Modeling $PM_{2.5}$ in Africa Using GEOS-Chem, Satellite Data, and Machine Learning

Benjamin Yang[1,2], Daniel Westervelt[2], Zhonghua Zheng[3], Garima Raheja[1,2], Allison Hughes[4], Emmanuel K-E Appoh[5], Victoria Owusu-Tawiah[6], and others[7]

[1]Department of Earth and Environmental Sciences, Columbia University, [2]Lamont-Doherty Earth Observatory, [3]Department of Earth and Environmental Sciences, The University of Manchester, [4]Department of Physics, University of Ghana, [5]Ghana Environmental Protection Agency, [6]Meteorology and Climate Science, Kwame Nkrumah University of Science and Technology, and [7]CAMS-Net and AfriqAir teams

1. Introduction

- Exposure to ambient **fine particulate matter ($PM_{2.5}$)** is a leading environmental risk factor for premature death.
- In Africa, surface $PM_{2.5}$ data is **sparse**, hindering pollution mitigation plans and human health improvement.
- Emerging hybrid surface air quality observation networks of **reference monitors** and **well-calibrated low cost sensors** provide an opportunity to evaluate and improve models over Africa.
- **Satellite data** from agencies such as the National Aeronautics and Space Administration (NASA) provide near-complete spatial coverage, but their columnar nature is imperfect representations of surface pollution.
- **Objectives:**
  1. Compare $PM_{2.5}$ predicted using the GEOS-Chem and machine learning (XGBoost) models.
  2. Evaluate the models against surface $PM_{2.5}$ monitoring sites across sub-Saharan Africa.

2. GEOS-Chem Model

- **GEOS-Chem** is a global 3D chemical transport model used freely to investigate atmospheric chemistry.
- Set up GCClassic v13.3.3 with a custom nested grid at 0.25° x 0.3125° horizontal resolution, 72 vertical layers, and GEOS-FP meteorology.
- Used 2013 DICE-Africa emissions inventory from Marais and Wiedenmeyer (2016). Scaling emissions to 2021 had little impact on modeled concentrations.
- Ran simulations (October 2020 - December 2021) on Columbia's HPC "Ginsburg" on a 32-core node.

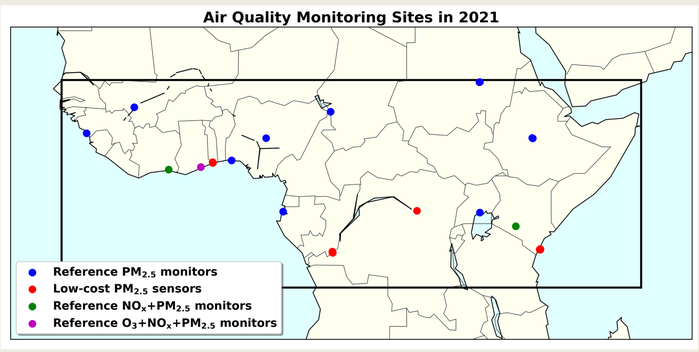


Figure 1. GEOS-Chem model domain (box) including 14 reference-grade (blue, green, purple) and 13 low-cost $PM_{2.5}$ monitoring sites (red). Note that some cities have multiple sites (points overlap).

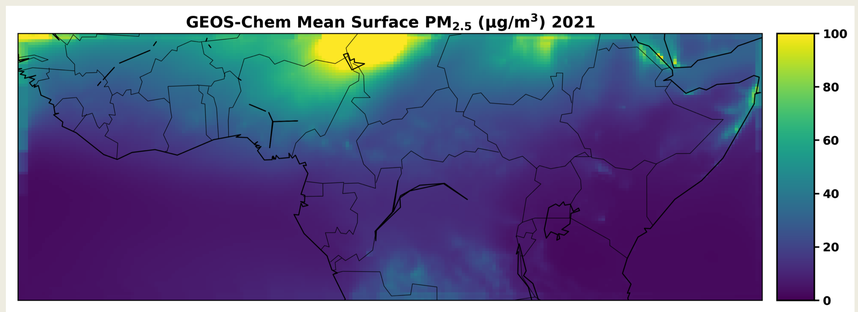


Figure 2. Annual mean $PM_{2.5}$ over the domain. The northern part is particularly impacted by Harmattan (dust season). There are faint pollution hot spots representing major cities.

