A machine learning model to estimate ambient $PM_{2.5}$ concentrations in industrialized highveld region of South Africa

Danlu Zhang[b], Linlin Du[a], Wenhao Wang[a], Qingyang Zhu[a], Jianzhao Bi[a], Noah Scovronick[a], Mogesh Naidoo[c], Rebecca M Garland[c,d,e], Yang Liu[a]

[a] Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

[b] Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

[c] Council for Scientific and Industrial Research, Pretoria 0001, South Africa

[d] Unit for Environmental Sciences and Management, North-West University, Potchefstroom 2520, South Africa

[e] Department of Geography, Geo-informatics and Meteorology, University of Pretoria, Pretoria 0001, South Africa

Highlights

- We developed a random forest model to estimate daily $PM_{2.5}$ concentrations at 1 km^2 resolution in South Africa.
- Our model captured seasonal trends and spatial patterns of $PM_{2.5}$ with relatively high accuracy.
- High $PM_{2.5}$ levels were identified in low-income settlements and industrial areas in western Mpumalanga.
- PM2.5 levels decreased in north of Gauteng province after the implementation of new air quality standard.

Abstract

Exposure to fine particulate matter ($PM_{2.5}$) has been linked to a substantial disease burden globally, yet little has been done to estimate the population health risks of $PM_{2.5}$ in South Africa due to the lack of high-resolution $PM_{2.5}$ exposure estimates. We developed a random forest model to estimate daily $PM_{2.5}$ concentrations at 1 km^2 resolution in and around industrialized Gauteng Province, South Africa, by combining satellite aerosol optical depth (AOD), meteorology, land use, and socioeconomic data. We then compared $PM_{2.5}$ concentrations in the study domain before and after the implementation of the new national air quality standards. We aimed to test whether machine learning models are suitable for regions with sparse ground observations such as South Africa and which predictors played important roles in $PM_{2.5}$ modeling. The cross-validation R^2 and Root Mean Square Error of our model was 0.80 and 9.40 $\mu g/m^3$, respectively. Satellite AOD, seasonal indicator, total precipitation, and population were among the most important predictors. Model-estimated $PM_{2.5}$ levels successfully captured the temporal pattern recorded by ground observations. Spatially, the highest annual $PM_{2.5}$ concentration appeared in central and northern Gauteng, including northern Johannesburg and the city of Tshwane. Since the 2016 changes in national $PM_{2.5}$ standards, $PM_{2.5}$ concentrations have decreased in most of our study region, although levels in Johannesburg and its surrounding areas have remained relatively constant. This is anadvanced $PM_{2.5}$ model for South Africa with high prediction accuracy at the daily level and at a relatively high spatial resolution. Our study provided a reference for predictor selection, and our results can be used for a variety of purposes, including epidemiological research, burden of disease assessments, and policy evaluation.

Keywords: $PM_{2.5}$, MAIAC AOD, random forest, air quality standard, South Africa

1. Introduction

Fine particulate matter ($PM_{2.5}$, airborne particles with an aerodynamic diameter of less than 2.5 μm) is a ubiquitous air pollutant that harms human health and wellbeing. Numerous epidemiological studies of both short- and long-term exposure have reported strong associations with adverse health outcomes, including premature mortality and morbidity from a range of diseases (Atkinson et al., 2014; Burnett et al., 2018; Liu et al., 2019). Populations living in developing regions, in particular, are often exposed to levels of particulate matter greatly exceeding WHO standards, and an estimated 4.2 million premature death worldwide are attributable to ambient $PM_{2.5}$ (Cohen et al., 2017; World Health Organization, 2016).

In addition to natural sources such as biomass burning, dust storms, and ocean spray, $PM_{2.5}$ and its precursors are emitted from several anthropogenic sources, including industrial activities, power generation, vehicle traffic, agriculture burning, and household fuel use (Tucker, 2000). The diverse emission sources and secondary production that occurs in the atmosphere results in complex distributions of $PM_{2.5}$ in space and time (Seinfeld and Pandis, 2016). Current air quality monitoring networks are often insufficient to quantify $PM_{2.5}$ exposure and health risk at the local level even in high-income countries. Routine monitoring is even more sparse or nonexistent in many low- and middle-income countries (Brauer et al., 2016), however satellite data can be used to fill this gap.

The past decade has seen the increasing application of satellite remote sensing products such as aerosol optical depth (AOD) to estimate surface $PM_{2.5}$ concentrations. AOD measures the light extinction of aerosol particles at a given wavelength as it passes through the atmospheric column. Although AOD is often strongly correlated with surface $PM_{2.5}$ concentration, this relationship is nonlinear and modified by various factors such as meteorology, particle vertical

distribution, and particle chemical composition (Hoff and Christopher, 2009; Liu et al., 2005; Sorek-Hamer et al., 2020). Over the past two decades, a number of statistical models have been proposed to capture these relationships at different spatial and temporal scales in order to improve prediction accuracy and robustness, including linear mixed-effects models (Ma et al., 2016b), geographically weighted regression (GWR) (Ma et al., 2014), generalized additive models (Strawa et al., 2013), Bayesian downscaler (Chang et al., 2013), and multi-stage models (Kloog et al., 2014). Most recently, machine learning models such as random forests (Brokamp et al., 2018; Hu et al., 2017) and neural network (Li et al., 2020) have shown high prediction accuracy. These advanced satellite-driven models are useful tools to fill the data gaps left by sparse ground monitors networks and enable more comprehensive assessments of $PM_{2.5}$ exposure and its associated health effects. Random forest model is a good choice for areas where advanced models have never been built to explore the spatiotemporal patterns of $PM_{2.5}$ because it not only has high prediction accuracy but also provides guidance for predictor selection which is very helpful for future research.

As part of the 2004 National Environmental Management: Air Quality Act (Act No. 39 of 2004), approximately 130 ground ambient air quality stations have been established or, for those already in existence, incorporated into the national reporting of air quality levels on the South African Air Quality System (SAAQIS, http://saaqis.environment.gov.za/). These stations monitor criteria pollutants and precursors including PM_{10}, $PM_{2.5}$, carbon monoxide (CO), nitrogen oxides (NOx, NO, and NO_2), ozone (O_3), and sulfur dioxide (SO_2), as well as meteorological factors (Gwaze and Mashele, 2018; South African Air Quality System). Most of these ground stations are in areas with poor air quality, such as low-income settlements that use solid fuel for cooking, heating or lighting (i.e., domestic burning activities), urban areas, areas near large roads, and industrial areas. These stations are situated within communities to assist with the assessment of

population exposure to air pollution. The ambient monitoring stations in South Africa are limited in spatial coverage, data availability, and data quality. Only 20 ground stations had available $PM_{2.5}$ data in our study domain. Hourly raw measurements were only available for 47% of the modeling days in our study period. After quality control, this percentage decreased to 40%.

The South African government promulgated National Ambient Air Quality Standards (NAAQS) for many criteria pollutants in 2009 (Department of Environmental Affairs, 2009), and in 2012 (Department of Environmental Affairs, 2012a), promulgated $PM_{2.5}$ standards. The PM_{10} and $PM_{2.5}$ standards were designed to become more stringent over time. Table S1 displays the $PM_{2.5}$ NAAQS with compliance date; ultimately (i.e., starting 1 January 2030), the 24-hr standard of 25 $\mu g/m^3$ will be equivalent to the WHO Air Quality Guideline, and the annual standard of 15 $\mu g/m^3$ would be equivalent to the WHO Interim Target 3 (World Health Organization, 2006).

Ambient air quality in South Africa is impacted by a range of sources including natural sources such as biomass burning, dust, lightning, and biogenic sources, as well as anthropogenic sources such as industry, vehicles, domestic burning, and waste burning (Wright et al., 2017). In South Africa, coal is the dominant energy resource, providing 69% of primary energy needs and more than 90% of electricity (The World Bank; Zulu et al., 2019). Emissions from other human activities, including industry, mining, mobile vehicle, domestic burning, waste burning, also contribute to $PM_{2.5}$ pollution, resulting in a significant public health burden (Katoto et al., 2019; Pacella et al., 2007; Wright et al., 2017).

Using monitoring data and interpolation with the Benefits Mapping and Analysis Program (BenMAP) model, Altieri and Keen (2019) estimated that if the WHO Guidelines for annual average $PM_{2.5}$ concentrations were met across South Africa, that 28,000 premature deaths (95th percentile CI 15 000 – 52 000) in South Africa could be avoided, with economic costs of over

$29 billion (~4.5% of South African's GDP). A key source of uncertainty in this estimate is from the limited coverage of PM measurements across South Africa.

The South African government declared three national air quality priority areas where ambient air quality does (or is projected to) exceed NAAQS to focus concerted and specific air quality measures to address the poor air quality (Department of Environmental Affairs, 2005). These Priority Areas are the Vaal Triangle Airshed Priority Area (VTAPA) (Department of Environmental Affairs and Tourism, 2006), the Highveld Priority Area (HPA) (Department of Environmental Affairs and Tourism, 2007), and the Waterberg-Bojanala Priority Area (WBPA) (Department of Environmental Affairs, 2012b) (Figure 1). The Priority Areas contain most of South Africa's large industrial hubs and coal-fired power plants. The Priority Areas also contain portions of the Gauteng Province, where the large mega-city conurbation of Johannesburg-Tshwane-Ekurhuleni is located, the domain of this study.

