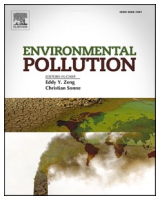
Environmental Pollution 310 (2022) 119883

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol



Ensemble averaging using remote sensing data to model spatiotemporal PM_{10} concentrations in sparsely monitored South Africa☆

Oluwaseyi Olalekan Arowosegbe [a,b], Martin Röösli [a,b], Nino Künzli [a,b], Apolline Saucy [a,b], Temitope C. Adebayo-Ojo [a,b], Joel Schwartz [c], Moses Kebalepile [d], Mohamed Fareed Jeebhay [e], Mohamed Aqiel Dalvie [e], Kees de Hoogh [a,b,*]

[a] Swiss Tropical and Public Health Institute, Allschwil, Switzerland
[b] University of Basel, Basel, Switzerland
[c] Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[d] Department for Education Innovation, University of Pretoria, Pretoria, South Africa
[e] Centre for Environmental and Occupational Health Research, School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa

ARTICLE INFO

Keywords:
Ensemble averaging
Particulate matter
Satellite observations
Machine learning

ABSTRACT

There is a paucity of air quality data in sub-Saharan African countries to inform science driven air quality management and epidemiological studies. We investigated the use of available remote-sensing aerosol optical depth (AOD) data to develop spatially and temporally resolved models to predict daily particulate matter (PM_{10}) concentrations across four provinces of South Africa (Gauteng, Mpumalanga, KwaZulu-Natal and Western Cape) for the year 2016 in a two-staged approach. In stage 1, a Random Forest (RF) model was used to impute Multiangle Implementation of Atmospheric Correction AOD data for days where it was missing. In stage 2, the machine learner algorithms RF, Gradient Boosting and Support Vector Regression were used to model the relationship between ground-monitored PM_{10} data, AOD and other spatial and temporal predictors. These were subsequently combined in an ensemble model to predict daily PM_{10} concentrations at 1 km × 1 km spatial resolution across the four provinces. An out-of-bag R^2 of 0.96 was achieved for the first stage model. The stage 2 cross-validated (CV) ensemble model captured 0.84 variability in ground-monitored PM_{10} with a spatial CV R^2 of 0.48 and temporal CV R^2 of 0.80. The stage 2 model indicated an optimal performance of the daily predictions when aggregated to monthly and annual means. Our results suggest that a combination of remote sensing data, chemical transport model estimates and other spatiotemporal predictors has the potential to improve air quality exposure data in South Africa's major industrial provinces. In particular, the use of a combined ensemble approach was found to be useful for this area with limited availability of air pollution ground monitoring data.

1. Introduction

Exposure to ambient air pollution is linked with several adverse health outcomes and is a major environmental risk factor associated with about 5 million deaths in 2019 (Murray et al., 2020). The World Health Organization (WHO) reported that 87% of the 3 million deaths estimated to be attributable to ambient air pollution in 2012 occurred in low and middle income countries (LMICs) (World Health Organization, 2016). Recently new findings from Northern America (Pinault et al., 2017; Shi et al., 2020) and Europe (Stafoggia et al., 2022; Strak et al., 2021) have provided evidence that adverse health effects are associated with air pollution even at levels less then national and international standards. This adds to the growing body of evidence supporting the revised 2021 WHO Air Quality Guidelines, where, for example, the guideline value for annual mean of particulate matter less than or equal to 10 μm in aerodynamic diameter (PM_{10}) was lowered from 20 $\mu g/m^3$ to 15 $\mu g/m^3$. In addition to higher emission of air pollutants in LMICs, barriers to an improved air quality in these regions include gaps in infrastructure, lack of data openness, unwillingness to share data to not hinder economic perspectives and capacity for air quality management (Mak and Lam, 2021; World Health Organization, 2021). The health impact of exposure to air pollution could be related to the current

☆ This paper has been recommended for acceptance by Admir Créso Targino.

* Corresponding author. Swiss Tropical and Public Health Institute, Allschwil, Switzerland.
E-mail address: c.dehoogh@swisstph.ch (K. de Hoogh).

https://doi.org/10.1016/j.envpol.2022.119883
Received 14 April 2022; Received in revised form 29 July 2022; Accepted 30 July 2022
Available online 3 August 2022

epidemiological transition of diseases from communicable diseases to non-communicable diseases in sub-Saharan African (SSA) countries (Adebayo-Ojo et al., 2022; Gouda et al., 2019; Koné et al., 2019). However, the limited number of monitoring sites for air pollutants in SSA countries has been a major challenge in investigating the association between exposure to air pollutants and adverse health outcomes (Amegah, 2018; Amegah and Agyei-Mensah, 2017). Air pollution monitoring sites in South Africa are sparsely distributed. These sites are mostly located in the designated air pollution priorities areas based on historical evidence of poor ambient air quality (Department of Environmental Affairs, 2016). These areas includes the Highveld, the Vaal triangle, the South Durban Basin and Waterberg areas located in four different provinces (Gauteng, Mpumalanga, Western Cape and KwaZulu-Natal) of South Africa (Arowosegbe et al., 2021b; Feig et al., 2019; Tshehla and Wright, 2019). There is also a substantial gap in historical and current air quality measurement data in South Africa due to inadequate technical and financial capacity to continuously operate these sites. Our previous work in South Africa used the most complete air pollutant monitoring data set, PM_{10}, to compare methods to impute missing daily PM_{10} concentrations across sites located in four provinces of Gauteng, Mpumalanga, Western Cape and KwaZulu-Natal (Arowosegbe et al., 2021b). In addition, large differences in PM_{10} concentrations exist between these provinces with monitoring sites in Gauteng exceeding the WHO air quality guideline 24-h PM_{10} concentration of 45 $\mu g/m^3$ 38% of the time between 2010 and 2017 compared to only 3% in Western Cape. Across the provinces, PM_{10} concentrations were highest in the winter months between June and August (Arowosegbe et al., 2021a).

Long- and short-term spatially varying air pollution data is important for air pollution mitigation strategies and epidemiological studies to protect the health of vulnerable populations. However, there are relatively few reference monitoring networks globally to capture the variation in air pollution around where people live and work (Martin et al., 2019). Consequently, a number of approaches including dispersion modeling, interpolation and land-use regression modeling have been used for long-term air pollutant exposure assessment in epidemiological studies (Bertazzon et al., 2015; Eeftens et al., 2012; Gulliver and Briggs, 2011; Wong et al., 2004). To better capture the spatial and temporal variation of air pollution required for epidemiological studies, hybrid statistical models have been implemented by several studies (de Hoogh et al., 2018; Mandal et al., 2020; Schneider et al., 2020; Stafoggia et al., 2019). Hybrid statistical models, for example, leverage the spatial and temporal coverage of satellite retrieved Aerosol Optical Depth (AOD) which quantifies the amount of light extinction by absorption or scattering that occurs in the column when light passes through suspended particles (Hoff and Christopher, 2009).

Recently, machine learning algorithms have been used to explore the relationship between ground-monitored air pollution data, AOD, spatial and temporal predictors (e.g. land use and meteorology). Machine learning algorithms are increasingly being used to model air pollution levels because of their ability to capture the underlying relationship between ground-monitored air pollution data and spatiotemporal predictors (de Hoogh et al., 2018; Mandal et al., 2020; Schneider et al., 2020; Sorek-Hamer et al., 2020; Stafoggia et al., 2019). Several variants of the hybrid statistical models have been implemented mostly in developed countries with good ground-monitored data to model long-term air pollution exposures (de Hoogh et al., 2016) and short-term air pollution exposures (de Hoogh et al., 2018; Lee et al., 2011; Stafoggia et al., 2019). Many previous air pollution modeling studies have either used single statistical models at different stages of their modeling approach or selected the best model out of several models to estimate air pollution concentrations (Bertazzon et al., 2015; Stafoggia et al., 2019; Stafoggia et al., 2020; Stafoggia et al., 2017). The application of machine algorithms to model PM_{10} concentration across South Africa presents an opportunity to assess the performance of this method in an area with limited ground-level monitoring data. Despite the flexibility and predictive performance of machine learning algorithms, these models are prone to overfitting especially when characterizing spatial predictors (Meyer et al., 2018). To improve the predictions from individual algorithms, ensemble averaging of different machine learning algorithms has been utilized in air pollution exposure modeling studies. Ensemble averaging takes advantage of the strengths of the individual machine learning algorithms to improve the accuracy of models predictions (Di et al., 2019; Mandal et al., 2020; Shtein et al., 2019).

