phase Reactive Uptake of Dicarbonyls as a Source of Organic Aerosol over Eastern North America. *Atmos. Environ.* **2009**, *43* (10), 1814−1822.
(56) Shen, S.; Li, C.; van Donkelaar, A.; Jacobs, N.; Wang, C.; Martin, R. V. Enhancing Global Estimation of Fine Particulate Matter Concentrations by Including Geophysical a Priori Information in Deep Learning. *ACS EST Air* **2024**, *1* (5), 332−345.
(57) Jack, D. W.; Asante, K. P.; Wylie, B. J.; Chillrud, S. N.; Whyatt, R. M.; Ae-Ngibise, K. A.; Quinn, A. K.; Yawson, A. K.; Boamah, E. A.; Agyei, O.; Mujtaba, M.; Kaali, S.; Kinney, P.; Owusu-Agyei, S. Ghana Randomized Air Pollution and Health Study (GRAPHS): Study Protocol for a Randomized Controlled Trial. *Trials* **2015**, *16* (1), No. 420.

https://doi.org/10.1021/acsestair.4c00366
ACS EST Air XXXX, XXX, XXX−XXX