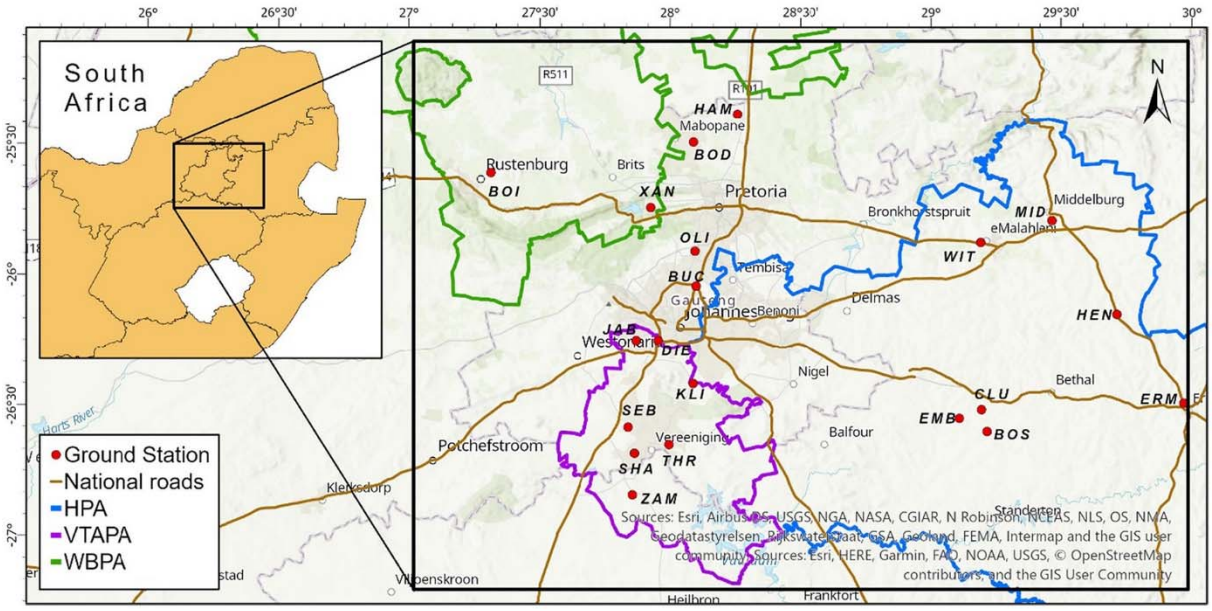

Figure 1. Study Domain and Ground Monitoring Stations.

Ambient air quality levels in the study domain often exceed South African National Ambient Air Quality Standards, with exceedances driven by high $PM_{2.5}$, PM_{10}, and ozone concentrations (Gregor et al., 2019; Kogieluxmie and Venkataraman, 2019; uMoya-NILU, 2017; Venter et al., 2012). A decreasing trend in $PM_{2.5}$ concentrations has been found at some monitoring sites in the HPA and the VTAPA, though at the current rate, compliance with current standards will take years at some sites and decades at others. PM concentrations generally peak in the dry winter season, driven by increases in both emissions (e.g., domestic burning, wind-blown dust, biomass burning) and stagnant meteorological conditions (Hersey et al., 2015; Tyson et al., 1988; Xulu et al., 2020). In general, PM concentrations are higher in low-income settlements compared to monitoring sites in industrial and middle-class residential areas (Hersey et al., 2015). Large-scale regional biomass burning impacts this region in the late winter and spring, leading to peak Aerosol Optical Depth (AOD) levels (Archibald et al., 2010; Duncan et al., 2003; Hersey et al., 2015; Queface et al., 2011; Tesfaye et al., 2011).

The ability to accurately estimate exposure to ground-level $PM_{2.5}$ is essential to study its adverse health effects and assess the effectiveness of air pollution control measures. In this study, we built a 1 km^2 resolution daily $PM_{2.5}$ concentration model in Gauteng Province and the surrounding areas from 2014 to 2018, based on satellite AOD, meteorological fields, land use data, and socioeconomic variables. This is an advanced model rarely developed in South Africa or elsewhere in Africa. Our goal is to explore the feasibility of developing machine learning models in this region with sparse and incomplete ground measurements and gain insight on important predictors of PM2.5 levels. We also assessed the change in regional $PM_{2.5}$ levels before and after the implementation of new national air quality standards.

2. Data and Method

2.1. Study Area

Our study area is approximately 200 x 230 km^2 in the northeast of South Africa on the Highveld Plateau (average altitude ~1700 meters), covering all of Gauteng Province, western Mpumalanga Province, and the eastern part of the North-West Province, as shown in Figure 1. Gauteng Province, with a population of approximately 15 million people, is the country's economic engine, contributed more than a third of South Africa's GDP in 2017 (http://www.statssa.gov.za/?p=11092). Gauteng Province contains the large mega-city conurbation of Johannesburg-Tshwane-Ekurhuleni; Johannesburg is the country's largest city, and Tshwane is its administrative capital. The domain contains part or all of the three declared Priority Areas, all of which have large industrial centers, including the areas around Rustenburg, Vereeniging, and eMalahleni. The domain contains 13 of the country's 15 coal-fired power stations, many of which are in Mpumalanga Province. The population and population density vary greatly across the domain, with peaks in the urban areas of Gauteng Province, with minimums in small cities and rural areas across the other Provinces.

The domain includes a variety of land uses including industrial, mining, urban centers, small cities, and agriculture. The majority of the domain is the grassland biome, with savanna at the northern part (Mucina et al., 2006). The meteorological conditions are often poor for dispersion of pollutants, with stagnant conditions, inversions, and recirculation of pollution most common in winter (Tyson et al., 1988). Precipitation in South Africa is highly seasonal, and this domain is part of the summer rainfall region.

2.2. Ground measurements

There were 20 ground monitoring stations included in our study (Figure 1). Hourly $PM_{2.5}$ data were provided by the South African Weather Service through a request submitted through South Africa Air Quality Information System (SAAQIS, http://saaqis.environment.gov.za/). These data then underwent quality control, with the negative values and repeating values (more than three consecutive identical values) examined and removed if found to be anomalous. Daily averages were only calculated when 75% or more of the hourly measurements were available (Table S2).

2.3. Satellite AOD and Gap Filling

The Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm uses time series analysis and a combination of pixel and image-based processing for Moderate Resolution Imaging Spectroradiometer (MODIS) measurement to get a higher spatial resolution (from 25 to 1 km2) and improve the accuracy of aerosol retrievals (Lyapustin et al., 2018). MAIAC AOD at 550 nm from the Terra (overpass at 10:30 am local time) and Aqua (overpass at 1:30 pm local time) satellites were downloaded from Level-1 and Atmosphere Archive & Distribution System Distributed Active Archive Center (https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/science-domain/maiac/). To improve the coverage of AOD, we developed a customized approach to combine Aqua and Terra AOD retrievals. First, we fitted a simple linear regression between Aqua AOD and Terra AOD by season and used the estimated regression coefficients to estimate the missing Aqua AOD for those grids with only Terra AOD, and vice versa. Secondly, the Aerosol Robotic Network (AERONET) L2 measurements from Pretoria_CSIR-DPSS site, the quality assured

ground based remote sensing aerosol network (https://aeronet.gsfc.nasa.gov/), were used to validate the gap-filled AOD observations. The AERONET AOD at 550 nm within 30 minutes of MAIAC measurement was computed based on AOD at 440 nm and Angstrom exponent (α) of wavelength range 440-675 nm, as Equation 1 shows. We developed a linear mixed-effect model, including season-specific random effects, between the AERONET site AOD and matched pixel AOD and used the resulting regression coefficients to correct gap-filled AOD data. Finally, the mean of validated Aqua and Terra AOD was calculated and used as the parameter in the $PM_{2.5}$ model. The aerosol optical depth and type was downloaded to show whether the smoke/dust model was used in MODIS AOD. In addition, an indicator of fire spot data, which is a type of cloud mask, was also captured from MODIS AOD.

$$AOD_{550nm} = AOD_{440nm} \times \left(\frac{550}{440}\right)^{-\alpha_{(440nm-675nm)}} \quad (1)$$

2.4. Meteorological and Land Use Data

Hourly meteorological data, including surface albedo, surface incident shortwave flux, evaporation from turbulence, total cloud fraction, total precipitation, wind speed, wind direction, planetary boundary layer (PBL) height, temperature, humidity, and surface pressure, with a spatial resolution of 0.25° latitude × 0.3125° longitude, were obtained from Goddard Earth Observing System Data Assimilation System GEOS-5 Forward Processing (GEOS-5 FP, https://gmao.gsfc.nasa.gov/GMAO_products/NRT_products.php). All data were converted to a daily mean value.