Hybrid statistical models have been identified as a potential solution to bridge the gap in ground-monitored air pollution data in LMICs, especially in SSA countries (Pinder et al., 2019). In this study, we developed a hybrid statistical model based on ensemble averaging for predicting daily PM_{10} concentrations at a 1 km × 1 km spatial resolution across four provinces of South Africa for the year 2016. The year 2016 was selected as it was the year with the largest available number of PM_{10} monitoring sites operating in recent years (i.e. between 2010 and 2017 the respective number of sites were: 21, 41, 42, 40, 39, 32, 46 and 41 sites). The performance of hybrid statistical models is largely dependent on the availability of air pollution monitoring data used to calibrate the models. Consequently, this study aims to explore the possibility of using remote-sensing data in combination with other spatial and temporal predictors and monitoring data to predict daily PM_{10} concentrations at 1 km × 1 km spatial resolution across four provinces of South Africa.

2. Materials and methods

2.1. Study area

South Africa is located at the southernmost tip of Africa. The surface area is 1,219,912 km^2, with an estimated population of 58.8 million (2019) (Department of Statistics South Africa, 2019). South Africa has a long coastline that stretches more than 2500 km along the Atlantic and Indian oceans. Its coastal plain is dominated by a plateau surrounded by a great escarpment. The central and eastern part of the plateau is known as the Highveld, which is between 1500 and 2100 m above sea level. The highest edge of the escarpment is the Mpumalanga province (Drakensberg) in the east from where it then extends south-west to Free State and Gauteng Provinces. Gauteng province, the smallest province with a land area of 18,176 km^2, has the largest population of approximately 15 million (about 26% of the total South Africa population) and is bordered to the east by Mpumalanga. Mpumalanga is home to most of South Africa's coal factories and is bordered by KwaZulu-Natal to the south. The coastal province of Western Cape occupies a land area of 129,462 km^2. South Africa is characterized with distinct climatic conditions; the eastern part of the country has a tropical climate while the south-western part has a Mediterranean climate with year-round wind. These climatic features coupled with a mountainous escarpment influence the spatial and temporal pattern of air pollutants across the country. South Africa has four climatic seasons: Autumn (March–May), Winter (June–August), Spring (September–November) and Summer (December–February).

2.2. PM_{10} monitoring data

PM_{10} hourly data were collected from 46 monitoring sites jointly maintained by the Department of Environmental Affairs, South Weather Services, provincial, local governments and private industries. Of those 46 sites, 19 sites are located in Gauteng province, 16 sites in Mpumalanga, 7 sites in Western Cape and 4 sites in KwaZulu-Natal (Fig. 1). The data were obtained from the South African Air Quality Information System (https://saaqis.environment.gov.za/. Accessed on October 22, 2018). Data quality checks were undertaken for each monitoring station including removing outliers defined as negative values or observations greater or less than four times the interquartile range of each monitoring sites. Hourly PM_{10} data were aggregated to daily values if 75% of hourly data were valid. For this study, 2 Gauteng province sites, 9 Mpumalanga

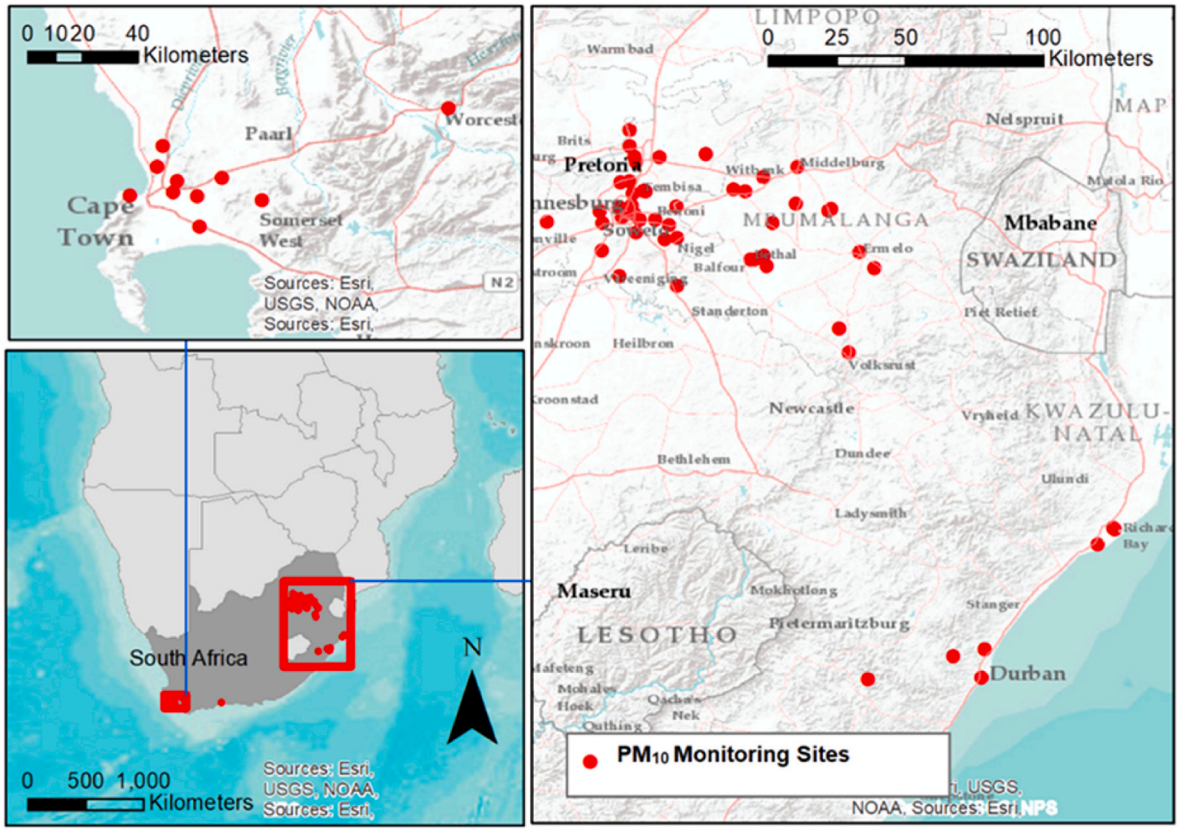


Fig. 1. The spatial distribution of PM_{10} Monitoring Sites.

sites, 4 Western Cape sites and 3 KwaZulu-Natal sites had at least 70% of annual PM_{10} data. Missing daily PM_{10} values for these sites were imputed as explained in our previous paper (Arowosegbe et al., 2021b). In brief, we imputed missing daily PM_{10} concentration by combining spatial and temporal predictors with ground-level monitored PM_{10} concentrations at sites with at least 70% of annual PM_{10} data in a Random Forest (RF) model. In contrast to the distribution of the predictions from National and Provincial RF models, the site-specific models PM_{10} predictions distribution were more comparable to the observed PM_{10} concentration distribution (Arowosegbe et al., 2021b). The final monitoring dataset used in this study included measured and imputed daily PM_{10} concentrations for a total of 18 sites.

2.3. Spatial and temporal predictors

Table 1 presents the data used as predictor variables in this study. All analysis were performed at a 1 km x 1 km-grid covering the entire study area, and each predictor variable was calculated to this spatial scale. Geospatial analyses were performed in ESRI ArcGIS 10. The next section describes the data in more detail.

2.3.1. Aerosol optical depth (AOD)

Aerosol Optical Depth (AOD) is a columnar integrated value that quantifies the amount of light absorbed or scattered by suspended particles as it passes through the atmosphere. AOD serves as an indicative measurement of particles in the column of the atmosphere at a given time. The Multi-Angle Implementation of Atmospheric Correction (MAIAC) product of AOD from the Moderate Resolution Imaging (MODIS) instrument on the Terra and Aqua satellites provides daily AOD estimates (Lyapustin et al., 2011). The MAIAC AOD product is provided at 1 km × 1 km spatial resolution (from https://lpdaac.usgs.gov/products/mcd19a2v006/. Accessed on October 20, 2018). The Terra and Aqua satellites travel across South Africa at a different time; Terra between 09:00 and 11:00 local time and Aqua between 13:00 and 15:00. Due to the two different measurement times, we combined daily AOD measurements of wavelength 470 nm from both the Aqua and Terra satellites. We used measurements from Aqua and combined it with AOD 470 nm measurements from Terra when Aqua AOD 470 nm measurements were missing. Data quality checks were performed to remove spurious measurements of AOD from cloud masking, values adjacent to cloud, high uncertainty flags and values within a 2.5th percentile moving window variance. The final MAIAC AOD data set for input in stage 1 for the year 2016 contained 62% of all possible observations.