3. Machine Learning Model

- **Extreme Gradient Boosting (XGBoost)** is a relatively novel, highly efficient, and sparsity-aware machine learning algorithm for parallel decision tree learning (Chen and Guestrin, 2016). This algorithm uses regularization and tree pruning to avoid overfitting and improve generalization.
- We found XGBoost to be the best algorithm, outperforming random forest and multiple linear regression.
- The model was trained at **six cities (monitoring sites)**: Abidjan, Abuja, Accra, Bamako, Conakry, Lagos.
- Used **13 input features** to predict daily $PM_{2.5}$:
  - ✓ **ECMWF reanalysis v5 (ERA5):** surface pressure (*sp*), 2-m temperature (*t2m*), total precipitation (*tp*), 10-m u-component of wind (*u10*), 10-m v-component of wind (*v10*)
  - ✓ **TROPMOMI satellite retrievals of tropospheric trace gas columns and aerosol properties:** carbon monoxide (*co*), formaldehyde (*hcho*), nitrogen dioxide (*no2*), ozone (*o3*), sulfur dioxide (*so2*), absorbing aerosol index (*aai*), aerosol height (*ah*)
  - ✓ **MODIS MAIAC satellite retrievals:** aerosol optical depth (*aod*)

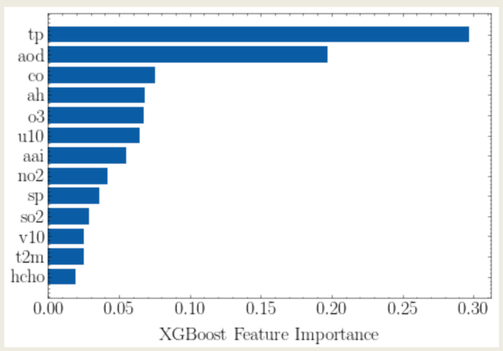


Figure 3. Built-in XGBoost feature importance scores. For this given model run, total precipitation and aerosol optical depth were the most important features. This suggests that precipitation is significant for removing pollutants from the atmosphere and confirms that AOD is a reasonable proxy for $PM_{2.5}$.

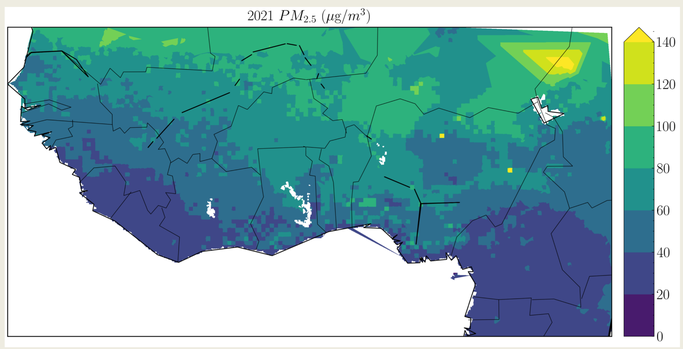
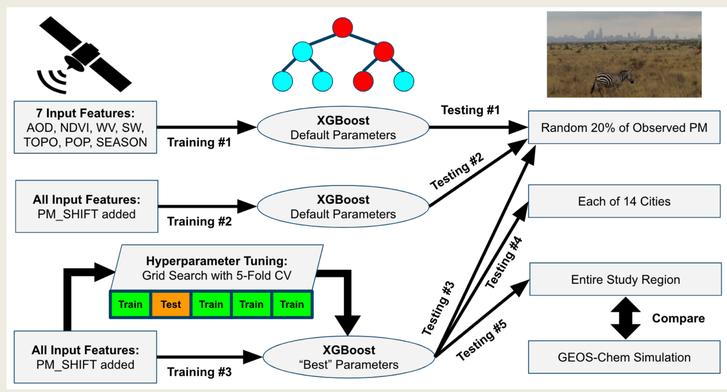
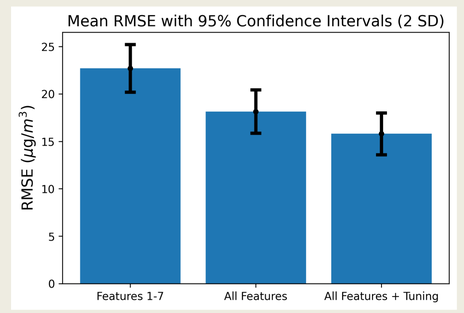


Figure 4. Annual mean $PM_{2.5}$ over West Africa based on meteorological and AOD features only. Again, we see higher values to the north (drier/dusty) and lower values to the south (wetter/vegetated).