The South African National Land-Cover map at 20m resolution was acquired by Sentinel 2 during the period from January 01, 2018 to December 31, 2018 (Department of Environmental Affairs and Department of Rural Development and Land Reform, 2019). The percentage of each

land cover type, such as woodland, grassland, wetland, residential area, rock surface, water surface, old fields, commercial land, and industrial land, was calculated by reclassifying 72 kinds of land use types. The Gridded Population of the World (GPW) version 4 was used for population counts and densities every five years from 2000 to 2020, at the resolution of 30 arc-seconds (Center for International Earth Science Information Network, 2016). Yearly population counts were estimated by linear interpolation. 30-meter elevation data were extracted from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model Version 3 (GDEM 003) (United States National Aeronautics and Space Administration and the Ministry of Economy Trade and Industry of Japan, 2019). The total main road and railway length within each pixel was calculated using ArcGIS.

2.5 Emission and Economic data

Emissions of domestic waste burning and fuel combustion for $PM_{2.5}$ at 3km resolution were developed using a method consistent with the second-generation Vaal Triangle Airshed Priority Area Air Quality Management Plan (Department of Environment Forestry and Fisheries, 2020). Open burning in residential areas was quantified based on available information. Both a top-down (for gas, paraffin, and coal) and a bottom-up (for wood) approach were used for domestic fuel use emissions(Department of Environment Forestry and Fisheries, 2020). Economic data were obtained at the municipal level, including income inequality, income poverty, a multi-dimensional poverty index (MPI), and variance of weighted deprivation scores (David et al., 2018). Income poverty indicates the percentage of residences with income lower than the upper income poverty line in 2011 (Statistics South Africa, 2019). Income equality is represented by the Gini coefficient. Municipalities with a higher Gini coefficient are considered

as high-income inequality. Four dimensions and ten indicators were used to calculate the MPI (David et al., 2018) and they were selected from the recent literature in measuring multidimensional poverty in South Africa (Frame et al., 2016; Statistics South Africa, 2014). MPI was high in poor municipalities and low in rich municipalities. Deprivation is determined by all the indicators at the household level and the household-level deprivation scores assigned by all individuals in the household (David et al., 2018). Weighted deprivation scores were calculated at the individual level.

2.6 Data Integration

The MAIAC AOD data grid at 1km resolution in sinusoidal projection was used as our modeling grid. The spatial alignment of the model predictors was performed as follows. The location of each ground air quality monitor in geographical coordinates was spatially joined to a MAIAC grid cell on the fly in ArcGIS using the nearest neighbor approach. For gridded meteorological parameters at a coarser resolution, the Euclidian distances between the centroids of the coarse resolution grid cells and the centroids of MAIAC grid cells were calculated in R, and the coarse resolution grid cells were mapped to the MAIAC grid cells using the inverse distance weighting approach (Shepard, 1968). Emissions data at 0.03° x 0.03° resolution and municipality-level economic data were re-projected to sinusoidal projection on the fly in ArcGIS, then assigned to each MAIAC grid cell using the nearest neighbor approach. Raster-based land use data at ~30 m resolution was re-projected to sinusoidal projection on the fly in ArcGIS, then the percentage of each major land use type was calculated for each MAIAC grid cell. Total main road and railway length were calculated in each MAIAC grid cell using polygon road network files. The model prediction data was compiled using the same procedure.

2.7. Random Forest Model

Random forest is a flexible, decision-tree based ensemble machine learning algorithm that can capture non-linear relationships and interactions between predictors (Breiman, 2001). It does not make restrictive assumptions of independence and population distribution. Compared with neural network models, results from RF are easier to interpret and can report predictor importance rankings to guide variable selection (Geng et al., 2018). The RF model randomly selects subset samples from all observations with replacements, and then builds multiple decision trees to reach a more accurate and stable prediction (Breiman, 2001). We set the number of decision trees and predictors in each node as 500 and 13, respectively, to achieve the balance of prediction accuracy and computational efficiency. To estimate daily $PM_{2.5}$ concentrations at 1 km resolution in our study domain, we trained a random forest with the response variable being daily mean $PM_{2.5}$ concentration at each monitoring station. Season variables were included in the analysis, including summer (Dec. – Feb.), winter (Jun.- Aug.), and spring (Sep. - Nov.). Since the national standard for ambient $PM_{2.5}$ concentration was changed on January 1st, 2016, we classified the year to a dichotomous variable before policy (2014-2015) and after policy (2016-2018). Other model predictors, as described above, included gap filled daily MAIAC AOD, meteorological factors, percent of land use type, population, elevation, road length, multidimensional poverty data, and emission data from waste and fuel burning. All predictors are listed in Table S3. We carried out 10-fold cross-validation (CV) to evaluate model performance, where we randomly divided the model training dataset into ten equal segments with nine used for training and one for prediction. We repeated this process ten times so that each $PM_{2.5}$ measurement was matched with a prediction. After this, we fitted a linear regression between

observations and predictions. The R^2 value and Root Mean Square Error (RMSE) were used to measure the performance of fitness, along with the intercept and slope.

2.8. Policy Analysis

The $PM_{2.5}$ estimation from the random forest model, divided into before and after the implementation of the more stringent ambient air quality standard for $PM_{2.5}$ (January 1st, 2016), was used to investigate if the change in the standard had an impact on ambient air quality. The percentage of the study area with a $PM_{2.5}$ concentration meeting the standard was calculated and the difference in $PM_{2.5}$ concentration between the two time periods was computed.

3. Results

3.1. Ground Measurements and Gap-filled AOD

Ground measurements were available from January 1st, 2014 through December 31st, 2018. A total of 14,927 daily $PM_{2.5}$ values were calculated from the hourly measurements, ranging from 72 to 1386 per monitoring site. The multi-year mean and standard deviation of daily $PM_{2.5}$ for all stations was 27.23 $\mu g/m^3$ and 20.95 $\mu g/m^3$, respectively, with a range of daily values of 0.91 – 263.88 $\mu g/m^3$.

There was substantial inter-monitoring site variability in the multi-year average $PM_{2.5}$ levels, with the lowest being 16.17 $\mu g/m^3$ in Middelburg and the highest 88.42 $\mu g/m^3$ in Olivenhoutbosch ("OLI" on Figure 1). Most stations had the same seasonal pattern, with lower levels in the summer (Dec-Feb) and higher levels in winter (Jun-Aug). This is consistent with previous studies (Hersey et al., 2015). The majority of monthly aggregated $PM_{2.5}$ measurements were less than 50 $\mu g/m^3$, while some stations had winter $PM_{2.5}$ observations approaching or

exceeding 100 μg/m^3, such as Olivenhoutbosch in 2017 and Xanadu in 2015. Nine of the ten highest daily $PM_{2.5}$ concentrations (> 200 μg/m^3) were recorded in Olivenhoutbosch during June and July 2017.

The gap-filling method increased temporal coverage of available AOD in our study domain from 48% for Aqua and 59% for Terra to 67% for the full gap-filled MAIAC AOD. During our 5-yr study period, the mean and standard deviation of gap-filled AOD of all station-days was 0.153 and 0.07, respectively, with a range of 0.041 to 0.58. The maximum monthly AOD was 0.23, which occurred in September 2015. High AOD readings always occurred in September through November each year which is consistent with a previous study over South Africa (Horowitz et al., 2017; Queface et al., 2011; Tesfaye et al., 2011).

3.2. Random Forest Model Performance

After matching all variables to the 1 km^2 fixed pixel, our final training dataset had 9,853 station-day observations and 40 predictors. The cross validation R^2 and Root Mean Square Error (RMSE) was 0.80 and 9.40 μg/m^3, respectively, indicating satisfactory performance from the random forest model (Figure 2). The slope and intercept of the univariate linear regression between $PM_{2.5}$ measurement and estimation were 1.13 and -3.66, respectively, indicating that the random forest model might slightly overestimate at low $PM_{2.5}$ concentrations and underestimate at high $PM_{2.5}$ values, especially when daily $PM_{2.5}$ concentration exceeds 150 μg/m^3. Since $PM_{2.5}$ measurements from Olivenhoutbosch station were always high, with the peak over 150 μg/m^3, we did a sensitive analysis to explore whether the model performed better when we removed this station. However, there was almost no change to the slope of the regression line. We also evaluated the model performance at the top 20% of all station-day observations and its

corresponding predictions. The CV R^2 and RMSE was 0.69 and 14.36 μg/m^3, respectively, along with the regression slope of 1.01 and intercept of -8.81 (Figure S1), indicating good model prediction accuracy and a low overall bias as high concentration levels.