2.3.2. Spatial and temporal predictors

Meteorological variables play an important role in the dispersion of air pollutants (De Visscher, 2013; Laña et al., 2016). We used daily global climate reanalysis of total precipitation, temperature, boundary layer height, vertical velocity, the component of the horizontal wind towards the east (U wind component) and the component of the horizontal wind towards north (V wind component) from the European Center for Medium-Range Weather Forecasts Reanalysis 5th Generation (ERA5) climate reanalysis dataset at a spatial resolution of 0.125° × 0.125° (approximately 10 km × 10 km) for the year 2016. We extracted Copernicus Atmosphere Monitoring Service (CAMS) Reanalysis daily columnar ensembles estimates of PM_{10}, nitrogen dioxide and ozone at a spatial resolution of 0.125° × 0.125° (approximately 10 km × 10 km) from the CAMS data store (https://ads.atmosphere.copernicus.eu./Accessed on October 30, 2018). Bilinear resampling was used for the spatially coarse meteorological and CAMS datasets (10 km × 10 km) to downscale to our 1 km x 1 km-grid using information from the four nearest grid cells values of these variables.

The spatial variables used for this study were calculated at a 1 km × 1 km grid covering the study area. The 2018 South Africa National Land cover dataset with 72 land use classes were reclassified into the five main categories (1) residential area, (2) industrial area, (3) built-up

Table 1
Description of spatial and temporal predictors.

Variable	Description	Source	Resolution
Population density	Mean population within 1 km × 1 km grid cell	SEDAC	~1 km
Land cover	South Africa National Land Cover 2018 densities (summary of meters within the grid cells by land cover categories of Natural, Built-up, Residential, Agricultural, Industrial)	South Africa Department of Environmental Affairs.	20 m
Light at night	1 km × 1 km Intersected aggregate	VIIRS-DNB	750 m
Impervious surface	1 km × 1 km Intersected aggregate after removing no data, clouds, shadows data	NOAA	30 m
Elevation	1 km × 1 km intersected aggregate of mean elevation	SRTM Digital Elevation Database	90 m
Roads	Summary of road length distance to nearest road type: major roads and other roads	OpenStreetMap	Lines
Climate zones	Cold interior, Temperate interior, Hot interior, Temperate coastal, Subtropical coastal, Arid interior	South Africa Bureau of Standards 2005	6 Zones
Copernicus Atmosphere Monitoring Service (CAMS) ensemble estimates of AOD	Daily CAMS ensemble estimates of AOD bilinear resampled at 1 km × 1 km	Copernicus Atmosphere Monitoring Service (CAMS)	10 km × 10 km
Meteorological variables (daily modelled planetary boundary layer height, temperature, precipitation, wind speed, wind direction, relative humidity, vertical velocity)	Daily global ECMWF re-analysis estimates bilinear resampled at 1 km × 1 km	ERA5-reanalysis	10 km × 10 km
Modelled Tropospheric estimates of NO_2, PM_{10}, O_3	Daily Chemical transport model estimate bilinear resampled at 1 km × 1 km	Chemical transport model Copernicus Atmosphere Monitoring Service (CAMS)	10 km × 10 km

Abbreviations: SEDAC (Socioeconomic Data and Applications Center), VIIRS-DNB(Visible Infrared Imaging Radiometer Suite-Day/Night Band), NOAA(National Oceanic and Atmospheric Administration, SRTM (Shuttle Radar Topography Mission), ERA-5 (European Center for Medium-Range Weather Forecasts Reanalysis 5th Generation(Hersbach et al., 2020)).

area, (4) water bodies and (5) agricultural area. Sum of major road and sum of other road length was calculated for each 1 km x 1 km-grid cell using road data extracted from OpenStreetMap. Similarly, population density at each grid cell was calculated based on the data extracted from the Socioeconomic Data and Application Center (SEDAC). Other spatial variables such as the light at night were extracted from Visible Infrared Imaging Radiometer Suite-Day/Night Band (VIIRS-DNB) and averaged at the 1 km × 1 km spatial resolution. Impervious surface and elevation data were respectively obtained from the National Oceanic and Atmospheric Administration and the Shuttle Radar Topography Mission Digital Elevation databases.

2.4. Statistical methods

We implemented a multi-stage machine learning modeling approach aimed at 1) imputing missing MAIAC AOD data using modelled estimates of CAMS AOD and 2) modeling the ground-monitored PM_{10} with AOD data, meteorological predictors, land use and land cover predictors. The calibrated model was then used to predict daily PM_{10} concentration at 1 km × 1 km grid cells over the four provinces of South Africa. In this study, we applied three machine learning algorithms at different stages of the analysis (Fig. S1).

2.5. Stage 1

We developed a model to impute missing MAIAC AOD data. The percentage of missing satellite-AOD measurements in South Africa, mainly caused by cloud cover, was 38%for the year 2016. We explored the statistical relationship between MAIAC AOD 0.47 μm wavelength, modelled co-located CAMS AOD estimates (469 nm, 550 nm, 670 nm, 865 nm and 1240 nm) day of the year, latitude and longitude using an optimized RF model:

$$PredMAIAC.AOD_{i,\,t} = MAIAC.AOD_{i,\,t} \sim f(CAMS.AOD_{i,\,t,z1-5} + day\ of\ the\ year + latitude_i + longitude_i) \quad (1)$$

where $PredMAIAC.AOD_{i,t}$ is the predicted MAIAC AOD 0.47 μm at grid cell *i*, on day *t*; MAIAC. $AOD_{i,t}$ is the target variable representing MAIAC AOD 0.47 μm wavelength estimates at grid *i* on day *t*; CAMS.AOD estimates the main predictor at grid cell *i*, on day *t*, at five wavelengths (z = 0.47 μm, 0.55 μm, 0.67 μm, 0.87 μm and 1.24 μm); day_of_the_year from 1 to 366; $latitude_i$ and $longitude_i$ represent the coordinates of grid cell centroid *i*.

2.6. Stage 2

A predictive model for daily PM_{10} concentrations was constructed by exploring its relationship with spatial and temporal predictors and AOD estimates from stage 1. We used an ensemble averaging approach using three different machine learning learners. The learners were RF (Breiman, 2001; Kwok and Carter, 1990), support vector regression (SVR) (Vapnik, 1999; Vapnik et al., 1997) and extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016). We selected tree based learners (RF and XGBoost) and SVR to account for complex non-linear relationship and patterns in explaining the variation in PM_{10} concentrations across the four South African provinces. We also implemented ensemble averaging of the predictions from the individual learners using a RF models that included the longitude and latitude of the 1 km × 1 km-grid cells to prevent the overfitting of the individual models. All the individual models were trained on the training data and optimized models were achieved through grid search, learners' internal parameter tuning and cross-validation processes. The RF parameter tuning includes grid search for the number of variables used to split each tree (mtry). Random variables of 2, 4, 6, 8, 10 and 12 were assessed in the grid search. We also searched for the number of random trees from 100 to 500 trees for an optimized model. The XGBoost model parameters grid space of maximum tree depth ranged from 4 to 14, maximum child weight from 2 to 10 and the subsample ratio from 0.4 to 0.9 were assessed to select the optimized model. The sigma and gamma values of the SVM were also selected based on grid search.

The individual learners were defined as:

$$YY\_PredPM_{10i,t} = PM_{10i,\,t} \sim f(SPT_{1i,t,\,\ldots\,,} SPT_{10i,t,\ldots,} SP_{20i,\ldots,} SP_{24i}) \quad (2)$$

where YY_ $PredPM_{10i,t}$ stands for RF, XGBoost or SVR, $PredPM_{10i,t}$ is the predicted PM_{10} at grid cell i, on day t; $PM_{10i,t}$ is the ground-monitored PM_{10} at the monitoring site in grid cell i on day t. $SPT1\text{-}10_{i,t}$ are spatio-temporal predictor variables numbering 1–10 in grid cell i and on day t and SP_i represent spatial predictor variables numbering between 20 and 24 in grid cell i on day t.

The RF averaging meta-model was defined as:

$$PredPM_{10i,t} = PM_{10i,t} \sim f(RF_{predPM_{10i,t}} + XGBoost_{predPM_{10i,t}} + SVR_{predPM_{10i,t}} + latitude_i + longitude_i) \quad (3)$$

where $PredPM_{10i,t}$ is the ensemble averaged predicted PM_{10} at grid cell i, on day t; $PM_{10i,t}$ is the ground-monitored PM_{10} at the monitoring site in grid cell i on day t, while $RF\_predPM_{10i,t}$, $XGBoost\_predPM_{10i,t}$ and $SVR\_predPM_{10i,t}$ are the predicted PM_{10} concentrations in grid cell i on day t from RF, XGBoost and SVR respectively. The $latitude_i$ and $longitude_i$ represent the coordinates of grid cell centroid i.

