## Supplementary information for…

# First measurements of ambient $PM_{2.5}$ in Kinshasa, Democratic Republic of Congo and Brazzaville, Republic of Congo using field-calibrated low cost sensors

**Celeste McFarlane[1,2], Paulson Kasereka Isevulambire[3], Raymond Sinsi Lumbuenamo[4], Arnold Murphy Elouma Ndinga[5], Ranil Dhammapala[6], Xiaomeng Jin[7], V. Faye McNeill[2,8], Carl Malings[9,10], R Subramanian[9,11,12], Daniel M. Westervelt[1,13,1]**

[1]*Lamont-Doherty Earth Observatory of Columbia University, New York, USA*
[2]*Columbia University, Department of Chemical Engineering, New York, USA*
[3]*Ecole Régionale postuniversitaire d'Aménagement et de Gestion Intégrés des Forêts et Territoire tropicaux (ERAIFT) Kinshasa, Democratic Republic of Congo*
[4]*World Bank Group, Kinshasa, Democratic Republic of Congo*
[5]*Département de chimie, Université Marien Ngouabi, Brazzaville, Republic of Congo*
[6]*Washington State Department of Ecology, Washington, USA*
[7]*Department of Chemistry, University of California Berkeley, USA*
[8]*Columbia University, Department of Earth and Environmental Sciences, New York, USA*
[9]*OSU-EFLUVE - Observatoire Sciences de l'Univers-Enveloppes Fluides de la Ville à l'Exobiologie, Université Paris-Est-Créteil, France*
[10]*NASA Postdoctoral Program Fellow, Goddard Space Flight Center, Greenbelt, Maryland, USA*
[11]*Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, PA 15218, USA*
[12]*Kigali Collaborative Research Center, Kigali, Rwanda*
[13]*NASA Goddard Institute for Space Studies, New York, USA*

[1] *Corresponding author. Tel: 1-845-365-8194
*Email address*: danielmw@ldeo.columbia.edu

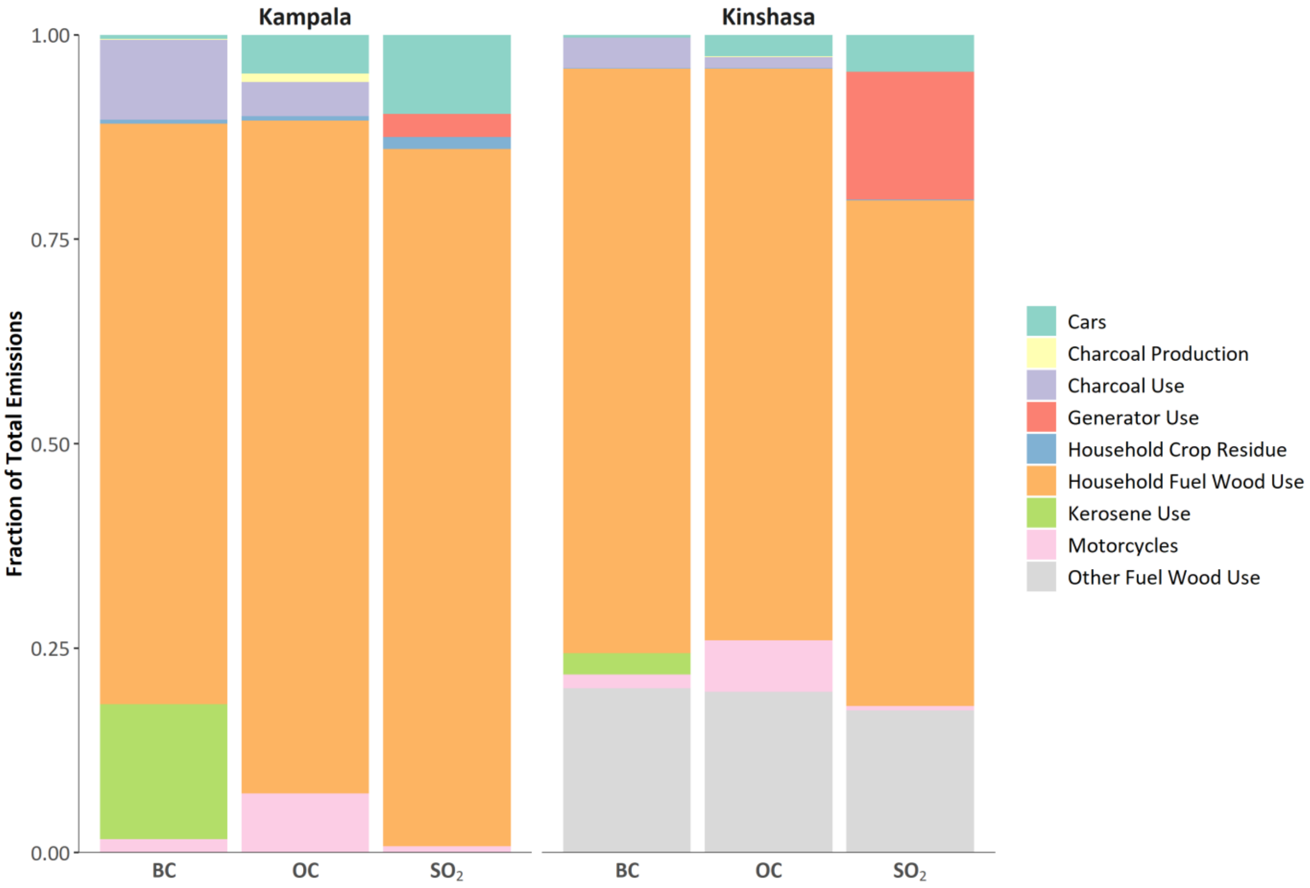

Fig. S1. Relative contribution of emission sectors to black carbon (BC), organic carbon (OC), and sulfur dioxide (SO2) in Kampala and Kinshasa from the DICE-Africa inventory for 2013.