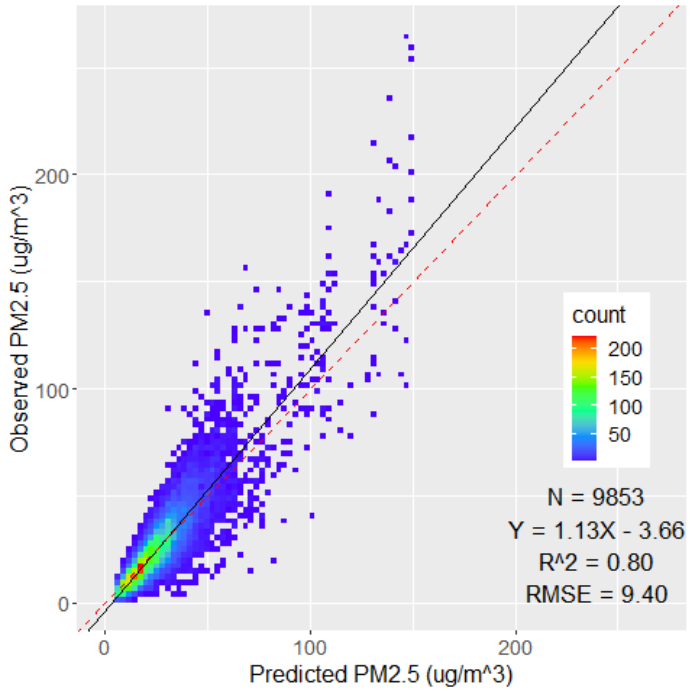

Figure 2. Scatter Plot for 10-fold Cross-validation of Daily $PM_{2.5}$. Red dashed line is the 1:1 line.

The importance ranking of random forest model predictors is provided in Figure S2, a measurement of the predictive power of independent variables. The indicator of season was the most important predictor, followed by gap filled MAIAC AOD, total precipitation, population, and the indicator of policy.

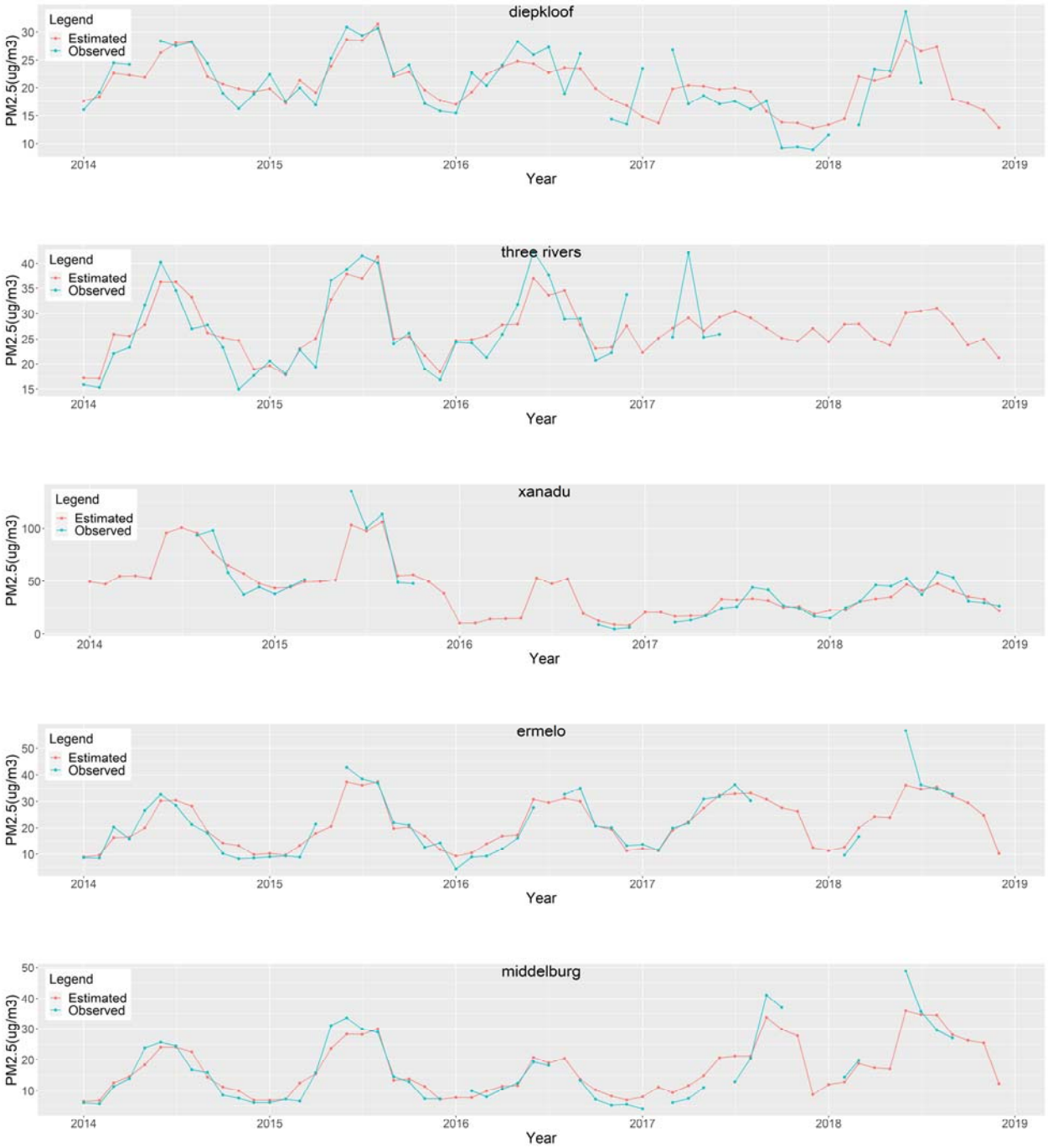

Figure 3. Observed (blue) and Estimated (red) Monthly $PM_{2.5}$ Concentration for five representative stations.

3.3. Model Predicted $PM_{2.5}$ Temporal and Spatial Patterns

Figure 3 shows the time series plots for monthly mean $PM_{2.5}$ ground measurements and our model predictions at five representative stations: Diepkloof, Three Rivers, Xanadu, Ermelo, and Middelburg. The Diepkloof station is located within a densely populated area close to the main roads into Johannesburg from the southwest. The Three Rivers station is located in a middle-class suburb at the southernmost of Gauteng Province and within the footprint of Eskom's Lethabo Power Station. Xanadu site is located in the peri-urban area west of the City of Tshwane in the Northwest Province with fewer strong local sources, though it is impacted by the outflow from the City of Tshwane. The Ermelo and Middelburg stations are located in low-income and middle-income residential areas in Mpumalanga Province, respectively, with similar land cover types. Time series plots for other stations are in Figure S3. Our model captured seasonal trends well, showing the highest $PM_{2.5}$ levels in winter (Jun- Aug) and the lowest levels in summer (Dec – Feb) across stations. While our model tended to underestimate at very high monthly $PM_{2.5}$ concentrations, i.e., close to 35 $\mu g/m^3$, the absolute difference between ground measurements and model predictions was generally less than 5 $\mu g/m^3$ at the monthly level.

Figure 4 shows the spatial distribution of model-predicted seasonal mean $PM_{2.5}$ concentrations at 1 km resolution. There is seasonal and spatial variability of $PM_{2.5}$, with the highest values centered around populated Gauteng Province. The lowest concentrations were generally seen in Mpumalanga Province. There is also an area to the west of Johannesburg that had lower $PM_{2.5}$ than its surrounding areas. The highest $PM_{2.5}$ was between the cities of Johannesburg and Tshwane. Northern Gauteng and Mpumalanga Provinces in the southwestern part of our study domain also had high $PM_{2.5}$ levels. The estimated mean $PM_{2.5}$ concentration was highest in winter, ranging from 21 $\mu g/m^3$ to 123 $\mu g/m^3$, with the area-averaged mean above 50

μg/m^3. The $PM_{2.5}$ in spring and fall were similar with an area-average mean around 35 μg/m^3. $PM_{2.5}$ concentrations in summer were the lowest and varied from 8 μg/m^3 to 67 μg/m^3 with an area-average mean value around 26 μg/m^3. In winter, the highest $PM_{2.5}$ values occurred in the northwest of the study domain whereas in the eastern part, estimated $PM_{2.5}$ was generally much lower and always less than 30 μg/m^3. In spring, summer, and fall, the estimated $PM_{2.5}$ was high at the junction of Mpumalanga and Gauteng north of Johannesburg. $PM_{2.5}$ concentration was less than 25 μg/m^3 on east of the study domain across all seasons.