We included latitude and longitude as additional predictors to the individual learners predictions to allow the RF meta-model to capture and account for the variation in the performance of the individual learners in space. If one learner does better in Gauteng province but another in Western Cape province, the RF meta-model will capture the underlying interaction, thus, allowing some level of weighting when averaging the predictions of the individual learners. The final averaged ensemble model was used to predict daily PM_{10} concentrations across the four provinces at 1 km × 1 km. All statistical analyses were implemented in R open source programming software using the Caret package, version 4 (R Core Team (2018)).

2.7. Statistical performance

We evaluated the performance of the Stage 1 RF model by assessing the relationship between observed AOD and predicted AOD estimates in the two-third training dataset and the one-third out-of-bag (OOB) sample. The percentage of variation of AOD captured by the RF model, the R squared (R^2), the root mean squared prediction error (RMSPE), the intercept and the slope of the linear regression between the observed and predicted AOD were computed as the performance metrics.

For Stage 2 models, a ten-fold cross validation was conducted by building the model on 90% of the PM_{10} data and assessing the ensemble model prediction on the hold out 10% PM_{10} data. Spatial performance was assessed through leave-location-out cross-validation (LLO CV). Site ID was used as the splitting criterion and the models were divided into ten folds to compute the models spatial performance. A model was trained on data from all but one-fold of sites (n−1). The hold-out folds were iteratively used to estimate the prediction errors of these models to predict for sites not included in the training folds dataset. For temporal cross-validation, day of the year was used to divide the dataset into 10 folds and temporal leave-time-out cross-validation (LTO CV) was used to assess the model's performance in time.

3. Results

3.1. Stage 1 imputation of AOD data

The stage 1 model performance was evaluated by comparing MAIAC AOD observations and model predictions in the OOB samples. The estimated percentage of variability (R^2) captured by the RF model in the OOB samples was 0.96 (RMSPE = 0.014, intercept = −0.001, slope = 1.01). The stage 1 model metrics suggest a good fit between the valid observed and the predicted AOD 470 nm. Fig. 2 shows a map of predicted AOD 470 nm for June 6, 2016 for Gauteng province. Example AOD prediction maps for the other three provinces are presented in (S2 – S4). The spatial coverage of valid MAIAC AOD values in South Africa in 2016 ranged from 43% in July to 80% in December (Table S1). The distribution of the valid MAIAC AOD data was not markedly different across the months. However, the month of September recorded the highest values of AOD (mean of 0.15).

3.2. Stage 2 calibrating PM_{10} with AOD and spatial-temporal data

Fig. 3 shows scatter plots between predicted and observed PM_{10} concentrations of the spatial, temporal and overall cross validation of the ensemble model. The overall R^2 of 0.81 suggest good correlation between ground-level PM_{10} and ensemble model PM_{10} predictions. The ensemble performed well temporally (R^2 of 0.80) but less so spatially (R^2 of 0.48). The cross-validated performance metrics of the individual models compared to the ensemble model is presented in Table 2. Of the

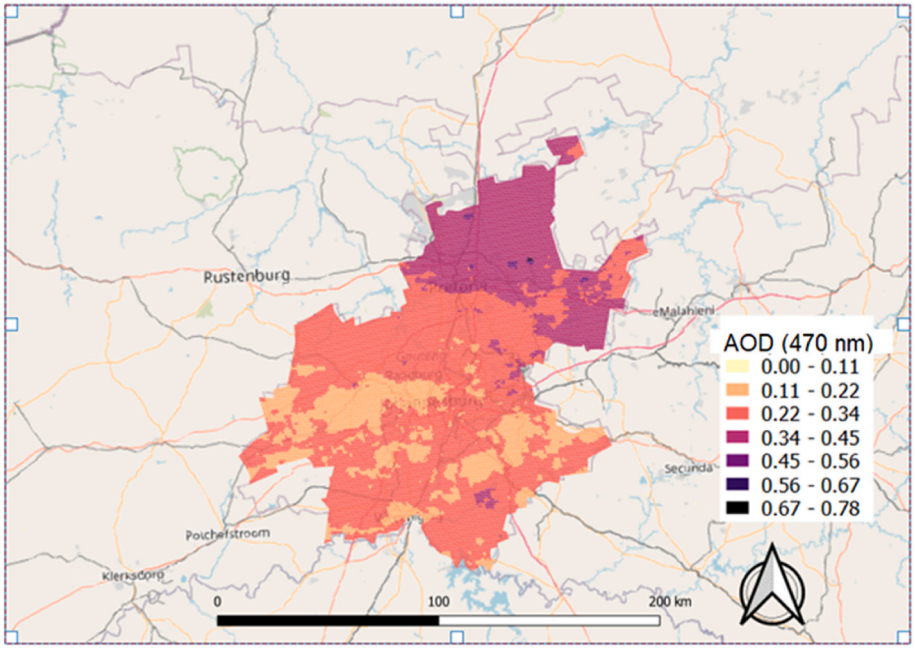


Fig. 2. Gauteng prediction map of AOD 470 nm for June 6, 2016.

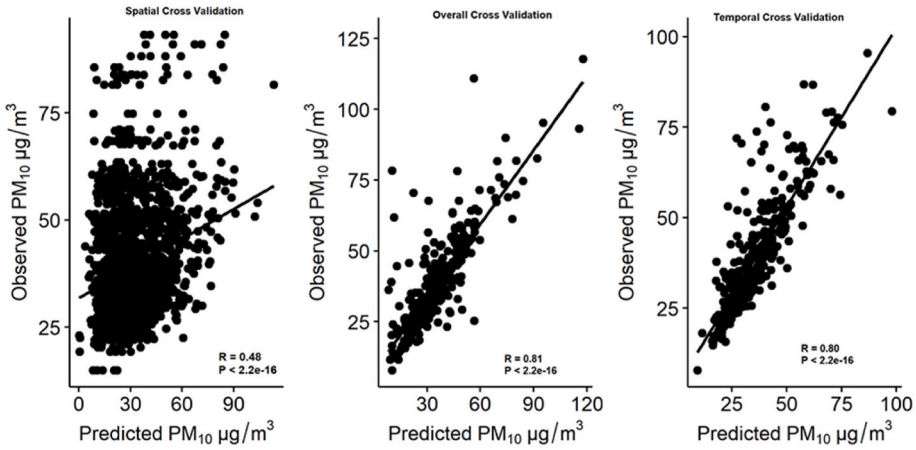


Fig. 3. Scatter plots between predicted and observed PM_{10} concentrations of the spatial, temporal and overall cross validation of the ensemble model.

Table 2
Cross-validated Performance Measures of the Different Stage 2 Models for 2016: R^2 (percent of explained variability). Overall root mean squared error (RMSE in $\mu g/m^3$), spatial and temporal R^2 and RMSE are reported for the ensemble averaged model.

Model	CV	R^2	RSME
Ensemble	Total	0.81	11.4
	Spatial	0.48	20.5
	Temporal	0.80	12.3
RF	Total	0.79	12.0
	Spatial	0.34	23.3
	Temporal	0.78	12.9
XGBOOST	Total	0.81	11.4
	Spatial	0.36	23.9
	Temporal	0.78	12.7
SVR	Total	0.77	12.6
	Spatial	0.14	31.0
	Temporal	0.76	12.3

three machine learning algorithms, the model performance of the XGBoost marginally outperformed RF and SVR. In principle, XGBoost sequentially optimizes weak trees to improve their performance. This might explain the better performance of our XGBoost model. The ensemble model monthly mean PM_{10} predictions follow the observed monthly mean PM_{10} temporal trends across the four provinces (Fig. 4). Fig. 5 shows the annual mean PM_{10} concentrations estimated at 1 km × 1 km resolution for the four provinces. The spatial distribution of the annual PM_{10} concentrations highlights highly populated and industrialized areas of Gauteng province. Our models identified Johannesburg, Soweto and areas around the Vaal Triangle as PM_{10} pollution hotspots in Gauteng province. Similarly, the Highveld areas of Secunda, Middelburg, Kriel, eMalahleni and Hendrina emerged as PM_{10} pollution hotspots in Mpumalanga province. The cities of Cape Town and Durban are highlighted as PM_{10} pollution hotspots in Western Cape and KwaZulu-Natal provinces respectively. The predicted concentrations of PM_{10} in Western Cape and KwaZulu-Natal provinces were lower compared to those in Gauteng and Mpumalanga provinces (Fig. S8). To illustrate the monthly variation in predicted PM_{10} concentrations, Fig. 6 shows seasonal patterns in the monthly mean PM_{10} concentrations for Gauteng province (see Supplementary Figs. S5–S7 for the monthly mean maps of Mpumalanga, KwaZulu-Natal and Western Cape Provinces). PM_{10} concentrations were highest during the winter months from June to September, peaking in September. The percentage improvement of the models for each variables included in the Stage 2 models are ranked in Fig. 7. The relative importance of each predictor quantifies the amount of error reduced when used by the models. For ease of interpretation, the importance score of each predictor was standardized from 0 to 100% by dividing each predictor importance score by the highest importance score of the predictors and multiply by 100 using R package Caret. Fig. 7 shows that the most important predictor was relative humidity, closely followed by CAMS_PM10.