4. Machine Learning vs. GEOS-Chem

- Designated a GEOS-Chem 2° x 2.5° global simulation as the **testbed** for asserted $PM_{2.5}$ observations.
- Trained a new XGBoost model using a total of **eight input features** at each of **14 cities**:
  - ✓ **NASA satellite earth observations:** aerosol optical depth (*AOD*), normalized difference vegetation index (*NDVI*), water vapor (*WV*), reflected shortwave radiation (*SW*), elevation (*TOPO*), population density (*POP*)
  - ✓ **Engineered features:** season indicator (*SEASON*), one-month right-shifted $PM_{2.5}$ (*PM_SHIFT*)
  - ✓ **Cities:** Abidjan, Abuja, Accra, Addis Ababa, Bamako, Conakry, Kampala, Khartoum, Kigali, Kinshasa, Lagos, Libreville, Nairobi, N'Djamena





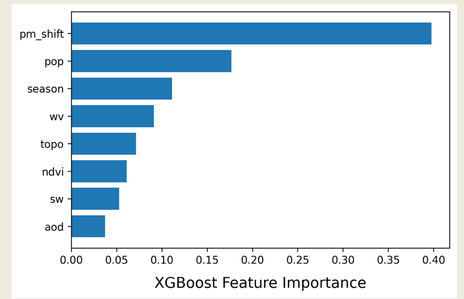


Figure 5. Flowchart on left illustrates XGBoost model training and testing steps. Top right plot shows the model tested on 20% of data (80% for training) for three experiments. Here, confidence intervals are defined as two standard deviations from 100 model runs. Bottom right plot displays feature importance from the third experiment. The model performs better when *PM_SHIFT* is introduced, and this feature ranks as most important.

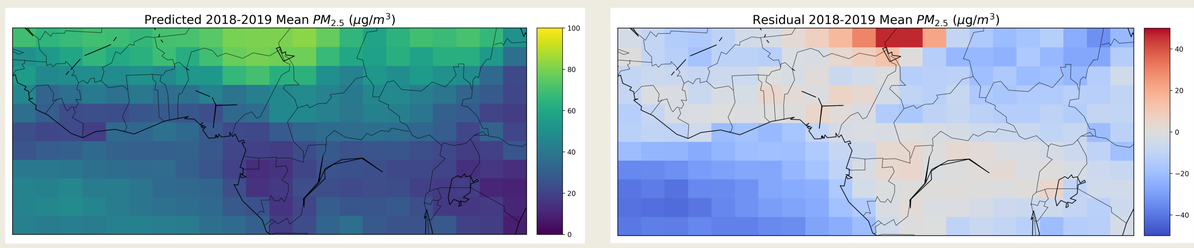


Figure 6. Annual mean predicted (left) and residual (right) $PM_{2.5}$ over the domain. Residual is "observed" (GEOS-Chem) minus predicted (XGBoost). The XGBoost model underestimates $PM_{2.5}$ by over 50 µg m^{-3} in the semi-arid Sahel region of Niger and Chad, influenced more by Saharan dust. On the other hand, $PM_{2.5}$ is overestimated over the relatively clean, remote Atlantic Ocean. The model performs better in the vincinity of cities where it was trained (e.g. a residual of -0.003 µg m^{-3} in northeast Ghana).

5. Evaluating GEOS-Chem and Machine Learning Against Obs

- GEOS-Chem mostly **underestimates** observed surface $PM_{2.5}$. The modeled diurnal profiles are fairly flat. Model limitations may include coarse resolution and inadequate emissions.
- Overall, XGBoost has a **higher r^2** (coefficient of determination) and **lower MAE** (mean absolute error) when compared to observations. However, both models miss some peaks (features of hyperlocal emissions sources).