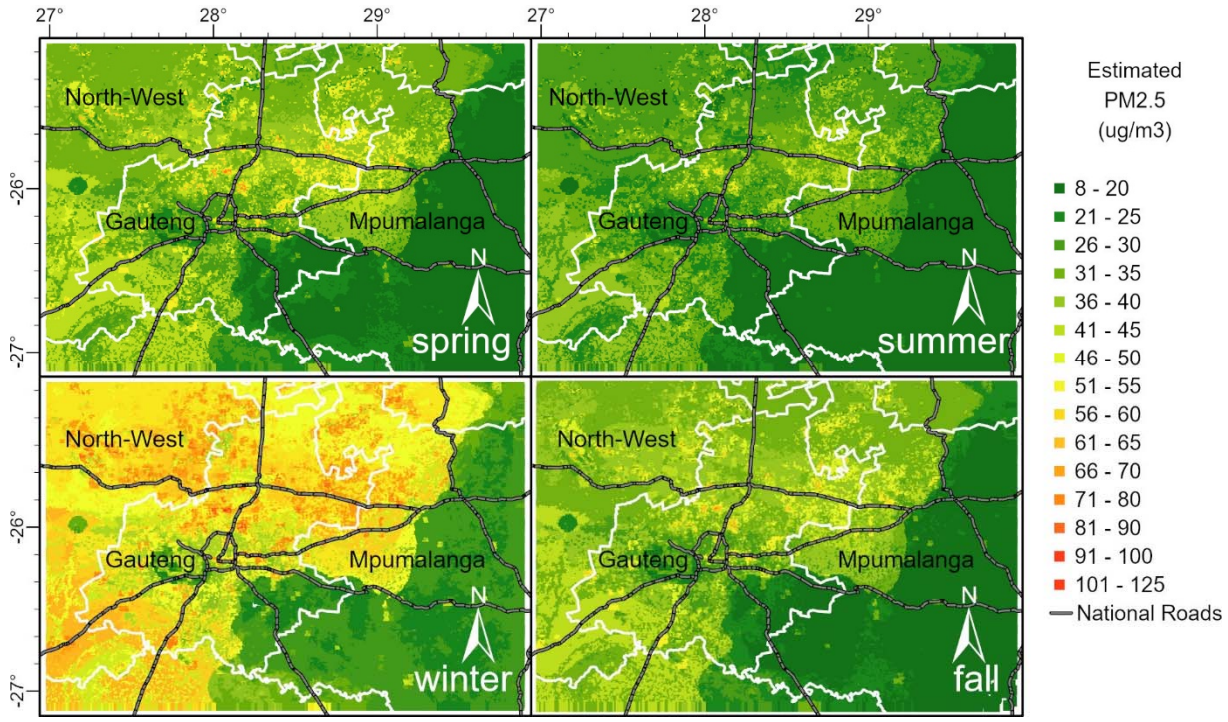

Figure 4. Five-year Average Seasonal Estimated ground-level $PM_{2.5}$ Concentration Map.

Figure S4 reports the annual average of $PM_{2.5}$ for each station in our study domain, which ranged from 15.97 μg/m^3 to 97.62 μg/m^3 and was similar between years. Similar to the seasonal averages, the lowest estimated annual averages were seen in the Mpumalanga and the south-eastern Gauteng, and the highest in and around Gauteng Province. The overall level of $PM_{2.5}$ was the lowest in 2016, but had similar levels in the following two years.

3.4. $PM_{2.5}$ Concentration Changes in Implementation of New Standard

During the full study period, only 31% of the study area had annual $PM_{2.5}$ concentrations below 25 μg/m^3 (the old standard), with 14% below 20 μg/m^3 (the new standard), indicating that the majority of the domain was out of compliance with the national $PM_{2.5}$ standards.

Figure 5 displays the difference in $PM_{2.5}$ concentration before (2014-215) and after (2016-2018) the change in the $PM_{2.5}$ air quality standard. $PM_{2.5}$ concentrations decreased in most of the study domain after this change. A reduction of 0.5 μg/m^3 to 2.0 μg/m^3 in $PM_{2.5}$ concentration occurred in the southern part of the study domain after the new standard came into effect. The reductions were even greater (above 2.0 μg/m^3) in the northern area, particularly in the northwest. However, Johannesburg and the surrounding areas did not experience an obvious change.

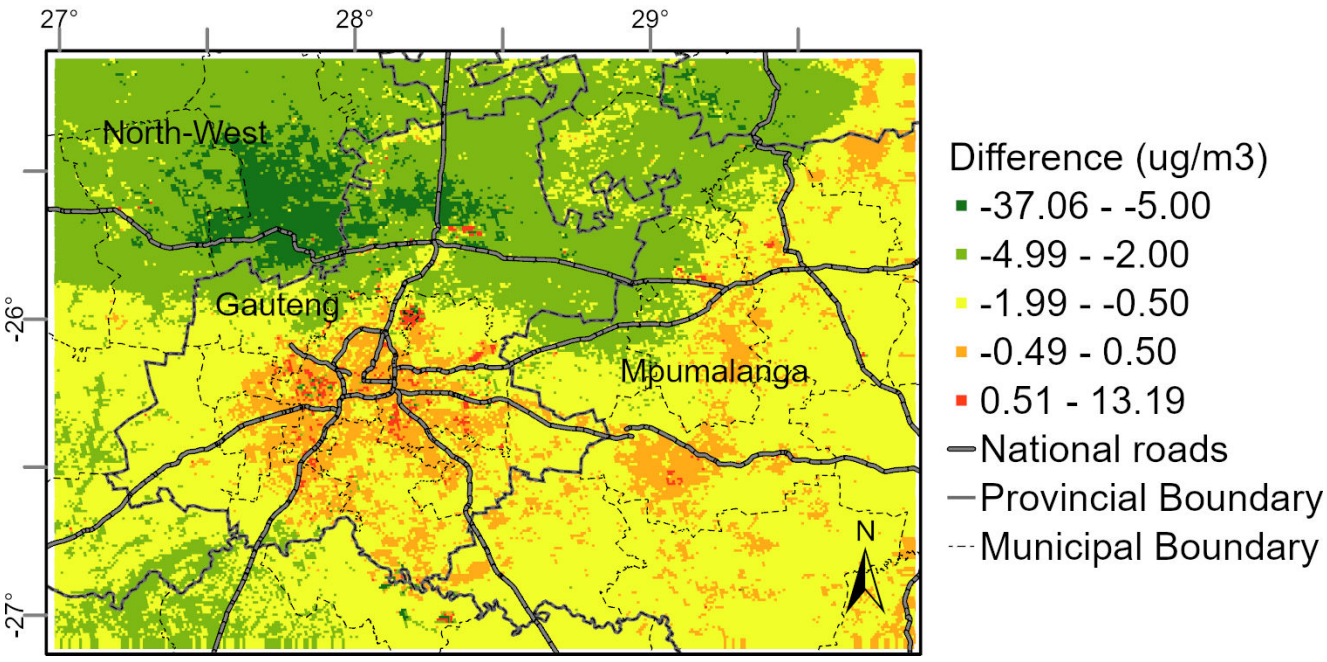

Figure 5. Difference in Annual $PM_{2.5}$ Concentration before (2014-215) and after (2016-2018) the new $PM_{2.5}$ standard.

4. Discussion

It is well known that fine particles are associated with a large burden of disease in South Africa (Altieri and Keen, 2019; Bauer et al., 2019; Global Burden of Disease, 2016; Lim et al., 2012; Wichmann et al., 2012). Nevertheless, the monitoring network in South Africa is uneven, and mainly located in densely populated urban centers or industrial areas. The monitors have variable data quality and data capture rates, hindering $PM_{2.5}$ exposure assessment for large segments of the South African population. Several modeling studies have estimated $PM_{2.5}$ concentration in the country. For example, Saucy et al. (2018) developed an annual land use regression model for outdoor $PM_{2.5}$ concentration in the Western Cape Province. Their R^2 of 0.21 indicates that much of the variability in $PM_{2.5}$ was not captured. Marais et al. (Marais et al., 2019) applied the GEOS-Chem model to simulate ambient $PM_{2.5}$ concentrations across Africa in 2012 and 2030, and found that $PM_{2.5}$ concentrations in 2012 for sites within this study's domain were overestimated by their model. Available gridded modeled datasets, such as those used in the Global Burden of Disease project, have large uncertainties when compared with available monitoring data (Garland et al., 2017; Luckson et al., 2020). The coarse spatial resolution and low predictive power of these simulated air quality data limit their applications in health effects research and policy assessment.

The random forest algorithm we developed was based on the 1 km resolution MAIAC AOD. This is anadvanced model for South Africa with high prediction accuracy at the daily level and at a relatively high spatial resolution. Our model is able to simulate the monthly and seasonal cycle across the domain well, showing the expected increase in ground-level $PM_{2.5}$ concentrations in the winter with stagnant air and little rain. Spatially, the highest $PM_{2.5}$ values were found within and around Gauteng Province, and a majority of the area did not meet the

national ambient air quality standards. Due to the large population and the large number of emission sources, it is expected that $PM_{2.5}$ concentrations would be high in Gauteng. The Vaal Triangle Airshed Priority Area (VTAPA) in southern Gauteng is known to have high air pollution levels due to large and numerous emissions sources. In addition, low-income settlements that use domestic burning for heating and cooking, and mining and industrial areas in western Mpumalanga also show high $PM_{2.5}$ levels. Interestingly, some high $PM_{2.5}$ concentrations were present far outside the urban centers, such as to the northwest through the southwest. There have been few to no measurements in these areas, and thus the performance of the model in this area is difficult to quantify but indicates the need for additional monitoring to better understand the $PM_{2.5}$ concentrations and spatial trends.