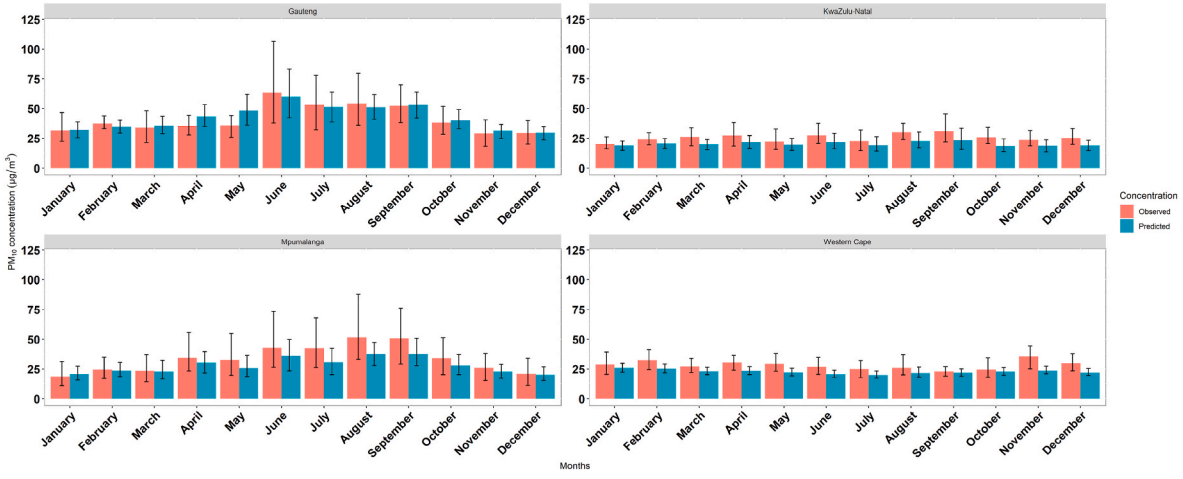


Fig. 4. Monthly observed versus predicted PM_{10} means. Error bars represent standard deviation of the mean.

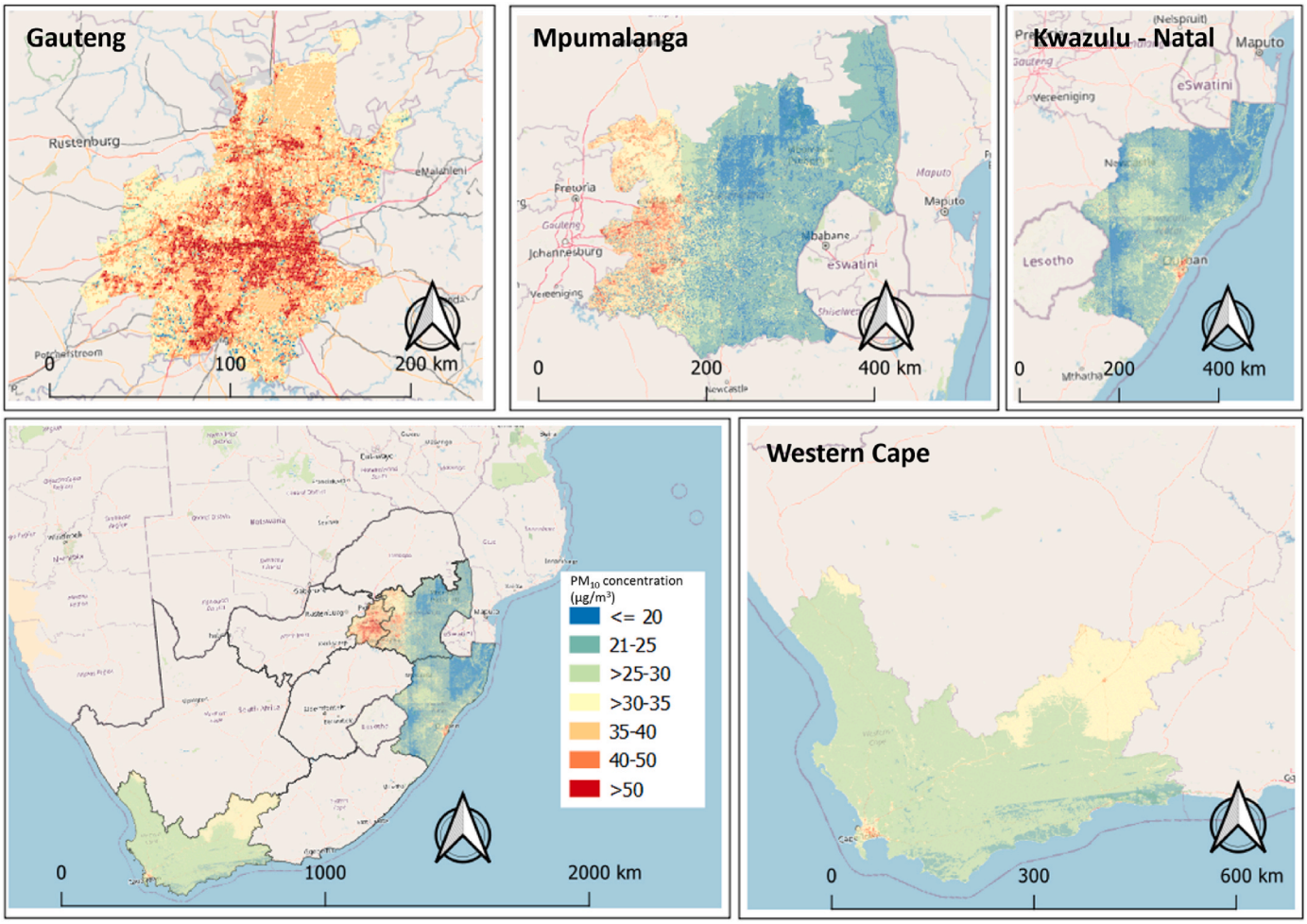


Fig. 5. Annual mean PM_{10} concentrations (µg/m^3) for 2016 at 1 km × 1 km grid cells aggregated from daily estimates.

4. Discussion

The application of AOD data to explain the variation in ground-monitored air pollutant has been explored in different countries because of its spatial coverage. In this study, about 38% of all possible AOD data were missing in 2016. The proportion of valid AOD data was high when compared to studies from the Northern Hemisphere. A Swiss study reported 80.2% missing AOD observation in Switzerland from 2003 to 2013 while a range of 67%–83% missing AOD observations was observed in Italy during the study period of 2013–2015 (de Hoogh et al., 2018; Stafoggia et al., 2019). The higher number of valid observations reported in this study was achieved due to the combination of the Aqua and Terra AOD products and favorable meteorological conditions in South Africa with fewer days, on average, with cloud cover in South Africa compared to Europe. The performance of the model used to impute missing AOD data suggested the model was able to capture about 96% variability in AOD with negligible error metrics. Our result is consistent with studies that have employed a similar approach in Great Britain and Italy with 98% and >94% percentage of variability in AOD captured respectively (Schneider et al., 2020; Stafoggia et al., 2019). The maximum value of AOD was recorded in September of 2016 in South Africa. This is comparable with results from a South African study on the regional and local characteristics of aerosols that also observed maximum values of AOD between August and October from 2000 to 2009 (Hersey et al., 2015). The high values of AOD reported during this period have been linked to the burning season in South Africa's neighboring countries of Mozambique and Zimbabwe. Both countries have been identified as the major source of aerosols transported to South Africa. In addition, August and October also coincide with increased windblown dust across South Africa (Hersey et al., 2015).

The missing 38% of AOD data, although a low percentage compared with other study regions, is not random, with the largest fraction of missing AOD data observed in the winter (June to August). The winter also coincides with the highest observed PM_{10} concentrations in the ground-level measurements due to increased use of fossil fuels e.g. for heating purpose (Hersey et al., 2015). This could potentially lead to bias in the predicted PM_{10} concentrations either over- or under-predicting. However, we also offered CAMS predicted PM_{10} which was higher ranked in the relative importance compared to AOD 470 nm (Fig. 7), which would have reduced the likelihood of potential bias in our estimates.