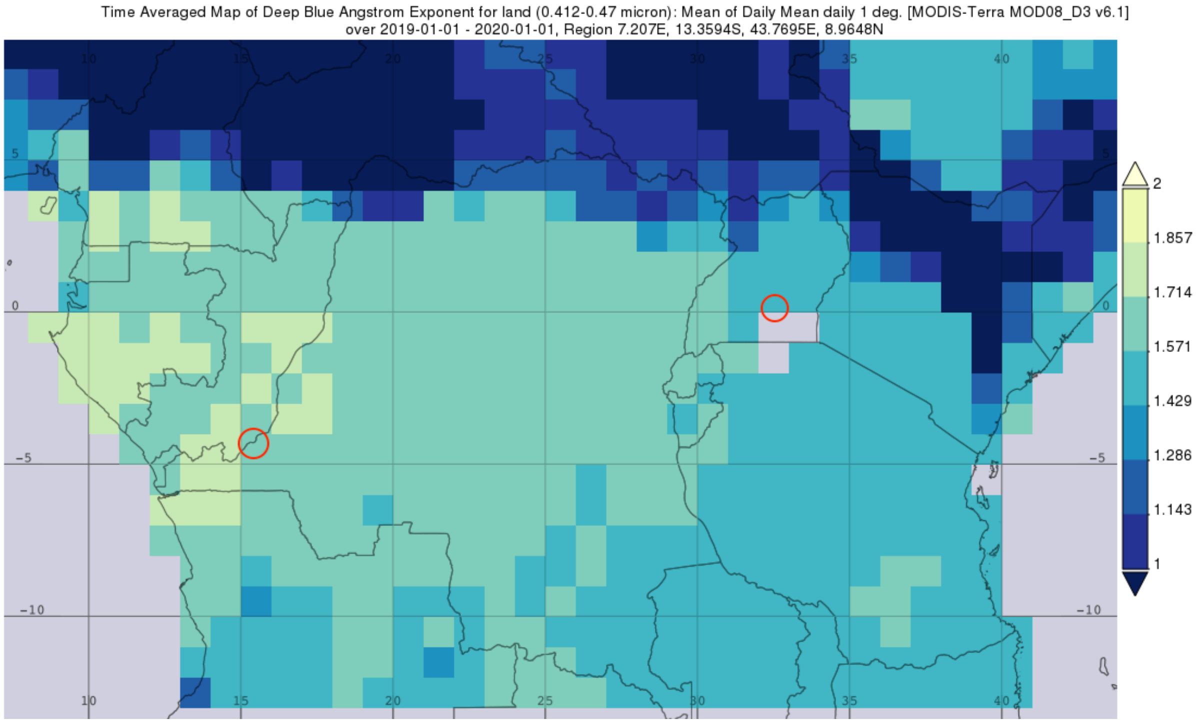
Time Averaged Map of Deep Blue Angstrom Exponent for land (0.412-0.47 micron): Mean of Daily Mean daily 1 deg. [MODIS-Terra MOD08_D3 v6.1]
over 2019-01-01 - 2020-01-01, Region 7.207E, 13.3594S, 43.7695E, 8.9648N
10
15
20
25
30
35
40
5
0
−5
−10
2
1.857
1.714
1.571
1.429
1.286
1.143
1

Figure S2. 2019 Annual mean MODIS Terra Deep Blue Angstrom Exponent over Central Africa. Red circles indicate approximate locations of Kinshasa/Brazzaville and Kampala.

***Supplemental methods:***

*Random Forest*

We also used a Random Forest (RF) approach to develop a calibration model for both the raw daily-averaged and hourly-averaged PurpleAir $PM_{2.5}$ data (Si et al., 2020.; Zimmerman et al., 2018). The advantage of the random forest approach is that it can better capture any non-linear relationship between the explanatory variables and the target species than MLR. The model consists of a user determined number of decision trees, each constructed with a random bootstrapped sample from the training data set. Each tree begins with an origin node (a specific condition which stratifies the data) that is split into sub-nodes after evaluating the response against a random subset of explanatory variables. The number of explanatory variables evaluated at each node is called $m_{try}$ and was set to one in this model. The algorithm then splits the tree using the explanatory variable found to be the strongest predictor of the response. Sub-nodes continue to be split until a terminal node is reached. The maximum number of sub-nodes and the minimum number of data points in each node are parameters that can be adjusted by the user to balance the accuracy and efficiency of the algorithm with the risk of overfitting the data set. In this model, the maximum number of sub-nodes was not restricted and the minimum number of data points per node was set to five.

In this model 500 trees were constructed using hourly values of PurpleAir $PM_{2.5}$ concentrations, temperature and relative humidity as explanatory variables. By using a large number of trees, each constructed with a different assortment of values, the random forest algorithm reduces the risk of overfitting. The higher the value of $m_{try}$, the greater the model accuracy against the training set and the greater the risk of overfitting. This is because as $m_{try}$ increases, the number of variables tested at each node increases allowing for a more detailed evaluation. This, however, decreases the randomness of the model which is important when evaluating overfitting. We employ a 75%/25% training versus validation split as with the MLR. This model was generated using the randomForest package in R. Using the Random Forest model to correct the raw daily $PM_{2.5}$ data towards FEM standard using the BAM-1020 results in a reduction of mean absolute error to 5.8 µg $m^{-3}$ from 14.8 µg $m^{-3}$ (see Table 1).