While a great number of $PM_{2.5}$ models have been reported worldwide, model performance largely depends on the availability of ground monitoring data. It has been shown that satellite based $PM_{2.5}$ estimates are very sensitive to the number of ground monitoring sites (Geng et al., 2018). Therefore, most $PM_{2.5}$ models reported in the literature were developed in regions with sufficient training data, such as the US, China, and Europe. There were no advanced $PM_{2.5}$ modeling studies in South Africa, providing us no reference on drivers of $PM_{2.5}$ distribution in this region. Our study explored the feasibility of developing a high-performance $PM_{2.5}$ spatiotemporal model in South Africa. The predictor importance rankings and relatively straightforward interpretation of the random forest algorithm allowed us to identify important and unique predictors of $PM_{2.5}$ in our study region. For example, incorporating emissions and economic indicators as well as land cover data specifically developed for South Africa improved model CV R^2 from 0.67 to ~0.80. Our findings would provide guidance to future development of more complex machine learning $PM_{2.5}$ models with less interpretability in this data-poor region. Our model was anadvanced algorithm to estimate daily $PM_{2.5}$ concentrations in South Africa with

comparable performance to the models reported in data rich regions such as the US (Hu et al., 2014; Liu et al., 2009) and China (He and Huang, 2018; Lin et al., 2018; Ma et al., 2016a). It enables the research of the air pollution health effects in South Africa and sets an example for similar efforts in Africa. Compared with regression-based models, random forest algorithm has better prediction accuracy and avoids overfitting (Hu et al., 2017). Although neural network models may achieve higher performance when fully trained with a large dataset (Hu et al., 2017), none of the published studies using neural networks offered any insight into the importance of various predictors. Most recently, model interpretablity tools started to merge for neural network models such as Captum, which can be used to better understand predictors importace (Kokhlikyan et al., 2020). However, Captum was developed for Python programs and no similar R packages are available to date.

One limitation of the model is that it slightly underestimates $PM_{2.5}$ when concentrations are high concentrations, and overestimate the low levels. One possible reason is that AOD – the second most important predictor – is a quantitative measure for aerosols abundance in the entire atmospheric column. Previous research has found that the AOD - $PM_{2.5}$ relationship varies by aerosol composition and vertical profile. While $PM_{2.5}$ levels peak in winter in South Africa, AOD peaks in late winter and early spring and is driven strongly by regional biomass burning (Hersey et al., 2015; Queface et al., 2011; Tesfaye et al., 2011). Therefore, different aerosol composition and vertical profiles at different times of the year may alter the AOD – $PM_{2.5}$ relationship and affect model performance. Among the lowest 20% station-day observations, observed $PM_{2.5}$ concentrations in summer and winter accounted for 31.27% and 12.89%, respectively. The overall lower quality of AOD data in summer may also affect model performance at low values. Compared with other seasons, there were more missing AOD data in summer, and our AOD gap filling model had the lowest accuracy in summer, resulting in more uncertainty in the gap-filled

AOD values used in model training and prediction. Finally, there are limited ground-based monitoring sites outside of the urban and industrial areas to fully train the random forest to represent $PM_{2.5}$ levels in these regions.

While the policy analysis was limited due to the short time-span of the data, the variability in $PM_{2.5}$ concentrations during our study period, calculated from both ground monitoring and our model, shows a consistent trend, with lower concentrations after the implementation of the new standard (2016-2018) compared to the previous two years. A decrease was also observed in many other parts of the study area, though not for Johannesburg.

Although the model has a good fit and captures the temporal variability in the study area, it could be improved in the future given data from more monitoring stations and longer time-series. In our training dataset, half of the ground stations only had $PM_{2.5}$ concentration data available for less than half of the study period, which would influence prediction accuracy during the rest of the time. In addition, there was a significant amount of missing AOD values even after the gap filling procedure due to extensive cloud cover, which could bias the annual averages of $PM_{2.5}$ but could be corrected in future work given better and longer coverage. These would be important extensions of the model developments we have described here.

5. Conclusion

Our model is anadvanced, high-resolution model to estimate daily $PM_{2.5}$ concentration in South Africa, with a domain of 200 x 230 km^2 that covers the population-dense Gauteng Province and surrounding area. The model was able to reproduce the marked seasonal pattern characteristic of northeastern South Africa and has high prediction accuracy at the daily level with an overall cross-validation R^2 of 0.8. Since the change in the national $PM_{2.5}$ standard in

2016, we observed a reduction of $PM_{2.5}$ in most of our study region, although levels in Johannesburg and the neighboring areas have remained relatively constant. The purpose of the Multi-Angle Imager for Aerosols (MAIA) investigation is to track harmful particulate matter using satellite data. As our study domain was one of the primary target areas of South Africa, our research can help reproduce important particulate matter data. The satellites data missed the peak of $PM_{2.5}$ concentration, so we also use AERONET data to make the observed $PM_{2.5}$ concentration close to the surface $PM_{2.5}$. This study relied on the use of simple linear regression and mixed-effect models to gap-fill MAIAC AOD values. To further improve prediction capabilities, more accurate and robust models need to be developed. The $PM_{2.5}$ estimates derived from this model could be applied to future epidemiologic studies, burden of disease assessments, and other policy evaluations. Extensions may include broadening the modeling domain and improving model performance through longer and more spatially dispersed observational time-series or from improved satellite coverage. Future studies should further analyze the composition of $PM_{2.5}$ and its adverse effects on health and the environment.

Acknowledgment

This work was partially supported by the MAIA science team at the JPL, California Institute of Technology, led by D. Diner (Subcontract #1588347). NS was supported by the NIEHS-funded HERCULES Center (P30ES019776). RMG and MN were supported by a CSIR Parliamentary Grant. We thank the PI of the Pretoria_CSIR-DPSS site from AERONET for establishing and maintaining the site. The authors thank the Department of Environment, Forestry and Fisheries, the South African Weather Service, and the air quality network owners of the Cities of Johannesburg, Tshwane, Ekurhuleni, and Sasol for the air quality data.

References

Altieri, K.E., & Keen, S.L., 2019. Public health benefits of reducing exposure to ambient fine particulate matter in South Africa. *Sci Total Environ, 684*, 610-620

Archibald, S., Scholes, R.J., Roy, D.P., Roberts, G., & Boschetti, L., 2010. Southern African fire regimes as revealed by remote sensing %J International Journal of Wildland Fire*, 19*, 861-878

Atkinson, R.W., Kang, S., Anderson, H.R., Mills, I.C., & Walton, H.A., 2014. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax, 69*, 660-665

Bauer, S., Im, U., Mezuman, K., & Gao, C., 2019. Desert Dust, Industrialization, and Agricultural Fires: Health Impacts of Outdoor Air Pollution in Africa. Journal of Geophysical Research: Atmospheres

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R.V., Dentener, F., Dingenen, R.v., Estep, K., Amini, H., Apte, J.S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P.K., Knibbs, L.D., Kokubo, Y., Liu, Y., Ma, S., Morawska, L., Sangrador, J.L.T., Shaddick, G., Anderson, H.R., Vos, T., Forouzanfar, M.H., Burnett, R.T., & Cohen, A., 2016. Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013. *Environmental Science & Technology, 50*, 79-88

Breiman, L., 2001. Random Forests. *Machine Learning, 45*, 5-32

Brokamp, C., Jandarov, R., Hossain, M., & Ryan, P., 2018. Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model. *Environmental Science & Technology, 52*, 4173-4179

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C.A., 3rd, Apte, J.S., Brauer, M., Cohen, A., Weichenthal, S., Coggins, J., Di, Q., Brunekreef, B., Frostad, J., Lim,

S.S., Kan, H., Walker, K.D., Thurston, G.D., Hayes, R.B., Lim, C.C., Turner, M.C., Jerrett, M., Krewski, D., Gapstur, S.M., Diver, W.R., Ostro, B., Goldberg, D., Crouse, D.L., Martin, R.V., Peters, P., Pinault, L., Tjepkema, M., van Donkelaar, A., Villeneuve, P.J., Miller, A.B., Yin, P., Zhou, M., Wang, L., Janssen, N.A.H., Marra, M., Atkinson, R.W., Tsang, H., Quoc Thach, T., Cannon, J.B., Allen, R.T., Hart, J.E., Laden, F., Cesaroni, G., Forastiere, F., Weinmayr, G., Jaensch, A., Nagel, G., Concin, H., & Spadaro, J.V., 2018. Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proc Natl Acad Sci U S A, 115*, 9592-9597

Center for International Earth Science Information Network, 2016. Gridded Population of the World, Version 4 (GPWv4): Administrative Unit Center Points with Population Estimates. In Columbia University (Ed.)