Recently, the application of ensemble models has become more prominent (Di et al., 2019; Mandal et al., 2020; Shtein et al., 2019). The argument for the ensemble modeling approach is that by combining individual model estimates the individual biases of the different statistical models can be reduced. In this study we applied an ensemble approach using a generalized linear model to combine three models; RF, XGBoost and SVR. The overall CV R^2 of 0.81 of the ensemble model was within the range of 0.71–0.81 reported by the two Italian studies for the years 2006–2012 and years 2013–2015, and substantially higher than the R^2 of 0.64 reported in Sweden (Shtein et al., 2019; Stafoggia et al., 2020; Stafoggia et al., 2017). Like the suboptimal performance of our model (spatial R^2 of 0.48 in hold-out sites), the model fit (total R^2) of the Swedish study reduced to 0.50 in hold-out sites. The strong decrease in our model performance in space is possibly due to the limited number and the uneven distribution of the monitoring sites. The monitoring sites

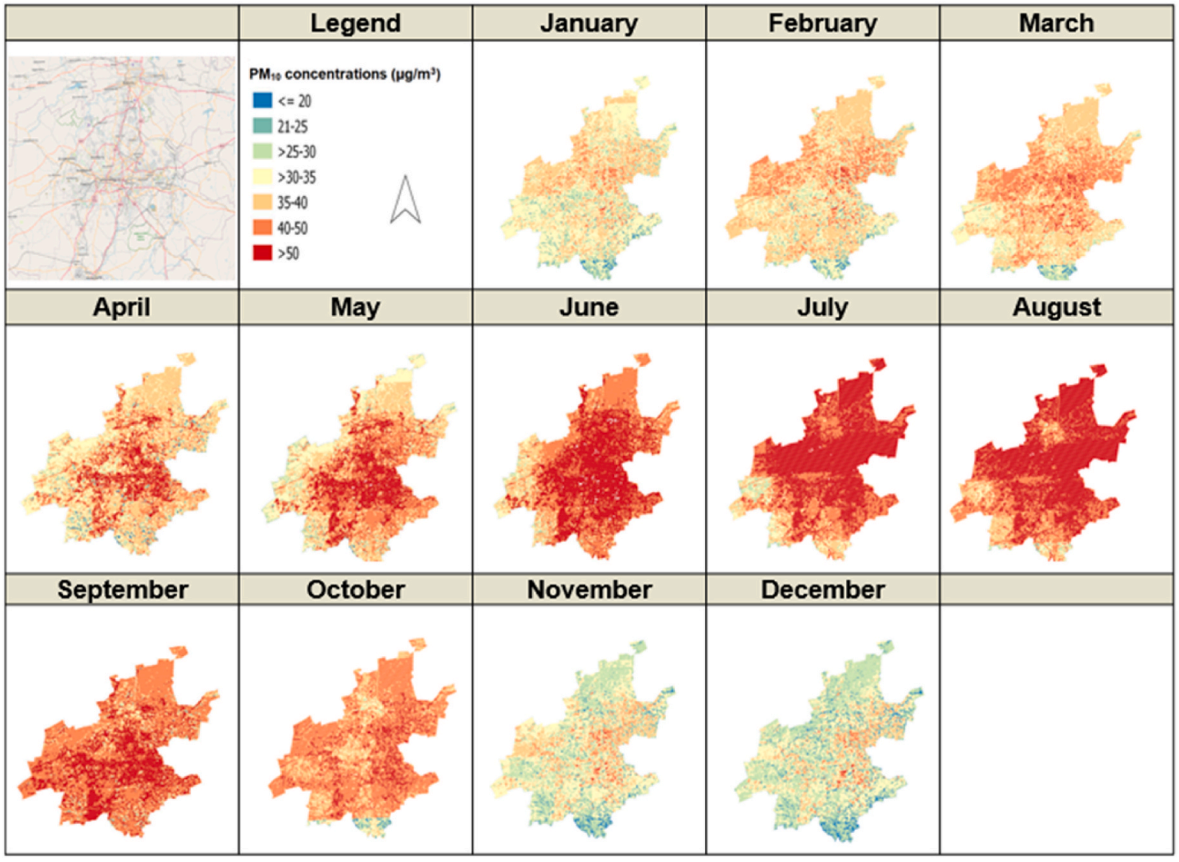


Fig. 6. Gauteng Province estimated monthly mean PM_{10} concentrations (µg/m^3) for 2016 at 1 km × 1 km grid cells aggregated from daily estimates.

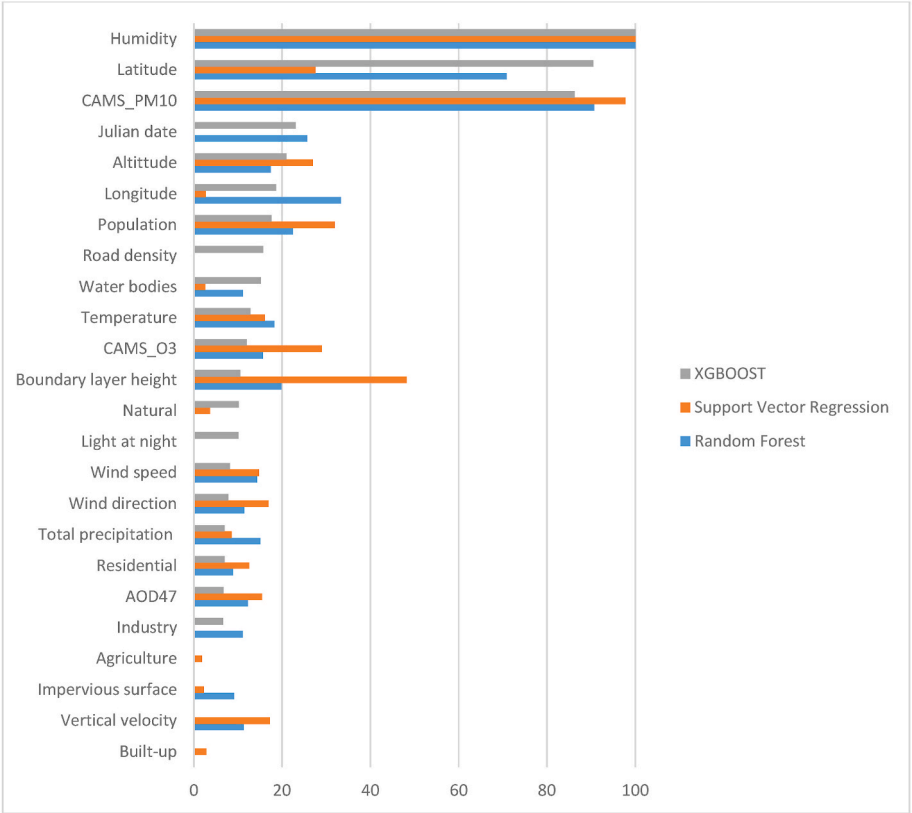


Fig. 7. Relative importance (scale from 0 to 100%) of the top 20 predictors from the individual models in Stage 2.

are located in high pollution priority areas and this might not be sufficient to capture the variation in PM_{10} concentrations beyond the spatial domains of the monitoring sites. In addition, due to lack of availability we may have missed important predictor variables to characterize PM_{10} concentrations in South Africa, for example, detailed emission data. Despite this, the geographical variation of the estimated PM_{10} concentrations aligns with the spatial pattern of PM_{10} concentrations presented in our previous study on the spatial and temporal characteristics of PM_{10} data. The potential to use AOD data to explain the variation of air pollution at ground level is dependent on its relationship with ground-monitored measurements. In this study, AOD did not emerge as a strong variable for explaining the variation in PM_{10} concentrations in South Africa. Hersey et al., (2015) also reported a poor correlation between $PM_{2.5}$ and PM_{10} and AOD in South Africa. The persistent and frequent dilution of South Africa's vertical column with plumes from biomass burning emissions from the tropics at stable layers between 3 and 5 km above the majority of South Africa has been posited for the poor correlation between AOD and ground-level PM (Campbell et al., 2003; Chand et al., 2009; Hersey et al., 2015; Tyson et al., 1996). Another reason is the likely inability of the satellite retrievals to differentiate between ground surface aerosol and concentrated aerosol layers from emissions released to the shallow boundary layer, related to geographical features, during the winter season in South Africa. Lastly, particulate matter concentrations in South Africa are influenced by morning and evening air pollution peak times. These peak times do not correspond to the different overpass times of the satellites in South Africa (Hersey et al., 2015).