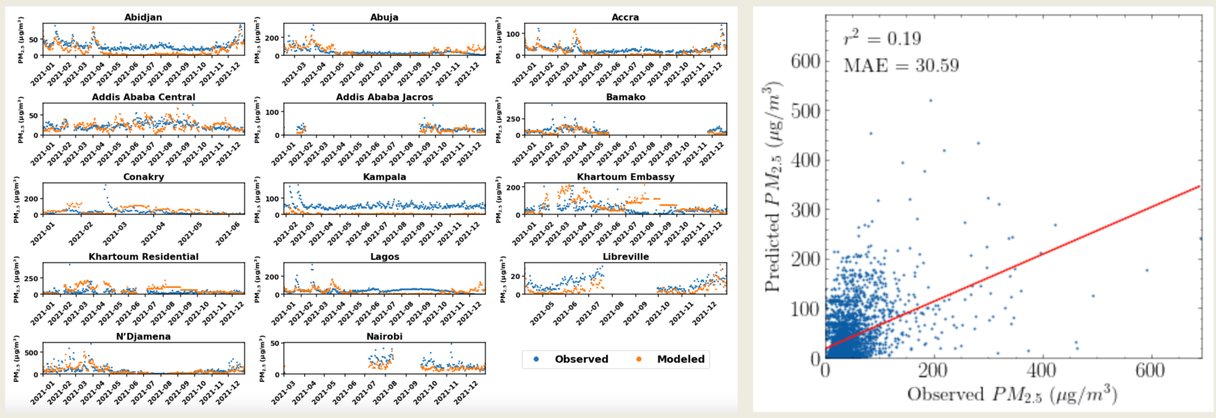


Figure 7. Left plots show daily mean surface $PM_{2.5}$ from GEOS-Chem (orange) and reference monitors (blue). Right plot is the model performance for daily mean $PM_{2.5}$ (points) at all 14 sites.

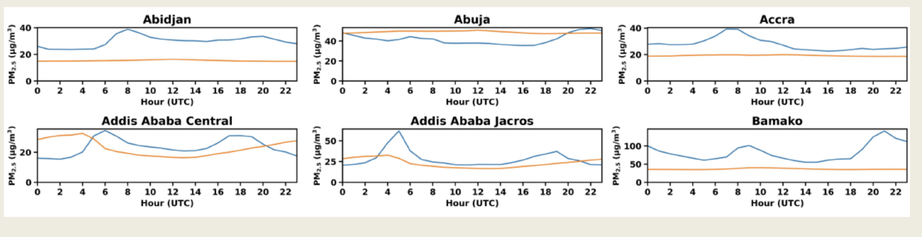


Figure 8. Diurnal profiles of $PM_{2.5}$ from GEOS-Chem (orange) and reference monitors (blue) at select cities.

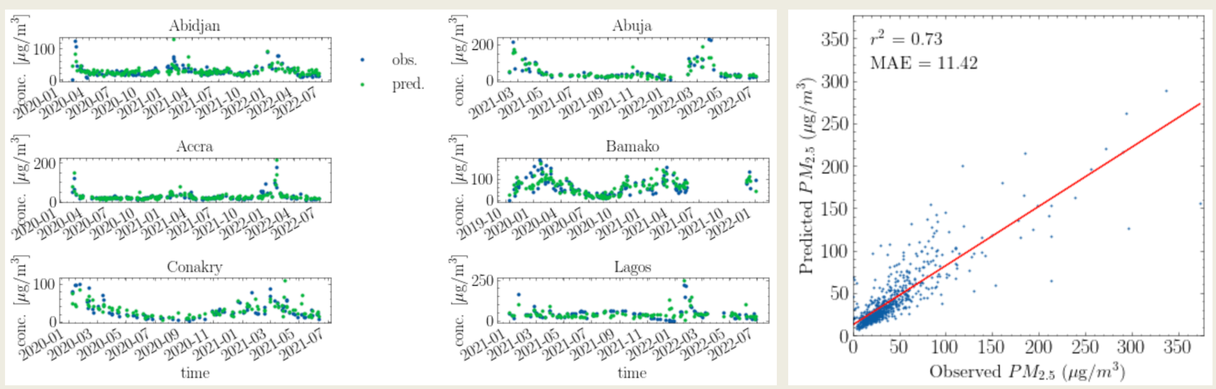


Figure 9. Left plots show XGBoost predicted $PM_{2.5}$ (green) compared to reference monitor observed $PM_{2.5}$ (blue). Right plot is model performance on test set (20% of entire dataset) for daily mean $PM_{2.5}$ (points) at all six cities.

6. Conclusions

- Our machine learning model performed better than GEOS-Chem **temporally** at available $PM_{2.5}$ monitoring sites but does not yet fully capture $PM_{2.5}$ **spatially** across sub-Saharan Africa.
- Machine learning is an **attractive alternative or supplement** to traditional chemical transport models, particularly if motivated by faster results, lower costs, and general ease of implementation.
- As **air quality monitoring networks expand** across Africa, providing more balanced data coverage, the machine learning spatial predictions are expected to improve.
- In the future, we aim to **create a new dataset** of satellite-derived, ground-truthed 1 km^2 daily surface $PM_{2.5}$ for policy evaluation in Africa. Our dataset will also enable cutting-edge research on **climate change impacts** on surface air quality levels in Africa.

Contact: benjamin.yang@columbia.edu (Westervelt Aerosol Group)

References:

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Marais, E. A., & Wiedinmyer, C. (2016). Air Quality Impact of Diffuse and Inefficient Combustion Emissions in Africa (DICE-Africa). Environmental Science & Technology, 50(19), 10739–10745. https://doi.org/10.1021/acs.est.6b02602