Chang, H.H., Hu, X., & Liu, Y., 2013. Calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentrations via statistical downscaling. *J Expos Sci Environ Epidemiol, 24*, 398-404

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., & Dandona, R.J.T.L., 2017. 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*, 389*, 1907-1918

David, A., Guilbert, N., Hamaguchi, N., Higashi, Y., Hino, H., Leibbrandt, M., & Shifa, M., 2018. Spatial poverty and inequality in South Africa: A municipality level analysis

Department of Environment Forestry and Fisheries, 2020. DRAFT SECOND GENERATION AIR QUALITY MANAGEMENT PLAN FOR VAAL TRIANGLE AIRSHED PRIORITY AREA

Department of Environmental Affairs, 2005. National Envrionment Management: Air Quality Act [No. 39 of 2004]. https://www.gov.za/sites/default/files/gcis_document/201409/a39-04.pdf

Department of Environmental Affairs, 2009. National Environmental Management: Air Quality Act of 2004 (Act No. 39 of 2004), National Ambient Air Quality Standards, Government Gazette No. 1210

Department of Environmental Affairs, 2012a. National Environmental Management: Air Quality Act of 2004 (Act No. 39 of 2004), National Ambient Air Quality Standards for Particulate Matter (PM2.5), Government Gazette No. 486

Department of Environmental Affairs, 2012b. NATIONAL ENVIRONMENTAL MANAGEMENT: AIR QUALITY ACT, 2004 (ACT NO. 39 OF 2004) DECLARATION OF THE WATERBERG NATIONAL PRIORITY AREA. https://www.environment.gov.za/sites/default/files/gazetted_notices/nemaqa_waterberg_declaration_g35435gen495.pdf

Department of Environmental Affairs, & Department of Rural Development and Land Reform, 2019. 2018 South African National Land-Cover Change Assessments, DEA E1434 Land-Cover

Department of Environmental Affairs and Tourism, 2006. DECLARATION OF THE VAAL TRIANGLE AIR-SHED PRIORITY AREA IN TERMS OF SECTION 18(1) OF THE NATIONAL ENVIRONMENTAL MANAGEMENT: AIR QUALITY ACT 2004, (ACT NO. 39 OF 2004) https://www.gov.za/sites/default/files/gcis_document/201409/28732b.pdf

Department of Environmental Affairs and Tourism, 2007. DECLARATION OF THE HIGHVELD ASPRIORITY AREA IN TERMS OF SECTION 18(1) OF THE NATIONAL ENVIRONMENTAL MANAGEMENT: AIR QUALITY ACT, 2004 (ACT NO. 39 OF 2004). https://www.gov.za/sites/default/files/gcis_document/201409/30518.pdf

Duncan, B.N., Martin, R.V., Staudt, A.C., Yevich, R., & Logan, J.A.J.J.o.G.R.A., 2003. Interannual and seasonal variability of biomass burning emissions constrained by satellite observations*, 108*, ACH 1-1-ACH 1-22

Frame, E., De Lannoy, A., & Leibbrandt, M., 2016. *Measuring multidimensional poverty among youth in South Africa at the sub-national level*.

Garland, R.M., Naidoo, M., Sibiya, B., & Oosthuizen, R., 2017. Air quality indicators from the Environmental Performance Index: potential use and limitations in South Africa. *Clean Air Journal, 27*

Geng, G., Murray, N.L., Chang, H.H., & Liu, Y., 2018. The sensitivity of satellite-based PM2.5 estimates to its inputs: Implications to model development in data-poor regions. *Environment International, 121*, 550-560

Global Burden of Disease, 2016. Global burden of air pollution. Institute for Health Metrics and Evaluation (IHME). In

Gregor, F., Rebecca, M.G., Seneca, N., Amukelani, M., & Marna Van der, M., 2019. Assessment of changes in concentrations of selected criteria pollutants in the Vaal and Highveld Priority Areas. *Clean Air Journal, 29*

Gwaze, P., & Mashele, S.H.J.C.A.J., 2018. South African Air Quality Information System (SAAQIS) mobile application tool: bringing real time state of air quality to South Africans*, 28*, 3-3

He, Q., & Huang, B., 2018. Satellite-based mapping of daily high-resolution ground PM2.5 in China via space-time regression modeling. *Remote Sensing of Environment, 206*, 72-83

Hersey, S., Garland, R.M., Crosbie, E., Shingler, T., Sorooshian, A., Piketh, S., Burger, R.J.A.c., & physics, 2015. An overview of regional and local characteristics of aerosols in South Africa using satellite, ground, and modeling data*, 15*, 4259

Hoff, R.M., & Christopher, S.A., 2009. Remote Sensing of Particulate Pollution from Space: Have We Reached the Promised Land? *J. Air Waste Manage. Assoc., 59*, 645-675

Horowitz, H.M., Garland, R.M., Thatcher, M., Landman, W.A., Dedekind, Z., Merwe, J.v.d., Engelbrecht, F.A.J.A.C., & Physics, 2017. Evaluation of climate model aerosol seasonal and spatial variability over Africa using AERONET*, 17*, 13999-14023

Hu, X., Belle, J., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., & Liu, Y., 2017. Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environmental Science & Technology, 51*, 6936-6944

Hu, X., Waller, L.A., Lyapustin, A., Wang, Y., & Liu, Y., 2014. Improving satellite-driven PM2.5 models with Moderate Resolution Imaging Spectroradiometer fire counts in the southeastern U.S. *Journal of Geophysical Research: Atmospheres, 119*, 11,375-311,386

Katoto, P.D.M.C., Byamungu, L., Brand, A.S., Mokaya, J., Strijdom, H., Goswami, N., De Boever, P., Nawrot, T.S., & Nemery, B., 2019. Ambient air pollution and health in Sub-Saharan Africa: Current evidence, perspectives and a call to action. *Environmental Research, 173*, 174-188

Kloog, I., Chudnovsky, A.A., Just, A.C., Nordio, F., Koutrakis, P., Coull, B.A., Lyapustin, A., Wang, Y., & Schwartz, J., 2014. A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmospheric Environment, 95*, 581-590

Kogieluxmie, G., & Venkataraman, S., 2019. A decadal analysis of particulate matter (PM2.5) and surface ozone (O3) over Vaal Priority Area, South Africa. *Clean Air Journal, 29*

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., & Yan, S.J.a.p.a., 2020. Captum: A unified and generic model interpretability library for pytorch

Li, L., Franklin, M., Girguis, M., Lurmann, F., Wu, J., Pavlovic, N., Breton, C., Gilliland, F., & Habre, R., 2020. Spatiotemporal imputation of MAIAC AOD using deep learning with downscaling. *Remote Sensing of Environment, 237*, 111584

Lim, S.S., Vos, T., Flaxman, A.D., Danaei, G., Shibuya, K., Adair-Rohani, H., Amann, M., Anderson, H.R., Andrews, K.G., Aryee, M., Atkinson, C., Bacchus, L.J., Bahalim, A.N., Balakrishnan, K., Balmes, J., Barker-Collo, S., Baxter, A., Bell, M.L., Blore, J.D., Blyth, F., Bonner, C., Borges, G., Bourne, R., Boussinesq, M., Brauer, M., Brooks, P., Bruce, N.G., Brunekreef, B., Bryan-Hancock, C., Bucello, C., Buchbinder, R., Bull, F., Burnett, R.T., Byers, T.E., Calabria, B., Carapetis, J., Carnahan, E., Chafe, Z., Charlson, F., Chen, H., Chen, J.S., Cheng, A.T., Child, J.C., Cohen, A., Colson, K.E., Cowie, B.C., Darby, S., Darling, S., Davis, A., Degenhardt, L., Dentener, F., Des Jarlais, D.C., Devries, K., Dherani, M., Ding, E.L., Dorsey, E.R., Driscoll, T., Edmond, K., Ali, S.E., Engell, R.E., Erwin, P.J., Fahimi, S., Falder, G., Farzadfar, F., Ferrari, A., Finucane, M.M., Flaxman, S., Fowkes, F.G., Freedman, G., Freeman, M.K., Gakidou, E., Ghosh, S., Giovannucci, E., Gmel, G., Graham, K., Grainger, R., Grant, B., Gunnell, D., Gutierrez, H.R., Hall, W., Hoek, H.W., Hogan, A., Hosgood, H.D., 3rd, Hoy, D., Hu, H., Hubbell, B.J., Hutchings, S.J., Ibeanusi, S.E., Jacklyn, G.L., Jasrasaria, R., Jonas, J.B., Kan, H., Kanis, J.A., Kassebaum, N., Kawakami, N., Khang, Y.H., Khatibzadeh, S., Khoo, J.P., Kok, C., Laden, F., Lalloo, R., Lan, Q., Lathlean, T., Leasher, J.L., Leigh, J., Li, Y., Lin, J.K., Lipshultz, S.E., London, S., Lozano, R., Lu, Y., Mak, J., Malekzadeh, R., Mallinger, L., Marcenes, W., March, L., Marks, R., Martin, R., McGale, P., McGrath, J., Mehta, S., Mensah, G.A., Merriman, T.R., Micha, R., Michaud, C., Mishra, V., Mohd Hanafiah, K., Mokdad, A.A., Morawska, L., Mozaffarian, D., Murphy, T., Naghavi, M., Neal, B., Nelson, P.K., Nolla, J.M., Norman, R., Olives, C., Omer, S.B., Orchard, J., Osborne, R., Ostro, B., Page, A., Pandey, K.D., Parry, C.D., Passmore, E., Patra, J., Pearce, N., Pelizzari, P.M., Petzold, M., Phillips, M.R., Pope,