Nonetheless, in the four provinces included in this study, the areas around the economic and industrial cities of these provinces recorded the highest PM_{10} concentrations estimates. The estimated annual mean PM_{10} concentration maps of the four provinces also suggest that concentrations in large parts of the Gauteng province are higher than WHO annual PM_{10} guideline of 15 $\mu g/m^3$ (World Health Organization, 2021). This is not surprising given that the Gauteng conurbation is the most densely populated province in South Africa with the highest density of anthropogenic emissions from all sources. Furthermore, we previously reported higher levels of PM_{10} concentrations in Gauteng province monitoring stations compared to the other three provinces (Arowosegbe et al., 2021a). A similar pattern was also reported for $PM_{2.5}$ by Zhang and colleagues (Zhang et al., 2021) showing high modelled $PM_{2.5}$ concentrations in Northern and Southern Gauteng of the Highveld region of South Africa. The models identified the PM_{10} pollution hotspots around the mining activities of Mpumalanga province, Southern Durban Basin industrial Basin of KwaZulu-Natal and Cape Town Metropolitan of Western Cape province. To demonstrate the seasonal pattern captured by our models, we found an increase in PM_{10} concentrations between May and September. This overlaps with the winter months when there is an increase in anthropogenic emissions due to increased use of coal for domestic and industrial purposes and the formation of surface inversion layers preventing the atmospheric mixing mechanism for the dispersion of pollutants (Hersey et al., 2015).

The ensemble approach used in this study performed well in characterizing PM_{10} concentrations across the four selected provinces of South Africa. However, we acknowledge the limited number of monitoring stations and ground-monitored PM_{10} data to calibrate these models. In addition, the distribution of the sparse monitoring stations impacted the stability of the models. The availability of emission data could have improved the performance of our models.

5. Conclusions

High quality air pollution exposure data to support health studies is lacking in many LMICs. With sparse air pollution monitoring data, we have shown - for the first time - that is possible to estimate daily PM_{10} concentrations for a whole year across four provinces of South Africa by leveraging remote sensing and novel spatiotemporal modeling approaches. Our spatiotemporal model was successful in capturing the day to day temporal variation, but was less efficient in characterizing the spatial contrast of PM_{10}. In particular, the chemical transport model variable, CAMS PM_{10}, was a highly influential predictor, and in our case more important than the satellite–derived variable MAIAC AOD. These variables should be considered as crucial predictors when modeling air pollution concentration in areas with limited ground monitoring networks. The potential of spatiotemporal models presented here, however, remains largely dependent on good air quality monitoring data as demonstrated by our study results. Therefore, efforts to improve air quality monitoring in SSA and other LMICs should be encouraged and supported to enable derivation of exposure data in these challenging settings.

Author statement

Oluwaseyi Olalekan Arowosegbea: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft preparation. **Martin Röösli**: Conceptualization, Supervision, Writing – review & editing, Methodology. **Nino Künzli**: Writing – review & editing. **Apolline Saucy**: Writing – review & editing, Methodology. **Temitope C Adebayo-Ojo**: Writing – review & editing. **Joel Schwartz**: Writing – review & editing, Methodology. **Moses Kebalepile**: Writing – review & editing. **Mohamed Fareed Jeebhay**: Writing – review & editing. **Mohamed Aqiel Dalvie**: Supervision, Writing – review & editing. **Kees de Hoogh**: Conceptualization, Supervision, Writing – review & editing, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This study is part of the Joint South Africa and Swiss Chair in Global Environmental Health

(SARChI), funded by the South African National Research Foundation (grant number 94883) and the Swiss State Secretariat for Education, Research, and Innovation. O.O.A. is a recipient of a Swiss Government Excellence Scholarship. The authors would like to thank the staff of the different monitoring stations and the staff at South Weather Services who provided data for this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2022.119883.

References

Adebayo-Ojo, T.C., Wichmann, J., Arowosegbe, O.O., Probst-Hensch, N., Schindler, C., Künzli, N., 2022. Short-term Joint effects of PM10, NO2 and SO2 on cardio-respiratory disease hospital admissions in Cape Town, South Africa. Int. J. Environ. Res. Publ. Health 19, 495.

Amegah, A.K., 2018. Proliferation of low-cost sensors. What prospects for air pollution epidemiologic research in Sub-Saharan Africa? Environ. Pollut. 241, 1132–1137.

Amegah, A.K., Agyei-Mensah, S., 2017. Urban air pollution in sub-saharan Africa: time for action. Environ. Pollut. 220, 738–743.

Arowosegbe, O.O., Röösli, M., Adebayo-Ojo, T.C., Dalvie, M.A., de Hoogh, K., 2021a. Spatial and temporal variations in PM10 concentrations between 2010–2017 in South Africa. Int. J. Environ. Res. Publ. Health 18, 13348.

Arowosegbe, O.O., Röösli, M., Künzli, N., Saucy, A., Adebayo-Ojo, T.C., Jeebhay, M.F., Dalvie, M.A., de Hoogh, K., 2021b. Comparing methods to impute missing daily

ground-level PM10 concentrations between 2010–2017 in South Africa. Int. J. Environ. Res. Publ. Health 18, 3374.
Bertazzon, S., Johnson, M., Eccles, K., Kaplan, G.G., 2015. Accounting for spatial effects in land use regression for urban air pollution modeling. Spatial and spatio-temporal epidemiology 14, 9–21.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Campbell, J.R., Welton, E.J., Spinhirne, J.D., Ji, Q., Tsay, S.C., Piketh, S.J., Barenbrug, M., Holben, B.N., 2003. Micropulse lidar observations of tropospheric aerosols over northeastern South Africa during the ARREX and SAFARI 2000 dry season experiments. J. Geophys. Res. Atmos. 108.
Chand, D., Wood, R., Anderson, T., Satheesh, S., Charlson, R., 2009. Satellite-derived direct radiative effect of aerosols dependent on cloud cover. Nat. Geosci. 2, 181–184.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
R Core Team, 2018. 2018 R: A language and environment for statistical computing, 2018. In: R Foundation for Statistical Computing. R.C.T., Vienna, Austria. Austria.
de Hoogh, K., Gulliver, J., van Donkelaar, A., Martin, R.V., Marshall, J.D., Bechle, M.J., Cesaroni, G., Pradas, M.C., Dedele, A., Eeftens, M., 2016. Development of West-European PM2. 5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. Environ. Res. 151, 1–10.
de Hoogh, K., Héritier, H., Stafoggia, M., Künzli, N., Kloog, I., 2018. Modelling daily PM2. 5 concentrations at high spatio-temporal resolution across Switzerland. Environ. Pollut. 233, 1147–1154.
De Visscher, A., 2013. Air Dispersion Modeling: Foundations and Applications. John Wiley & Sons.
Department of Environmental Affairs, 2016. In: Affairs, D.o.E. (Ed.), 2nd South Africa Environment Outlook (Pretoria).
Department of Statistics South Africa, 2019. P0302 - Mid-year Population Estimates. South Africa (This statistical release contains estimations of the population of South Africa and describes the methods used to compile these estimations).
Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M.B., Choirat, C., Koutrakis, P., Lyapustin, A.J.E.i., 2019. An Ensemble-Based Model of PM2. 5 Concentration across the Contiguous United States with High Spatiotemporal Resolution, vol. 130, 104909.
Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dėdelė, A., Dons, E., de Nazelle, A., 2012. Development of land use regression models for PM2. 5, PM2. 5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46, 11195–11205.
Feig, G., Garland, R.M., Naidoo, S., Maluleke, A., Marna, V.d.M., 2019. Assessment of changes in concentrations of selected criteria pollutants in the Vaal and Highveld priority areas. Clean Air J. 29.
Gouda, H.N., Charlson, F., Sorsdahl, K., Ahmadzada, S., Ferrari, A.J., Erskine, H., Leung, J., Santamauro, D., Lund, C., Aminde, L.N., 2019. Burden of non-communicable diseases in sub-saharan Africa, 1990–2017: results from the global burden of disease study 2017. Lancet Global Health 7, e1375–e1387.
Gulliver, J., Briggs, D., 2011. STEMS-Air: a simple GIS-based air pollution dispersion model for city-wide exposure assessment. Sci. Total Environ. 409, 2419–2429.
Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049.
Hersey, S.P., Garland, R.M., Crosbie, E., Shingler, T., Sorooshian, A., Piketh, S., Burger, R., 2015. An overview of regional and local characteristics of aerosols in South Africa using satellite, ground, and modeling data. Atmos. Chem. Phys. 15, 4259–4278.
Hoff, R.M., Christopher, S.A., 2009. Remote sensing of particulate pollution from space: have we reached the promised land? J. Air Waste Manag. Assoc. 59, 645–675.
Koné, B., Youssef Oulhote, A.M., Olaniyan, T., Kouame, K., Benmarhnia, T., Munyinda, N., Basu, N., Fobil, J.N., Etajak, S., Annesi-Maesano, I., 2019. Environmental health research challenges in Africa: insights from symposia organized by the ISEE Africa Chapter at ISES-ISEE 2018. Environmental Epidemiology 3.
Kwok, S.W., Carter, C., 1990. Multiple Decision Trees, Machine Intelligence and Pattern Recognition. Elsevier, pp. 327–335.
Laña, I., Del Ser, J., Padró, A., Vélez, M., Casanova-Mateo, C., 2016. The role of local urban traffic and meteorological conditions in air pollution: a data-based case study in Madrid, Spain. Atmos. Environ. 145, 424–438.
Lee, H., Liu, Y., Coull, B., Schwartz, J., Koutrakis, P., 2011. A novel calibration approach of MODIS AOD data to predict PM 2.5 concentrations. Atmos. Chem. Phys. 11, 7991–8002.
Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., Levy, R., Reid, J., 2011. Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. J. Geophys. Res. Atmos. 116.
Mak, H.W.L., Lam, Y.F., 2021. Comparative assessments and insights of data openness of 50 smart cities in air quality aspects. Sustain. Cities Soc. 69, 102868.
Mandal, S., Madhipatla, K.K., Guttikunda, S., Kloog, I., Prabhakaran, D., Schwartz, J.D., Team, G.H.I., 2020. Ensemble averaging based assessment of spatiotemporal variations in ambient PM2. 5 concentrations over Delhi, India, during 2010–2016. Atmos. Environ. 224, 117309.
Martin, R.V., Brauer, M., van Donkelaar, A., Shaddick, G., Narain, U., Dey, S., 2019. No one knows which city has the highest concentration of fine particulate matter. Atmos. Environ. X 3, 100040.
Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model. Software 101, 1–9.
Murray, C.J., Aravkin, A.Y., Zheng, P., Abbafati, C., Abbas, K.M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., 2020. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 396, 1223–1249.
Pinault, L.L., Weichenthal, S., Crouse, D.L., Brauer, M., Erickson, A., Donkelaar, A.v., Martin, R.V., Hystad, P., Chen, H., Finès, P., Brook, J.R., Tjepkema, M., Burnett, R.T., 2017. Associations between fine particulate matter and mortality in the 2001 Canadian census health and environment cohort. Environ. Res. 159, 406–415.
Pinder, R.W., Klopp, J.M., Kleiman, G., Hagler, G.S., Awe, Y., Terry, S., 2019. Opportunities and challenges for filling the air quality data gap in low-and middle-income countries. Atmos. Environ. 215, 116794.
Schneider, R., Vicedo-Cabrera, A.M., Sera, F., Masselot, P., Stafoggia, M., de Hoogh, K., Kloog, I., Reis, S., Vieno, M., Gasparrini, A., 2020. A satellite-based spatio-temporal machine learning model to reconstruct daily PM2. 5 concentrations across Great Britain. Rem. Sens. 12, 3803.
Shi, L., Wu, X., Danesh Yazdi, M., Braun, D., Abu Awad, Y., Wei, Y., Liu, P., Di, Q., Wang, Y., Schwartz, J., Dominici, F., Kioumourtzoglou, M.-A., Zanobetti, A., 2020. Long-term effects of PM2·5 on neurological disorders in the American Medicare population: a longitudinal cohort study. Lancet Planet. Health 4, e557–e565.
Shtein, A., Kloog, I., Schwartz, J., Silibello, C., Michelozzi, P., Gariazzo, C., Viegi, G., Forastiere, F., Karnieli, A., Just, A.C., 2019. Estimating daily PM2. 5 and PM10 over Italy using an ensemble model. Environ. Sci. Technol. 54, 120–128.
Sorek-Hamer, M., Chatfield, R., Liu, Y., 2020. Strategies for using satellite-based products in modeling PM2. 5 and short-term pollution episodes. Environ. Int. 144, 106057.
Stafoggia, M., Schwartz, J., Badaloni, C., Bellander, T., Alessandrini, E., Cattani, G., De'Donato, F., Gaeta, A., Leone, G., Lyapustin, A., 2017. Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. Environ. Int. 99, 234–244.
Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., De Hoogh, K., De'Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., 2019. Estimation of daily PM10 and PM2. 5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ. Int. 124, 170–179.
Stafoggia, M., Johansson, C., Glantz, P., Renzi, M., Shtein, A., Hoogh, K.d., Kloog, I., Davoli, M., Michelozzi, P., Bellander, T., 2020. A random forest approach to estimate daily particulate matter, nitrogen dioxide, and ozone at fine spatial resolution in Sweden. Atmosphere 11, 239.
Stafoggia, M., Oftedal, B., Chen, J., Rodopoulou, S., Renzi, M., Atkinson, R.W., Bauwelinck, M., Klompmaker, J.O., Mehta, A., Vienneau, D., Andersen, Z.J., Bellander, T., Brandt, J., Cesaroni, G., de Hoogh, K., Fecht, D., Gulliver, J., Hertel, O., Hoffmann, B., Hvidtfeldt, U.A., Jöckel, K.-H., Jørgensen, J.T., Katsouyanni, K., Ketzel, M., Kristoffersen, D.T., Lager, A., Leander, K., Liu, S., Ljungman, P.L.S., Nagel, G., Pershagen, G., Peters, A., Raaschou-Nielsen, O., Rizzuto, D., Schramm, S., Schwarze, P.E., Severi, G., Sigsgaard, T., Strak, M., van der Schouw, Y.T., Verschuren, M., Weinmayr, G., Wolf, K., Zitt, E., Samoli, E., Forastiere, F., Brunekreef, B., Hoek, G., Janssen, N.A.H., 2022. Long-term exposure to low ambient air pollution concentrations and mortality among 28 million people: results from seven large European cohorts within the ELAPSE project. Lancet Planet. Health 6, e9–e18.
Strak, M., Weinmayr, G., Rodopoulou, S., Chen, J., de Hoogh, K., Andersen, Z.J., Atkinson, R., Bauwelinck, M., Bekkevold, T., Bellander, T., Boutron-Ruault, M.-C., Brandt, J., Cesaroni, G., Concin, H., Fecht, D., Forastiere, F., Gulliver, J., Hertel, O., Hoffmann, B., Hvidtfeldt, U.A., Janssen, N.A.H., Jöckel, K.-H., Jørgensen, J.T., Ketzel, M., Klompmaker, J.O., Lager, A., Leander, K., Liu, S., Ljungman, P., Magnusson, P.K.E., Mehta, A.J., Nagel, G., Oftedal, B., Pershagen, G., Peters, A., Raaschou-Nielsen, O., Renzi, M., Rizzuto, D., van der Schouw, Y.T., Schramm, S., Severi, G., Sigsgaard, T., Sørensen, M., Stafoggia, M., Tjønneland, A., Verschuren, W. M.M., Vienneau, D., Wolf, K., Katsouyanni, K., Brunekreef, B., Hoek, G., Samoli, E., 2021. Long term exposure to low level air pollution and mortality in eight European cohorts within the ELAPSE project: pooled analysis. BMJ 374, n1904.
Tshehla, C., Wright, C.Y., 2019. 15 years after the national environmental management air quality act: is legislation failing to reduce air pollution in South Africa? South Afr. J. Sci. 115, 1–4.
Tyson, P., Garstang, M., Swap, R., Kallberg, P., Edwards, M., 1996. An air transport climatology for subtropical southern Africa. Int. J. Climatol. 16, 265–291.
Vapnik, V., 1999. The Nature of Statistical Learning Theory. Springer science & business media.
Vapnik, V., Golowich, S.E., Smola, A., 1997. Support vector method for function approximation, regression estimation, and signal processing. Adv. Neural Inf. Process. Syst. 281–287.
Wong, D.W., Yuan, L., Perlin, S.A., 2004. Comparison of spatial interpolation methods for the estimation of air quality data. J. Expo. Sci. Environ. Epidemiol. 14, 404–415.
World Health Organization, 2016. Ambient Air Pollution: a Global Assessment of Exposure and Burden of Disease. World Health Organization, Geneva.
World Health Organization, 2021. WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, Geneva.
Zhang, D., Du, L., Wang, W., Zhu, Q., Bi, J., Scovronick, N., Naidoo, M., Garland, R.M., Liu, Y., 2021. A machine learning model to estimate ambient PM2.5 concentrations in industrialized highveld region of South Africa. Rem. Sens. Environ. 266, 112713.