D., Pope, C.A., 3rd, Powles, J., Rao, M., Razavi, H., Rehfuess, E.A., Rehm, J.T., Ritz, B., Rivara, F.P., Roberts, T., Robinson, C., Rodriguez-Portales, J.A., Romieu, I., Room, R., Rosenfeld, L.C., Roy, A., Rushton, L., Salomon, J.A., Sampson, U., Sanchez-Riera, L., Sanman, E., Sapkota, A., Seedat, S., Shi, P., Shield, K., Shivakoti, R., Singh, G.M., Sleet, D.A., Smith, E., Smith, K.R., Stapelberg, N.J., Steenland, K., Stockl, H., Stovner, L.J., Straif, K., Straney, L., Thurston, G.D., Tran, J.H., Van Dingenen, R., van Donkelaar, A., Veerman, J.L., Vijayakumar, L., Weintraub, R., Weissman, M.M., White, R.A., Whiteford, H., Wiersma, S.T., Wilkinson, J.D., Williams, H.C., Williams, W., Wilson, N., Woolf, A.D., Yip, P., Zielinski, J.M., Lopez, A.D., Murray, C.J., Ezzati, M., AlMazroa, M.A., & Memish, Z.A., 2012. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet, 380*, 2224-2260

Lin, C.Q., Liu, G., Lau, A.K.H., Li, Y., Li, C.C., Fung, J.C.H., & Lao, X.Q., 2018. High-resolution satellite remote sensing of provincial PM2.5 trends in China from 2001 to 2015. *Atmospheric Environment, 180*, 110-116

Liu, C., Chen, R., Sera, F., Vicedo-Cabrera, A.M., Guo, Y., Tong, S., Coelho, M., Saldiva, P.H.N., Lavigne, E., Matus, P., Valdes Ortega, N., Osorio Garcia, S., Pascal, M., Stafoggia, M., Scortichini, M., Hashizume, M., Honda, Y., Hurtado-Diaz, M., Cruz, J., Nunes, B., Teixeira, J.P., Kim, H., Tobias, A., Iniguez, C., Forsberg, B., Astrom, C., Ragettli, M.S., Guo, Y.L., Chen, B.Y., Bell, M.L., Wright, C.Y., Scovronick, N., Garland, R.M., Milojevic, A., Kysely, J., Urban, A., Orru, H., Indermitte, E., Jaakkola, J.J.K., Ryti, N.R.I., Katsouyanni, K., Analitis, A., Zanobetti, A., Schwartz, J., Chen, J., Wu, T., Cohen, A., Gasparrini, A., & Kan, H., 2019. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities. *N Engl J Med, 381*, 705-715

Liu, Y., Paciorek Christopher, J., & Koutrakis, P., 2009. Estimating Regional Spatial and Temporal Variability of PM2.5 Concentrations Using Satellite Data, Meteorology, and Land Use Information. *Environmental Health Perspectives, 117*, 886-892

Liu, Y., Sarnat, J.A., Kilaru, A., Jacob, D.J., & Koutrakis, P., 2005. Estimating ground-level PM2.5 in the eastern united states using satellite remote sensing. *Environ. Sci. Technol., 39*, 3269-3278

Luckson, M., Roelof, B., & Stuart, J.P., 2020. Evaluating the potential of remote sensing imagery in mapping ground-level fine particulate matter (PM2.5) for the Vaal Triangle Priority Area. *Clean Air Journal, 30*

Lyapustin, A., Wang, Y., Korkin, S., & Huang, D.J.A.M.T., 2018. MODIS Collection 6 MAIAC algorithm*, 11*

Ma, Z., Hu, X., Huang, L., Bi, J., Liu, Y.J.E.s., & technology, 2014. Estimating ground-level PM2. 5 in China using satellite remote sensing*, 48*, 7436-7444

Ma, Z., Hu, X., Sayer, A., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L., & Liu, Y., 2016a. Satellite-Based Spatiotemporal Trends in PM2.5 Concentrations: China, 2004–2013. *Environmental Health Perspectives, 124*, 184-192

Ma, Z., Liu, Y., Zhao, Q., Liu, M., Zhou, Y., & Bi, J.J.A.E., 2016b. Satellite-derived high resolution PM2. 5 concentrations in Yangtze River Delta Region of China using improved linear mixed effects model*, 133*, 156-164

Marais, E.A., Silvern, R.F., Vodonos, A., Dupin, E., Bockarie, A.S., Mickley, L.J., Schwartz, J.J.E.S., & Technology, 2019. Air quality and health impact of future fossil fuel use for electricity generation and transport in Africa*, 53*, 13524-13534

Mucina, L., Rutherford, M.C., & Powrie, L.W., 2006. Vegetation Atlas of South Africa, Lesotho and Swaziland

Pacella, R., Cairncross, E., Witi, J., & Bradshaw, D., 2007. Estimating the burden of disease attributable to urban outdoor air pollution in South Africa in 2000. *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde, 97*, 782-790

Queface, A.J., Piketh, S.J., Eck, T.F., Tsay, S.-C., & Mavume, A.F., 2011. Climatology of aerosol optical properties in Southern Africa. *Atmospheric Environment, 45*, 2910-2921

Saucy, A., Röösli, M., Künzli, N., Tsai, M.-Y., Sieber, C., Olaniyan, T., Baatjies, R., Jeebhay, M., Davey, M., Flückiger, B.J.I.j.o.e.r., & health, p., 2018. Land use regression modelling of outdoor NO2 and PM2. 5 concentrations in three low income areas in the western cape province, South Africa*, 15*, 1452

Seinfeld, J.H., & Pandis, S.N., 2016. *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons

Shepard, D., 1968. A two-dimensional interpolation function for irregularly-spaced data. In, *Proceedings of the 1968 23rd ACM national conference* (pp. 517-524)

Sorek-Hamer, M., Chatfield, R., & Liu, Y., 2020. Review: Strategies for using satellite-based products in modeling PM2.5 and short-term pollution episodes. *Environment International, 144*, 106057

South African Air Quality System. http://saaqis.environment.gov.za/ (accessed. January 13th, 2021)

Statistics South Africa, 2014. The South African MPI: Creating a multidimensional poverty index using census data

Statistics South Africa, 2019. National Poverty Lines. http://www.statssa.gov.za/publications/P03101/P031012019.pdf

Strawa, A.W., Chatfield, R.B., Legg, M., Scarnato, B., & Esswein, R., 2013. Improving retrievals of regional fine particulate matter concentrations from Moderate Resolution Imaging

Spectroradiometer (MODIS) and Ozone Monitoring Instrument (OMI) multisatellite observations. *Journal of the Air & Waste Management Association, 63*, 1434-1446

Tesfaye, M., Sivakumar, V., Botai, J., & Mengistu Tsidu, G., 2011. Aerosol climatology over South Africa based on 10 years of Multiangle Imaging Spectroradiometer (MISR) data. *Journal of Geophysical Research: Atmospheres, 116*

The World Bank, 2015. Electricity Production from Coal Sources (% of total) - South Africa. https://data.worldbank.org/indicator/EG.ELC.COAL.ZS?locations=ZA (accessed. Januray 15th, 2021)

Tucker, W.G.J.F.P.T., 2000. An overview of PM2. 5 sources and control strategies*, 65*, 379-392

Tyson, P.D., Kruger, F.J., & Louw, C.W., 1988. Atmospheric pollution and its implications in the Eastern Transvaal highveld

uMoya-NILU, 2017. Review of the 2009 Gauteng Air Quality Management Plan, Air Quality Baseline Report, *Secondary Review of the 2009 Gauteng Air Quality Management Plan, Air Quality Baseline Report*

United States National Aeronautics and Space Administration, & the Ministry of Economy Trade and Industry of Japan, 2019. Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Model Version 3. In

Venter, A.D., Vakkari, V., Beukes, J.P., van Zyl, P.G., Laakso, H., Mabaso, D., Tiitta, P., Josipovic, M., Kulmala, M., Pienaar, J.J., & Laakso, L., 2012. An air quality assessment in the industrialised western Bushveld Igneous Complex, South Africa %J South African Journal of Science*, 108*, 1-10

Wichmann, J., Voyi, K.J.I.j.o.e.r., & health, p., 2012. Ambient air pollution exposure and respiratory, cardiovascular and cerebrovascular mortality in Cape Town, South Africa: 2001–2006*, 9*, 3978-4016

World Health Organization, 2006. Air Quality Guidelines, Global Update 2005, *Secondary Air Quality Guidelines, Global Update 2005*

World Health Organization, 2016. Ambient air pollution: A global assessment of exposure and burden of disease, *Secondary Ambient air pollution: A global assessment of exposure and burden of disease*

Wright, C., Garland, R., Thambiran, T., Forbes, P., Diab, R., & Oosthuizen, R., 2017. Air quality and human health impacts in southern Africa

Xulu, N.A., Piketh, S.J., Feig, G.T., Lack, D.A., & Garland, R.M., 2020. Characterizing Light-absorbing Aerosols in a Low-income Settlement in South Africa. *Aerosol and Air Quality Research, 20*, 1812-1832

Zulu, T., Aphane, O., Audat, T., & Olifant, V., 2019. The South Africa Energy Sector Report 2019. http://www.energy.gov.za/files/media/explained/2019-South-African-Energy-Sector-Report.pdf